Improving Prediction Performance and Model Interpretability through Attention Mechanisms from Basic and Applied Research Perspectives

KITADA, Shunsuke / 北田, 俊輔

(開始ページ / Start Page)
1
(終了ページ / End Page)
156
(発行年 / Year)
2023-03-24
(学位授与番号 / Degree Number)
32675甲第573号
(学位授与年月日 / Date of Granted)
2023-03-24
(学位名 / Degree Name)
博士 (工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)
(URL)
https://doi.org/10.15002/00026672

Doctoral Dissertation Reviewed by Hosei University

Improving Prediction Performance and Model Interpretability through Attention Mechanisms from Basic and Applied Research Perspectives

Shunsuke Kitada

March 2023

Abstract

With the dramatic advances in deep learning technology, machine learning research is focusing on improving the interpretability of model predictions as well as prediction performance in both basic and applied research. While deep learning models have much higher prediction performance than conventional machine learning models, the specific prediction process is still difficult to interpret and/or explain. This is known as the black-boxing of machine learning models and is recognized as a particularly important problem in a wide range of research fields, including manufacturing, commerce, robotics, and other industries where the use of such technology has become commonplace, as well as the medical field, where mistakes are not tolerated.

Focusing on natural language processing tasks, we consider interpretability as the presentation of the contribution of a prediction to an input word in a recurrent neural network. In interpreting predictions from deep learning models, much work has been done mainly on visualization of importance mainly based on attention weights and gradients for the inference results. However, it has become clear in recent years that there are not negligible problems with these mechanisms of attention mechanisms and gradients-based techniques. The first is that the attention weight learns which parts to focus on, but depending on the task or problem setting, the relationship with the importance of the gradient may be strong or weak, and these may not always be strongly related. Furthermore, it is often unclear how to integrate both interpretations. From another perspective, there are several unclear aspects regarding the appropriate application of the effects of attention mechanisms to real-world problems with large datasets, as well as the properties and characteristics of the applied effects. This dissertation discusses both basic and applied research on how attention mechanisms improve the performance and interpretability of machine learning models.

From the basic research perspective, we proposed a new learning method that focuses on the vulnerability of the attention mechanism to perturbations, which contributes significantly to prediction performance and interpretability. Deep learning models are known to respond to small perturbations that humans cannot perceive and may exhibit unintended behaviors and predictions. Attention mechanisms used to interpret predictions are no exception. This is a very serious problem because current deep learning models rely heavily on this mechanism. We focused on training techniques using adversarial perturbations, i.e., perturbations that dares to deceive the attention mechanism. We demonstrated that such an adversarial training technique makes the perturbation-sensitive attention mechanism robust and enables the presentation of highly interpretable predictive evidence. By further extending the proposed technique to semi-supervised learning, a general-purpose learning model with a more robust and interpretable attention mechanism was achieved.

From the applied research perspective, we investigated the effectiveness of the deep learning models with attention mechanisms validated in the basic research, are in real-world applications. Since deep learning models with attention mechanisms have mainly been evaluated using basic tasks in natural language processing and computer vision, their performance when used as core components of applications and services has often been unclear. We confirm the effectiveness of the proposed framework with an attention mechanism by focusing on the real world of applications, particularly in the field of computational advertising, where the amount of data is large, and the interpretation of predictions is necessary. The proposed frameworks are new attempts to support operations by predicting the nature of digital advertisements with high serving effectiveness, and their effectiveness has been confirmed using large-scale ad-serving data.

In light of the above, the research summarized in this dissertation focuses on the attention mechanism, which has been the focus of much attention in recent years, and discusses its potential for both basic research in terms of improving prediction performance and interpretability, and applied research in terms of evaluating it for real-world applications using large data sets beyond the laboratory environment. The dissertation also concludes with a summary of the implications of these findings for subsequent research and future prospects in the field.

Acknowledgments

"Play hard, study hard" (よく遊び、よく学べ in Japanese) is the wonderful policy of our Iyatomi Lab where I am a member. This dissertation was written under this policy with the support of many people. In fact, I spent two years of my master's course and three years of my Ph.D. course immersed in research so much that I could say that I *played* with my research. I would like to thank you for giving me the opportunity to *play* a lot.

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Hitoshi Iyatomi for his generosity in sharing his thoughts, motivation, enthusiasm, and knowledge with me. As per his valued *"the door is always open"* policy, he was always available to give me advice and help. I was not a good student by any means and I was often a problem/trouble for his. I would like to repay his by completing this dissertation. The days I spent discussing the research with you were vivid and wonderful.

I would like to thank Yoshifumi Seki. Dr. Seki is my supervisor in industry, if Prof. Iyatomi is a supervisor in academia. Writing the top conference paper with him changed my life, leading me to go on to the Ph.D. course. He also cared about many other things towards me and I learnt a lot from him. A large part of this dissertation contains advice I received from him, for which I am very grateful.

I have been blessed with wonderful seniors. I would like to thank Daiki Shimada, Ryunosuke Kotani, Yusuke Kawasaki, Erika Fujita, and Takatoshi Okumura. Before belonging to the lab, I heard some rumors that Mr. Shimada, Mr. Kawasaki, and Ms. Fujita were great students. I wanted to study and do research with them, so I applied to belong to Iyatomi Lab. I was fascinated by the work that Mr. Shimada and Mr. Kotani were doing on natural language processing (NLP) and thought it would be great if I could do similar interesting research. I still remember attending my first domestic conference with Mr. Okumura. Even now, Mr. Okumura sometimes treats me to dinner, and I really appreciate it. They are still my seniors whom I admire.

I have been blessed with wonderful lab members of the same grade. I would like to thank Hayato Arai, Michiaki Ito, Takumi Saikawa, Hiroki Tani, Yusuke Chayama, Chihiro Hasegawa, Atsushi Minagawa, Hayao Munesato, and Seiya Murabayashi. Most of the classmates who had become good friends in university became members of the same lab, and we were able to continue our friendly competition. Their hard work and technical skills are quite impressive, and seeing their hard work up close every day was inspiring and had a positive impact on me personally. Some of the lab members came together to start their own businesses, while I went on to the Ph.D. course and took a different path. I was able to run through the Ph.D. course, encouraged by their entrepreneurial efforts. I am proud to have been able to do research at the same time as the wonderful lab members.

I have been blessed with wonderful juniors. In particular, I would like to thank Shunta Nagasawa, Takumi Aoki, Kazuya Ohata, Tubasa Nakagawa, Yusuke Tsushima, and Sota Nemoto, with whom I wrote and presented papers. They are colleagues with whom we had daily discussions and conducted research together as members of the same NLP group. Writing the paper with them helped me grow a great deal. Additionally, I would also send my special thanks to Kaito Odagiri. He was particularly unique among my junior colleagues in the lab and was always helpful and insightful.

I have been blessed with wonderful international students. I would like to thank Huu Quan Cap, Mahmoud Daif, and Gent Imeraj. They are the best friends I have ever had in the same lab. I learned so much from them, including English and their culture, and they helped me grow a lot. I hope to continue to grow with them.

I would like to thank Dr. Kyosuke Nishida and Prof. Chihiro Shibata for the feedback they provided as part of my dissertation committee. Your suggestions and comments have taken my dissertation to a higher level. Thank you very much for taking time out of your busy schedule to review my dissertation.

Finally, I would like to appreciate my parents for their long-standing support, mainly the economical support. Thank you very much for your constant and generous support of my research, which I was always free to do. I am usually too embarrassed to say this, but I would like to take this opportunity to express my enormous gratitude.

Thanks to JASSO and the Research JSPS Fellowship for Young Scientists (DC2) for their financial support. These supports encouraged me to carry out my daily research with peace of mind. This work was supported by JSPS KAKENHI under Grant 21J14143.

List of Publications

The papers related to this dissertation are indicated by †.

Journal Papers

- **Shunsuke Kitada**, Yuki Iwazaki, Riku Togashi, and Hitoshi Iyatomi, "DM^{2}S^2: Deep Multi-Modal Sequence Sets with Hierarchical Modality Attention," IEEE Access, vol. 10, pp. 120023-120034, 2022, DOI: 10.1109/ACCESS.2022.3221812.

- † **Shunsuke Kitada**, and Hitoshi Iyatomi, "Making attention mechanisms more robust and interpretable with virtual adversarial training for semi-supervised text classification," Springer Applied Intelligence, 2022. DOI: 10.1007/s10489-022-04301-w.

- † **Shunsuke Kitada**, Hitoshi Iyatomi, and Yoshifumi Seki, "Ad Creative Discontinuation Prediction with Multi-Modal Multi-Task Neural Survival Networks," Applied Sciences 12, no. 7: 3594, 2022. DOI: 10.3390/app12073594.

- † **Shunsuke Kitada** and Hitoshi Iyatomi, "Attention meets perturbations: Robust and interpretable attention with adversarial training," IEEE Access, vol. 9, pp. 92974-92985, 2021, DOI: 10.1109/ACCESS.2021.3093456.

Conference Papers

- Kazuya Ohata, **Shunsuke Kitada** and Hitoshi Iyatomi. "Feedback is Needed for Retakes: An Explainable Poor Image Notification Framework for the Visually Impaired." Proceedings of the IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI. 2022.

- Tsubasa Nakagawa, **Shunsuke Kitada**, and Hitoshi Iyatomi. "Expressions Causing Differences in Emotion Recognition in Social Networking Service Documents." Proceedings of the 31st ACM International Conference on Information and Knowledge Management. 2022.

- Takumi Aoki, **Shunsuke Kitada**, and Hitoshi Iyatomi. "Text Classification through Glyph-aware Disentangled Character Embedding and Semantic Sub-character Augmentation." Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop. 2020.

- Mahmoud Daif, **Shunsuke Kitada**, and Hitoshi Iyatomi. "AraDIC: Arabic Document Classification Using Image-Based Character Embeddings and Class-Balanced Loss." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2020.

- † **Shunsuke Kitada**, Hitoshi Iyatomi, and Yoshifumi Seki. "Conversion prediction using multi-task conditional attention networks to support the creation of effective ad creatives." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.

- **Shunsuke Kitada**, Ryunosuke Kotani, and Hitoshi Iyatomi. "End-to-end text classification via image-based embedding using character-level networks." Proceedings of the 2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2018.

Preprints

- **Shunsuke Kitada**, and Hitoshi Iyatomi. “Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning.” arXiv preprint arXiv:1809.02568 (2018).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Deep learning [1] is one of the machine learning (ML) models and has contributed greatly to the current development of artificial intelligence (AI). Compared to traditional ML models such as decision trees [2] and support vector machines [3], DL models have been shown to dramatically improve prediction performance in various fields because they can automatically learn important features from data to realize a goal through the training. DL models, which do not require knowledge about the subject, have demonstrated predictive and generative abilities beyond human capabilities, especially in the fields of computer vision (CV) [4, 5] and natural language processing (NLP) [6, 7]. Because ML/DL models will be used more frequently in the future, it is necessary for users to be able to interpret the validity of the prediction results of these models and the basis for them, in terms of the reliability and practicality of the models.

While the recent DL model has excellent prediction performance, it is difficult to interpret and explain the model prediction due to the complex architecture of the model. This black box and/or not transparent nature is an important issue that needs to be resolved [8]. Explainable AI is a field that seeks to explain the predictions of ML/DL models. This field has been studied for more than 40 years [9, 10], and classical explainable AI provided explanations based on roles constructed by humans carefully. Recent DL models have complex neural network (NN) structures consisting of various nonlinear transformations, making a human interpretation of the internal inference process very difficult. The interpretability of the predictions is recognized as a particularly important issue in a wide range of research fields, including manufacturing [11], e-commerce [12], and robotics [13], where the use of DL models is becoming common, as well as in the medical field [14] and autonomous driving [15], where mistakes are not tolerated.

1.1 Objectives and Requirements for Explainable AI

Explainability/interpretability for ML/DL models can be taken to mean a variety of things. In this section, we clarify our position on the interpretation and usage of these terms for ML/DL models in this dissertation, with examples of what is stated in the literature in the related fields. Previous studies have attempted to clarify the purpose and requirements for explainable AI [16–18]. Additionally, there are examples of companies such as Google [1] and Amazon [2], which place ML at the center of their business, discussing the provision of explanations from the standpoint of providing their systems and services.

Explainable AI should consider the audience because the contents and details of the reasons to be presented by depending on the audience. As a promising definition of explainable AI, Arrieta et al. [16] define it as follows:

> *Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

Additionally, Arrieta et al. [16], in light of the previous explainable AI research, have identified the following nine goals for the research: (1) **trustworthiness**, (2) **causality**, (3) **transferability**, (4) **informativeness**, (5) **confidence**, (6) **fairness**, (7) **accessibility**, (8) **interactivity**, (9) **privacy awareness**.

We agree with the above definition of Arrieta et al. [16] and aim to achieve the following goal

[1] `https://cerre.eu/wp-content/uploads/2020/07/ai_explainability_whitepaper_google.pdf`

[2] `https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf`

for the audiences: "*building explainable AI that provides reasons and details that make understanding easier.*" As described below, this dissertation is divided into discussions of basic and applied perspectives, each of which aims to define and improve "model interpretability" as follows. In the basic research perspective, the goal is to provide an interpretation that makes the basis for the predictions of multiple interpretation methods identical. This goal focuses on (1) trustworthiness, which is the state of trust that the model will function as intended when the problem is solved, and (5) confidence, which is the state of stability in the behavior of the model. In the applied research perspective, assuming that the audience is the operator of the service, the goal is to provide an interpretation that better supports the operator's decision-making. This goal focuses on (4) informativeness, the state in which humans can extract the information necessary for decision-making when solving real-world problems.

Throughout this dissertation, we will focus on the interpretation of each word in the input sentence/document in the prediction of NLP models. By making it possible for audiences to interpret where the model contributes predictions to the input, we believe that the above definition of explainable AI is satisfied. The ability to interpret the contribution of inputs from the DL model is useful for validating the model behavior, analyzing errors, and making decisions when operating the model in the real world.

1.2 Explainable AI Methods to Interpret the Prediction Results

There are two major approaches to the black box nature of DL models: (1) designing transparent models [19–21] and (2) providing post-hoc explanations for model predictions [12, 22, 23]. For (1), the design of transparent models involves research into understanding the architecture of the model itself [18] and the learning algorithms of the model [21]. In DL models, these works have limited applicability to the target model architecture, making them difficult to apply to existing models. They typically have limited prediction performance as they attempt to provide human-interpretable explanations. For (2), the post-hoc explanation methods involve visualization of factors that influence predictions [23, 24], and providing analytical explanations with concrete examples if applicable [25]. The post-hoc approach is widely used today because it generally applies to DL models and is easier than designing transparent models.

In the following, we explain the visualization approaches of importance based on the gradient for the prediction result and the learned attention weights in the post-hoc explanation, which is nowadays the mainstream interpretation of the predictions. First, we define the mathematical formulas that will be used throughout this dissertation with respect to these approaches.

We consider a recurrent-neural network (RNN) model for NLP task. The input of the model is word sequence $\boldsymbol{x} = (x_1, x_2, \cdots, x_T) \in \mathcal{V}^T$ of the length T where the words taken from vocabulary $\mathcal{V}$. The output of the model is $\hat{\boldsymbol{y}} \in \mathbb{R}^{\mathcal{C}}$ corresponds to ground truth $\boldsymbol{y} \in \mathbb{R}^{\mathcal{C}}$, where $\mathcal{C}$ is the set of class labels. We introduce the following short notation for the word sequence $(x_1, x_2, \cdots, x_T)$ as $(x_t)_{t=1}^T$. Let $\boldsymbol{w}_t \in \mathbb{R}^d$ be a d-dimensional word embedding corresponding to x_t. We represent each word with the word embeddings to obtain $(\boldsymbol{w}_t)_{t=1}^T \in \mathbb{R}^{d \times T}$. The word embeddings is encoded with encoder **Enc** to obtain the m-dimensional hidden states:

$$\boldsymbol{h}_t = \mathbf{Enc}(\boldsymbol{w}_t, \boldsymbol{h}_{t-1}) \in \mathbb{R}^m, \tag{1.1}$$

where $\boldsymbol{h}_0$ is the initial hidden state, and it is regarded as a zero vector.

1.2.1 Gradient-based Approach

The gradient-based approach estimates the contribution of input $\boldsymbol{x}$ to ground truth $\boldsymbol{y}$ or prediction $\hat{\boldsymbol{y}}$ by computing the partial derivative of $\boldsymbol{x}$ with respect to $\boldsymbol{y}$ or $\hat{\boldsymbol{y}}$. Here we use the term gradient for $\partial \boldsymbol{y} / \partial \boldsymbol{x}$. The goal of the gradient-based approach is to estimate attribution maps $\boldsymbol{g}^c = (g_i^c)_{i=1}^T$ for each word. The attribution maps $\boldsymbol{g}^c$ are considered to capture the importance of each input word for a particular output class $c \in \mathcal{C}$.

Following the advent of AlexNet [4], shortly after DL first came into the limelight, an early gradient-based approach was proposed by Simonyan et al. [24] and has long supported the interpretation of DL model predictions. They used a formulation similar to the above to present the attribute map to the user as a saliency map based on its absolute value:

$$\boldsymbol{g}^{c}_{\text{saliency}} = \left| \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right|. \tag{1.2}$$

The core idea of gradient-based methods is to map gradient information *backward* into the input space based on labels or inference results.

Because gradient-based approaches are applicable to all DL models trained by backpropagation, they have been used for many years in various fields for insight into the internal workings of models and for error analysis [22, 23, 26]. Since then, gradient-based approaches such as GradCAM [22], DeepLIFT [26], and Integrated Gradient [23] have been proposed to provide interpretations of predictions that provide deeper human insight, different from the rule-based approaches that were common during classical explainable AI. There is a large body of literature claiming that gradient-based approaches can be used to explain the importance of input features [27, 28].

Gradient-based approaches have a variety of advantages. First, gradient-based methods are fast and efficient in the computation of attribution maps. It is easily scalable because it does not depend on the number of input features, and it can be easily computed with high performance with the support of deep learning frameworks. Second, it can be applied to existing models and any network architecture with very few lines of code by overriding the gradient of nonlinearity in the computational graph. It is easy to implement because there is no need to implement custom layers or operations.

On the other hand, there are limitations to gradient-based approaches. The biggest problem is that the attribution maps presented are often visually noisy [29]. The gradient-based approaches assume that small changes in input features cause small changes in predictions, which may not always be the case [30]. Additionally, they are sensitive to the choice of input baseline, which can affect the attribution scores [23, 31]. The gradient-based approaches are not able to capture interactions between features and, therefore, may not provide a complete explanation of the model's predictions [32].

1.2.2 Attention-based Approach

In recent DLs, attention mechanisms [6, 33] are used to focus more *attention* on specific parts of the input as the model processes it. The effectiveness of the mechanism was initially validated in machine translation [33] and image captioning [34], but it is now being expanded in a wide variety of tasks. The attention mechanism works by assigning weights to each part of the input.

The key component in the attention mechanism is an attention score function $\mathcal{S}(\cdot, \cdot)$. The score function maps a query $\boldsymbol{Q}$ and key $\boldsymbol{K}$ to attention score $\tilde{a}_t$ for the t-th word. For the NLP tasks, we consider $\boldsymbol{K} = (\boldsymbol{h}_t)_{t=1}^{T}$. The attention scores $\tilde{\boldsymbol{a}} = (\tilde{a}_t)_{t=1}^{T}$ projected to sum to 1 by a alignment function $\mathcal{A}$, which results in the attention weights $\boldsymbol{a}$ as follows:

$$\boldsymbol{a} = \mathcal{A}(\boldsymbol{K}, \boldsymbol{Q}) = \phi(\mathcal{S}(\boldsymbol{K}, \boldsymbol{Q})), \tag{1.3}$$

where ϕ is a projection function to probability, such as softmax. The weight $\boldsymbol{a}$ indicates the importance of that portion to the current task. These weights are then used to selectively focus on the most important parts of the input to generate output. We then compute weighted instance vector $\boldsymbol{h}_{\boldsymbol{a}}$ as a weighted sum of the hidden states:

$$\boldsymbol{h}_{\boldsymbol{a}} = \sum_{i=1}^{T} a_t \boldsymbol{h}_t \tag{1.4}$$

Finally, $\boldsymbol{h}_a$ is fed to a prediction layer, which outputs a probability distribution over the set of categories $\mathcal{C}$.

The score function conventionally uses the following additive attention formulation:

$$\mathcal{S}_{\text{add}}(\boldsymbol{K}, \boldsymbol{Q})_i = \boldsymbol{v}^{\top} \tanh\left(\boldsymbol{W}_1 \boldsymbol{K}_t + \boldsymbol{W}_2 \boldsymbol{Q}\right), \tag{1.5}$$

where $W_1, W_2 \in \mathbb{R}^{d \times M}$ and $\boldsymbol{v} \in \mathbb{R}^d$ are the learnable parameters. Currently, the Transformer and other variants often use the following scaled-dot product attention formulation:

$$\mathcal{S}_{\text{prod}}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{\boldsymbol{K}^\top \boldsymbol{Q}}{\sqrt{m}} \tag{1.6}$$

In this dissertation, each evaluation is carried out with additive attention formulation as the score function.

While the gradient-based approach can interpret the prediction of a model without any interpretation/explanation mechanism, the attention-based DL model allows for the interpretation of the model by visualizing the learned attention weights. Explanations based on learned weights of attention mechanisms [33] can be shown in a *forward* direction for DL model inputs, giving a clear explanation that appeals to human intuition. Since the attention mechanism introduced in conventional RNN is generally placed before the last layer (or the last layer of the encoder in encoder-decoder architecture), the attention weights are learned to emphasize the part that contributes to the final prediction. For the Transformer model [6], which has been attracting much interest in recent years, attention rollout [35] has been proposed, which visualizes the result of the matrix product of the attention scores.

1.3 Discussion on Post-hoc Explanation Approaches

Although the attention mechanism contributes significantly to predictive performance and model interpretability, researchers find that the mechanism still suffers in the predictive interpretation of the model. In a claim that surprised the field, Jain and Wallace [36] argued that "attention is not an explanation." They reported the following two major problems in a simple RNN-based model with the attention mechanism, especially by NLP task: (1) There is not necessarily a strong correlation between the regions estimated to be important by attention weights and those obtained by gradients. (2) Small perturbations to the attention mechanism lead to unintended predictive changes, and those adversarial perturbations that deceive the mechanism lead to large predictive errors. For (1), the authors reported that Kendall's rank correlations [37] of importance indicated by the learned attention weight and word importance calculated by the gradient show almost no correlation. For (2), they found that small perturbations to the attention mechanisms lead to unintended prediction changes, while adversarial perturbations that deceive the mechanism lead to large prediction errors.

Against the weak explanatory nature of the attention mechanisms reported above, refutational analyses and methods to overcome them have been proposed in recent years [38–41]. On the analytical side, Wiegreffe and Pinter [38] argued that Jain and Wallace [36]'s claim depends on the definition of explanation and that testing it requires considering different aspects of the model with attention mechanisms in a rigorous experimental setting. Furthermore, similar to Wiegreffe and Pinter [38], several studies argue that the effects of explanations by attention mechanisms vary depending on the definition of explanatory properties [39, 40, 42]. With regard to improving the interpretability of the attention mechanism, there are some studies that re-examined the structure of the mechanism itself [43] or proposed a training technique that makes the attention mechanism robust to perturbations *indirectly* [44]. On the other hand, to our knowledge, no technique has been proposed to *directly* address the vulnerability of attention mechanisms to perturbation, as pointed out in Jain and Wallace [36].

In the light of the definition of explainable AI in the Arrieta et al. [16], if each interpretation approach provides a different interpretation, it will have a negative impact on the audience, especially in the trustworthiness and confidence. As described above, various studies have been conducted on the interpretability of DL models, focusing on the gradient- and attention-based approaches. However, a comprehensive synthesis of these multiple approaches to obtain reliable and robust interpretations is an important basic research direction, and there remains room for further research.

1.4 Interpreting Deep Learning Models in Real-World Applications

While the above discussion focused on the interpretation of DL models in the basic research perspectives, this section describes the provision of interpretation in DL models that are operated in the real world. To begin with, we emphasize that there are still few studies that have validated using real-world datasets under a practical situation. We speculate that this may be because information about applications in the real world is often confidential. Therefore, DL-based models, including attention-based models, have been developed in fields such as machine translation [6, 33], machine reading comprehension [7], and image classification [45]. However, the development and evaluation of such models have been carried out on relatively small and well-organized data sets, limited to so-called laboratory environments. One of the root causes of the limitation is the lack of publicly available dataset, which is important for the research area. The effectiveness of DL models outside of these data-available areas remains to be developed. So far, attention-based DL models have been reported to perform well on various tasks. On the other hand, it remains unclear how the DL models perform on noisier, more imbalanced, and diverse real-world data that may deviate from the benchmark dataset.

The literature evaluating its interpretability in DL models operating in the real world is even less extensive than the literature evaluating its performance on real-world datasets. According to Arrieta et al. [16] and the industry tech giants, Explainable AI is expected to be an AI that provides such audience/operators with details and reasons to clarify their own behavior or to facilitate understanding. Furthermore, there is a certain need to build a DL model that can provide high prediction performance and prediction evidence, which have not been achieved by conventional ML models. In sum, there is currently limited research on developing models that can be interpreted for practical use of large real-world data. This point remains an important applied research issue.

1.5 The Proposals of this Dissertation

1.5.1 Basic Research Perspectives

In terms of the basic research for this dissertation, Chapters 2 and 3 describe the vulnerability of the attention mechanism, which is essential to DL models, to perturbations and countermeasures against it. We further discuss the interpretability of our technique after its application. The majority of prior studies have suggested that interpretation is possible by investigating where the attention mechanisms assign large weight to the model inputs. On the other hand, we believed that the mechanism was vulnerable to noise, which would negatively affect prediction performance and interpretability.

To overcome the above challenges, a natural idea is: to introduce adversarial perturbation to the attention mechanism that is vulnerable to perturbation so that it learns to be robust to noise. We focus on adversarial training (AT) [46] to deal with adversarial examples [47] that produce inaccurate model output. AT [46] was first proposed in the field of image recognition as a method to overcome the weak point where the model can be fooled by input small noise/perturbation that is imperceptible to humans. Since then, AT has been introduced in the NLP field and has demonstrated its usefulness in areas which is difficult for the DL model to predict, even when the text is mixed with metaphorical expressions that are understandable to humans (e.g., fake news detection [48]). While AT in NLP is often applied to the input space, its effects when applied to attention mechanisms remain unclear.

The basic research perspective of this dissertation proposes a new training technique that focuses on the vulnerability to perturbations of the attention mechanism and contributes significantly to prediction performance and prediction interpretation. We consider using AT, in which perturbations are applied to deceive the mechanisms, exploiting adversarial perturbations. By employing the proposed technique, DL model will be trained to pay stronger attention to areas that are more important for prediction, which is expected to not only improve prediction performance but also improve model interpretability.

Adversarial Training for Attention Mechanism

Inspired by AT [46], a powerful regularization technique for enhancing model robustness, we aim to overcome the vulnerability of the attention mechanism to perturbations. In Chapter 2, we propose a general training technique for natural language processing tasks, including AT for attention (Attention AT) and more interpretable AT for attention (Attention iAT). The proposed techniques improved the prediction performance and the model interpretability by exploiting the mechanisms with AT. In particular, Attention iAT boosts those advantages by introducing adversarial perturbation, which enhances the difference in the attention of the sentences. Evaluation experiments with ten open datasets revealed that AT for attention mechanisms, especially Attention iAT, demonstrated (1) the best performance in nine out of ten tasks and (2) more interpretable attention (i.e., the resulting attention correlated more strongly with gradient-based word importance) for all tasks. Additionally, the proposed techniques are (3) much less dependent on perturbation size in AT.

Virtual Adversarial Training for Attention Mechanism

AT has successfully reduced the disadvantage of being vulnerable to perturbations to the attention mechanisms by considering adversarial perturbations, as shown in Chapter 2. However, this technique requires label information, and thus, its use is limited to supervised settings. In Chapter 3, we explore the approach of incorporating virtual AT (VAT) [49] into the attention mechanisms, by which adversarial perturbations can be computed even from unlabeled data. To realize this approach, we propose two general training techniques, namely VAT for attention mechanisms (Attention VAT) and "interpretable" VAT for attention mechanisms (Attention iVAT), which extend AT for attention mechanisms to a semi-supervised setting. In particular, Attention iVAT focuses on the differences in attention; thus, it can efficiently learn clearer attention and improve model interpretability, even with unlabeled data. Empirical experiments based on six public datasets revealed that our techniques provide better prediction performance than conventional AT-based as well as VAT-based techniques, and stronger agreement with evidence that is provided by humans in detecting important words in sentences. Moreover, our proposal offers these advantages without needing to add the careful selection of unlabeled data. That is, even if the model using our VAT-based technique is trained on unlabeled data from a source other than the target task, both the prediction performance and model interpretability can be improved.

1.5.2 Applied Research Perspectives

In terms of applied research for this dissertation, Chapters 4 and 5 describe research on applications of ML/DL models in NLP tasks. In particular, we focus on applications to the field of computational advertising [50], which has a significant impact on business. The field of computational advertising is a relatively new and important business-related research topic, dealing with very large online data volumes of the order of 100 million. Display advertising is a type of online advertising in which an advertiser pays a publisher to display graphical material on the publisher's web page or application. This graphical material, which primarily contains images and text, is commonly referred to as ad creative and serves to effectively provide product information to consumers who are willing to buy. The market for digital advertising has been expanding enormously and is expected to grow further in the future. In fact, the IAB internet advertising revenue report 2021 [51] states that "digital advertising revenue increased 35.4% year over year, the highest growth since 2006." Since it is difficult to operate ad data due to its large scale manually, a methodology is expected to provide operational support by means of a computer to distribute highly effective ads and discontinue ineffective ads.

Mainstream research in online advertising estimates click-through rate (CTR) and conversion rates (CVR) for users of the ads being served. For such tasks, various methods based on ML/DL have recently emerged [52–56], including the variants of factorization machines [57–60] that can consider feature interactions. While these have been evaluated and assessed on anonymized ad-serving benchmark datasets such as Criteo[3] and Avazu[4] and have reported some improvement in prediction performance, very few studies have evaluated them on other data, including real-world data [61, 62]. This is because ad data is generally very complex in terms of rights involving a large

[3] http://labs.criteo.com/2013/12/download-terabyte-click-logs/

[4] https://www.kaggle.com/c/avazu-ctr-prediction

number of stakeholders, making it difficult to disclose. Additionally, there are very few studies on the prediction of effectiveness for ads with no delivery performance, the so-called cold start setting [63].

The applied aspect of this dissertation describes the practical application of NLP technology to the computational advertising field and discuss the prediction performance and model interpretability of the proposed frameworks. Chapter 4 describes a framework for evaluating ad creatives that are more effective in terms of serving. Although it is important to evaluate in advance what kind of ad creatives should be served, it is also important to provide evidence as to why the ad creatives are effective in order to improve newly created ad creatives and judge the validity of such evaluations. Chapter 5 describes a framework for automating the discontinuation of ad creatives with less serving effectiveness. Predicting the timing of ad discontinuation itself is important. Additionally, providing some evidence as to why the ad creative is discontinued at a certain time is an important indicator for advertisers and ad operators to consider whether to trust the prediction. This is a major factor in confidence in the DL model.

Operational Support for More Effective Ad Creatives

Accurately predicting conversions in advertisements is generally a challenging task because such conversions do not occur frequently. In Chapter 4, we propose a new framework to support creating high-performing ad creatives, including the accurate prediction of ad creative text conversions before serving them to the consumer. The proposed framework includes the key ideas needed to train the model: multi-task learning and conditional attention mechanism. The multi-task learning is an idea to improve prediction performance by simultaneously predicting clicks and conversions in order to overcome the difficulty of prediction due to data imbalance. The conditional attention mechanism is a new mechanism that can take into account the attribute values of the ad creative, such as the genre of the ad creative and the gender of the delivery target, to further improve the performance of conversion prediction. We evaluated the proposed framework on actual large-scale serving history data and confirmed that these ideas improved the performance of conversion prediction.

The conditional attention mechanism incorporated in the proposed framework is capable of interpreting word expressions that contribute to the effectiveness of ad serving. Attention highlighting using the learned conditional attention mechanism predicts conversions in advance, taking into account the attribute values set at the time of ad submission, and simultaneously visualizes the importance of the words that contributed to the prediction. By observing the attention highlighting when the proposed method predicts a high number of conversions for a prototype ad creative, it is possible to confirm the word expressions that better fits the target ad attributes.

Operational Support for Less Effective Ad Creatives

Discontinuing ad creatives at an appropriate time is one of the most important ad operations that can have a significant impact on sales. Such operational support for ineffective ads has been less explored than that for effective ads. After pre-analyzing 1,000,000 real-world ad creatives, we found that there are two types of discontinuation: short-term (i.e., cut-out) and long-term (i.e., wear-out). In Chapter 5, we propose a practical prediction framework for the discontinuation of ad creatives with a hazard function-based loss function inspired by survival prediction. Our framework accurately predicts the appropriate timing for the discontinuation of two types of digital advertisements, short-term and long-term. The framework consists of two main techniques: (1) a two-term estimation technique with multi-task learning and (2) a click-through rate-weighting technique for the loss function. We evaluated our framework using 1,000,000 real-world ad creatives, including 10 billion scale impressions.

The attention mechanism incorporated in the proposed framework allows us to interpret the word expressions that contributes to ad discontinuation. While various factors determine ad discontinuation, we expect that short- and long-term discontinuation in the ad text will show different trends. Specifically, in the short-term discontinuation, it is possible to identify word expressions in which the user did not show interest. On the other hand, in the long-term discontinuation, it is possible to confirm word expressions that are no longer in season. Based on the interpretation of these word expressions, the operator can make a final decision to discontinue the ad creatives. Unfortunately, due to restrictions on information disclosure, quantitative evaluation of interpretability by specific attention visualization is not possible. Meanwhile, we report that the performance of the framework is sufficient to support ad operators, and the word-by-word visualization the framework provides has significant advantages that could support discontinuation.

1.6 Thesis Structure

The remainder of this thesis is organized as follows. The first two are chapters on aspects of basic research, and the next two are chapters on aspects of applied research. The last chapter is the conclusion of this thesis:

- **Chapter 2** describes adversarial training for attention mechanisms.
- **Chapter 3** describes virtual adversarial training for attention mechanisms.
- **Chapter 4** describes conversion prediction for ad creatives.
- **Chapter 5** describes discontinuation prediction for ad creatives.
- **Chapter 6** provides a discussion of the applicability, interpretability, and development of the proposed method throughout this dissertation.
- **Chapter 7** contains the conclusion section of this dissertation.

Chapter 2

Adversarial Training for Attention Mechanisms

A serious problem was pointed out in Jain and Wallace [36] that attention mechanism, a key component of current deep learning models, is vulnerable to perturbations. This chapter 2 aims to address the problem by employing adversarial training to attention mechanism. An overview of attention mechanism and the background of our proposed technique to its vulnerabilities is described in Section 2.1. Related work on attention mechanisms and adversarial training is described in Section 2.2. We propose new adversarial training techniques for attention mechanism (Attention AT) and interpretable one (Attention iAT) in Section 2.3 that make attention mechanisms robust to perturbations. Our proposed techniques introduce perturbations that deceive the attention mechanism during training, and we present the experimental conditions under which we compare our techniques to the conventional AT for word embedding in Section 2.4. We confirm in Section 2.5 that our techniques perform better in terms of prediction performance and interpretability than the conventional techniques that introduce perturbations to word embeddings. We present in Section 2.6 the comparison and discussion of the proposed techniques in terms of AT for attention and word embeddings, and in terms of the type of perturbation. Finally, we summarize this chapter in Section 2.7.

2.1 Background

Attention mechanisms [33] are widely applied in NLP field through DNNs. As the effectiveness of attention mechanisms became apparent in various tasks [41, 64–68], they were applied not only to RNNs but also to CNNs. Moreover, Transformers [6] which make proactive use of attention mechanisms have also achieved excellent results. However, it has been pointed out that DNN models tend to be locally unstable, and even tiny perturbations to the original inputs [47] or attention mechanisms can mislead the models [36]. Specifically, Jain and Wallace [36] used a practical bidirectional RNN (BiRNN) model to investigate the effect of attention mechanisms and reported that learned attention weights based on the model are vulnerable to perturbations.[1]

The Transformer [6] and its follow-up models [70, 71] have self-attention mechanisms that estimate the relationship of each word in the sentence. These models take advantage of the effect of the mechanisms and have shown promising performances. Thus, there is no doubt that the effect of the mechanisms is extremely large. However, they are not easy to train, as they require huge amounts of GPU memory to maintain the weights of the model. Recently, there have been proposals to reduce memory consumption [72], and we acknowledge the advantages of the models. On the other hand, the application of attention mechanisms to DNN models, such as RNN and CNN models, which have been widely used and do not require relatively high training requirements, has not been sufficiently studied.

In this chapter, we focus on improving the robustness of commonly used BiRNN models (as described detail in Section A.1) to perturbations in the attention mechanisms. Furthermore, we demonstrate that the result of overcoming the vulnerability of the attention mechanisms is an improvement in the prediction performance and model interpretability.

To tackle the models' vulnerability to perturbation, Goodfellow et al. [46] proposed adversarial training (AT) that increases robustness by adding adversarial perturbations to the input and the training technique forcing the model to address its difficulties. Previous studies [46, 73] in the

[1] In Jain and Wallace [36], the vulnerability of attention mechanisms to perturbations is confirmed with an RNN-based model [36]. In this chapter, we focus on the model, and Transformer [6]-based model such as BERT [69] and their successor models [70, 71] will be future work.

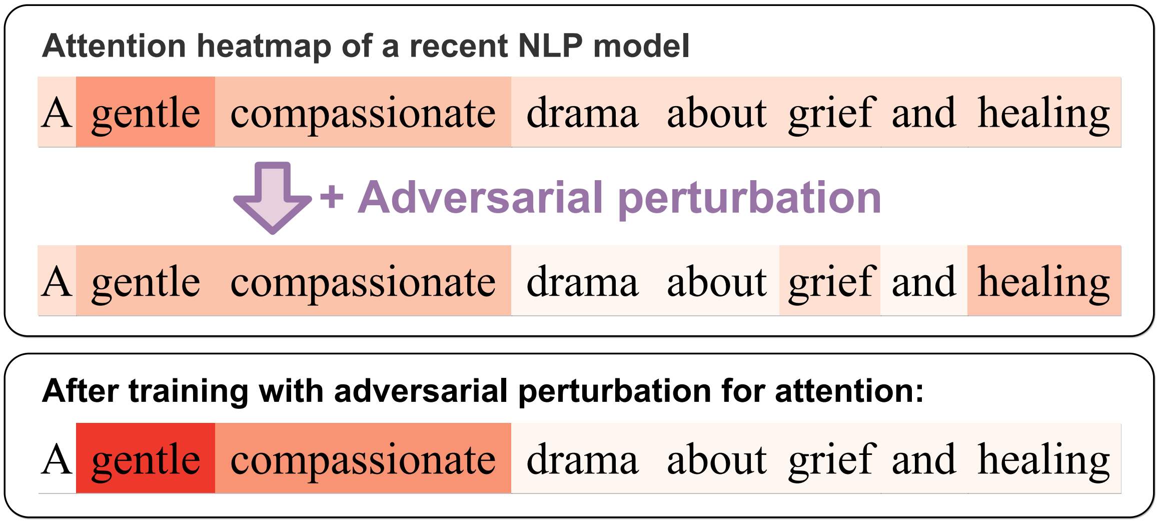

Attention heatmap of a recent NLP model
A gentle compassionate drama about grief and healing
+ Adversarial perturbation
A gentle compassionate drama about grief and healing
After training with adversarial perturbation for attention:
A gentle compassionate drama about grief and healing

Figure 2.1: An example of an attention heatmap for a BiRNN model with attention mechanisms and the model with attention mechanisms trained with adversarial training from the Stanford Sentiment Treebank (SST) [74]. The proposed adversarial training for attention mechanisms helps the model learn cleaner attention.

image recognition field have theoretically explained the regularization effect of AT and shown that it improves the robustness of the model for unseen images.

AT is also widely used in the NLP field as a powerful regularization technique [75–78]. In pioneering work, Miyato et al. [75] proposed a simple yet effective technique to improving the text classification performance by applying AT to a word embedding space. Later, interpretable AT (iAT) was proposed to increase the interpretability of the model by restricting the direction of the perturbations to existing words in the word embedding space [76]. The attention weight of each word is considered an indicator of the importance of each word [79], and thus, in terms of interpretability, we assume that the weight is considered a higher-order feature than the word embedding. Therefore, AT for attention mechanisms that adds an adversarial perturbation to deceive the attention mechanisms is expected to be more effective than AT for word embedding.

From motivations above, we propose a new general training technique for attention mechanisms based on AT, called *adversarial training for attention* (Attention AT) and *more interpretable adversarial training for attention* (Attention iAT). The proposed techniques are the first attempt to employ AT for attention mechanisms. The proposed Attention AT/iAT is expected to improve the robustness and the interpretability of the model by appropriately overcoming the adversarial perturbations to attention mechanisms [80–82]. Because our proposed AT techniques for attention mechanisms is model-independent and a general technique, it can be applied to various DNN models (e.g., RNN and CNN) with attention mechanisms. Our technique can also be applied to any similarity functions for attention mechanisms, e.g, additive function [33] and scaled dot-product function [6], which is famous for calculating the similarity in attention mechanisms.

To demonstrate the effects of these techniques, we evaluated them compared to several other state-of-the-art AT-based techniques [75, 76] with ten common datasets for different NLP tasks. These datasets included binary classification (BC), question answering (QA), and natural language inference (NLI). We also evaluated how the attention weights obtained through the proposed AT technique agreed with the word importance calculated by the gradients [24]. Evaluating the proposed techniques, we obtained the following findings concerning AT for attention mechanisms in NLP:

- AT for attention mechanisms improves the prediction performance of various NLP tasks.
- AT for attention mechanisms helps the model learn cleaner attention (as shown in Figure 2.1) and demonstrates a stronger correlation with the word importance calculated from the model gradients.
- The proposed training techniques are much less independent concerning perturbation size in AT.

Especially, our Attention iAT demonstrated the best performance in nine out of ten tasks and more interpretable attention, i.e., resulting attention weight correlated more strongly with the gradient-based word importance [24]. The implementation required to reproduce these techniques and the evaluation experiments are available on GitHub.[2]

[2] `https://github.com/shunk031/attention-meets-perturbation`

2.2 Related Work

2.2.1 Attention Mechanisms

Attention mechanisms were introduced by Bahdanau et al. [33] for the task of machine translation. Today, these mechanisms contribute to improving the prediction performance of various tasks in the NLP field, such as sentence-level classification [64], sentiment analysis [41], question answering [65], and natural language inference [66]. There are a wide variety of attention mechanisms; for instance, additive [33] and scaled dot-product [6] functions are used as similarity functions.

Attention weights are often claimed to offer insights into the inner workings of DNNs [79]. However, Jain and Wallace [36] reported that learned attention weights are often uncorrelated with the word importance calculated through the gradient-based method [24], and perturbations interfere with interpretation. In this chapter, we demonstrate that AT for attention mechanisms can mitigate these issues.

2.2.2 Adversarial Training

AT [46, 47, 83] is a powerful regularization technique that has been primarily explored in the field of image recognition to improve the robustness of models for input perturbations. In the NLP field, AT has been applied to various tasks by extending the concept of adversarial perturbations, e.g., text classification [75, 76], part-of-speech tagging [77], and machine reading comprehension [78, 84]. As mentioned earlier, these techniques apply AT for word embedding. Other AT-based techniques for the NLP tasks include those related to parameter updating [85] and generative adversarial network (GAN)-based retrieval-enhancement method [86]. Our proposal is an adversarial training technique for attention mechanisms and is different from these methods.

Miyato et al. [75] proposed Word AT, a technique that applied AT to the word embeddings. The adversarial perturbations are generated according to the back-propagation gradients. These perturbations are expected to regularize the model. Since then, Sato et al. [76] proposed Word iAT, and it has been known to achieve almost the same performance as Word AT that does not expect interpretability [76]. The Word iAT technique aims to increase the model's interpretability by determining the perturbation's direction so that it is closer to other word embeddings in the vocabulary. Both reports demonstrated improved task performance via AT. However, the specific effect of AT on attention mechanisms has yet to be investigated. In this chapter, we aim to address this issue by providing analyses of the effects of AT for attention mechanisms using various NLP tasks.

AT is considered to be related to other regularization techniques (e.g., dropout [87], batch normalization [88]). Specifically, dropout can be considered a kind of noise addition. Word dropout [89] and character dropout [90], known as wildcard training, are variants for NLP tasks. These techniques can be considered random noise for the target task. In contrast, AT has been demonstrated to be effective because it creates particularly vulnerable perturbations that the model is trained to overcome [46].

It has been reported that DNN models that introduce adversarial training to overcome adversarial perturbations capture human-like features [80–82]. These features help to make the prediction of DNN models easier to interpret for humans. In this chapter, we demonstrate that the proposed AT to attention mechanisms provides cleaner attention that is more easily interpreted by humans.

2.3 Adversarial Training for Attention Mechanisms

The main contribution of this paper is to explore the idea of employing AT for attention mechanisms. In this chapter, we propose a new training technique for attention mechanisms based on AT, called Attention AT and Attention iAT. The proposed techniques aim to achieve better regularization effects and to provide better interpretation of attention in the sentence. These techniques are the first application of AT to the attention in each word, which is expected to be more interpretable, with reference to AT for word embeddings [75] and a technique more focused on interpretability [76]. In this chapter, we generate adversarial perturbations based on the model described in Section A.1.

2.3.1 Attention AT: Adversarial Training for Attention

We describe the proposed Attention AT, which features adversarial perturbations in the attention mechanisms rather than in the word embeddings [75, 76]. The adversarial perturbation on the mechanisms is defined as the worst-case perturbation on attention mechanisms of a small bounded norm ϵ that maximizes loss function $\mathcal{L}$ of the current model:

$$\boldsymbol{r}_{\texttt{AT}} = \underset{\boldsymbol{r}:||\boldsymbol{r}||_2 \leq \epsilon}{\operatorname{argmax}} \, \mathcal{L}(X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}), \tag{2.1}$$

where $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$ is the input sequence with attention score $\tilde{\boldsymbol{a}}$, its perturbation $\boldsymbol{r}$, $\boldsymbol{y}$ is the target output, and $\hat{\boldsymbol{\theta}}$ represents the current model parameters. We apply the fast gradient method [69, 75], i.e., first-order approximation to obtain an approximate worst-case perturbation of norm ϵ, through a single gradient computation as follows:

$$\boldsymbol{r}_{\texttt{AT}} = \epsilon \frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \text{where } \boldsymbol{g} = \nabla_{\tilde{\boldsymbol{a}}} \mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}). \tag{2.2}$$

ϵ is a hyper-parameter to be determined using the validation dataset. We find this $\boldsymbol{r}_{\texttt{AT}}$ against the current model parameterized by $\hat{\boldsymbol{\theta}}$ at each training step and construct an adversarial perturbation for attention score $\tilde{\boldsymbol{a}}$:

$$\tilde{\boldsymbol{a}}_{\texttt{adv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\texttt{AT}}. \tag{2.3}$$

2.3.2 Attention iAT: Interpretable Adversarial Training for Attention

We describe the proposed Attention iAT for further boosting the prediction performance and the interpretability of NLP tasks. Rather than utilizing AT to attention mechanisms (as described in Section 2.3.1), Attention iAT effectively exploits differences in the attention to each word in a sentence for the training. As a result, this technique provides cleaner attention in the sentence and improves the interpretability of the attention. These effects contribute to improving the performance of various NLP tasks.

In terms of formulation, the proposed Attention iAT is analogous to interpretable AT for word embeddings (Word iAT) [76], which increases the interpretability of AT for word embeddings in formulas. However, the implications and effects for training the model are very different; in the proposed Attention iAT, the *attention difference enhancement*, described later, enhances the difference in attention for each word. The difference and its effect will be explained later in this section and discussed in Section 2.6.3.

Suppose $\tilde{a}_t$ denotes the attention score corresponding to the t-th position in the sentence. We define the difference vector $\boldsymbol{d}_t$ as the difference between the attention to the t-th word $\tilde{a}_t$ in a sentence and the attention to any k-th word $\tilde{a}_k$:

$$\boldsymbol{d}_t = (d_{t,k})_{k=1}^{T} = (\tilde{a}_t - \tilde{a}_k)_{k=1}^{T}. \tag{2.4}$$

$T = T_S$ in single sequence task, and $T = T_P$ in a pair sequence task. By normalizing the norm of the vector, we define a normalized difference vector of the attention for the t-th word:

$$\tilde{\boldsymbol{d}}_t = \frac{\boldsymbol{d}_t}{||\boldsymbol{d}_t||_2}. \tag{2.5}$$

The number of dimensions in d_k is the number of the vocabulary (fixed length) for Word iAT, while the dimension of words in a sentence (variable length) for Attention iAT. The dimensionality of d_t in Attention iAT is much smaller compared to Word iAT.[3] We define perturbation $r(\boldsymbol{\alpha}_t)$ for attention to the t-th word with trainable parameters $\boldsymbol{\alpha}_t = (\alpha_{t,k})_{t=1}^{T} \in \mathbb{R}^T$ and the normalized difference vector of the attention $\tilde{\boldsymbol{d}}_t$ as follows:

$$r(\boldsymbol{\alpha}_t) = \boldsymbol{\alpha}_t^{\top} \cdot \tilde{\boldsymbol{d}}_t. \tag{2.6}$$

By combining $\boldsymbol{\alpha}_t$ for all t, we can calculate perturbation $\boldsymbol{r}(\boldsymbol{\alpha})$ for the sentence:

$$\boldsymbol{r}(\boldsymbol{\alpha}) = \{r(\boldsymbol{\alpha}_t)\}_{t=1}^{T}. \tag{2.7}$$

[3] This normalization is done on a sentence-by-sentence basis, so it does not matter that the dimension is varies ($T << V$).

Then, similar to $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$ in Eq. 2.1, we introduce $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{\alpha})}$ and seek the worst-case weights of the difference vectors that maximize the loss functions as follows:

$$\boldsymbol{r}_{\texttt{iAT}} = \underset{\boldsymbol{\alpha}:||\boldsymbol{\alpha}||\leq\epsilon}{\operatorname{argmax}} \mathcal{L}(X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{\alpha})}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}). \tag{2.8}$$

Contrary to Attention iAT, in Word iAT, the difference $d_{t,k}$ in Eq. 2.4 is defined as the distance between the t-th word in a sentence and the k-th word in the vocabulary in the word embedding space. Based on the distance, Word iAT determines the direction of perturbation for the t-th word as a linear sum of the word directional vectors in the vocabulary. In contrast, Attention iAT does not compute the distance to word embeddings in the vocabulary. Instead, this technique computes the difference in attention to other words in the sentence and determines the direction of the perturbation. The adversarial perturbation of Attention iAT, defined in this way, works to increase the difference in attention to each word. We call this process in Attention iAT as *attention difference enhancement*. Owing to the process, Attention iAT improves the interpretability of attention and contributes to the performance of the model's prediction. The detail discussions are shared in Section 2.6.3.

For computational efficiency, we calculate the interpretable adversarial perturbation by applying the same approximation method as in Eq. 2.2:

$$\boldsymbol{r}_{\texttt{iAT}} = \epsilon \frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \text{where } \ \boldsymbol{g} = \nabla_{\boldsymbol{\alpha}} \mathcal{L}(X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{\alpha})}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}). \tag{2.9}$$

Then, similar to Eq. 2.3, we construct a perturbated example for attention score $\tilde{\boldsymbol{a}}$:

$$\tilde{\boldsymbol{a}}_{\texttt{iadv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\texttt{iAT}}. \tag{2.10}$$

Table 2.1: Dataset statistics. We used the three most well-known NLP tasks for evaluation: binary classification (BC), question answering (QA), and natural language inference (NLI). We split the dataset into training, validation, and test sets. We performed preprocessing as shown at `https://github.com/successar/AttentionExplanation` in the same manner as Jain and Wallace [36]. See the details in Appendix B.1. Jain and Wallace [36] split the dataset into only a training set and a test set, so we did not get the same result.

Task	Dataset			# class	# train	# valid	# test	# vocab	Avg. # words
Binary Classification (BC)	SST		[74]	2	6,920	872	1,821	13,723	18
	IMDB		[91]	2	17,186	4,294	4,353	12,485	171
	20News		[92]	2	1,145	278	357	5,986	110
	AGNews		[93]	2	51,000	9,000	3,800	13,713	35
Question Answering (QA)	CNN news		[94]	584	380,298	3,924	3,198	70,192	773
	bAbI	Task 1	[95]	6	8,500	1,500	1,000	22	39
		Task 2		6	8,500	1,500	1,000	36	98
		Task 3		6	8,500	1,500	1,000	37	313
Natural Language Inference (NLI)	SNLI		[96]	3	549,367	9,842	9,824	20,979	22
	Multi NLI		[97]	3	314,161	78,541	19,647	53,112	34

2.3.3 Training a Model with Adversarial Training

At each training step, we generate adversarial perturbation in the current model. To this end, we define the loss function for adversarial training as follows:

$$\tilde{\mathcal{L}} = \underbrace{\mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \boldsymbol{\theta})}_{\substack{\text{The loss from} \\ \text{unmodified examples}}} + \underbrace{\lambda \mathcal{L}(X_{\tilde{\boldsymbol{a}}_{\texttt{ADV}}}, \boldsymbol{y}; \boldsymbol{\theta})}_{\substack{\text{The loss from} \\ \text{its adversarial examples}}} , \tag{2.11}$$

where λ is the coefficient that controls the balance between two loss functions. Note that $X_{\tilde{\boldsymbol{a}}_{\texttt{ADV}}}$ can be $X_{\tilde{\boldsymbol{a}}_{\texttt{adv}}}$ for Attention AT or $X_{\tilde{\boldsymbol{a}}_{\texttt{iadv}}}$ for Attention iAT.

2.4 Experiments

In this section, we describe the evaluation tasks and datasets, the details of the models, and the evaluation criteria.

2.4.1 Tasks and Datasets

We evaluated the proposed techniques using the open benchmark tasks (i.e., four BCs, four QAs, and two NLIs) used in Jain and Wallace [36]. In our experiment, we added MultiNLI [97] as an additional NLI task for more detailed analysis (see the details in Appendix B.1). Table 3.2 presents the statistics for all datasets. We split the dataset into a training set, a validation set, and a test set.[4] We performed preprocessing, including tokenization with spaCy[5], mapping out vocabulary words to a special `<unk>` token, and mapping all words with numeric characters to `qqq` in the same manner as Jain and Wallace [36].

2.4.2 Model Settings

We compared the two proposed training techniques to four conventional training techniques. They were implemented using the same model architecture as described in Section A.1. Following Jain and Wallace [36], we used bi-directional long short-term memory (LSTM) [98] as the BiRNN-based encoder, including $\mathbf{Enc}$, $\mathbf{Enc}_P$, and $\mathbf{Enc}_Q$. A total of six training techniques were evaluated in the experiments:

- **Vanilla** [36]: a model with attention mechanisms trained without the use of AT.
- **Word AT** [75]: word embeddings trained with AT.
- **Word iAT** [76]: word embeddings trained with iAT.
- **Attention RP**: attention to the word embeddings is trained with random perturbation (RP).
- **Attention AT** (**Proposed**): attention to the word embeddings is trained with AT.
- **Attention iAT** (**Proposed**): attention to the word embeddings is trained with iAT.

We implemented the training techniques above using the AllenNLP library with Interpret [99, 100]. Through the experiments, we set the hyper-parameter $\lambda = 1$ related to AT or iAT in Eq. 3.14. To

[4] Jain and Wallace [36] split the dataset into only a training set and a test set, so we did not get the same results.

[5] spaCy · Industrial-strength Natural Language Processing in Python `https://spacy.io/`

ensure a fair comparison of the training techniques, we followed the configurations (e.g., initialization of word embedding, hidden size of the encoder, optimizer settings) used in the literature [36] (see the details in Appendix D.1).

Note that while Jain and Wallace [36] used a test set to adjust the model's hyper-parameters, we used a validation set. In adversarial training, the Allentune library [101] was used to adjust hyper-parameter ϵ, and we report the test scores for the model with the highest validation score.

2.4.3 Evaluation Criteria

First, we compared the prediction performance of each model for each task. As an evaluation metric of the prediction performance, we used the F1 score[6], accuracy, and the micro-F1 score for the BC, QA, and NLI, respectively, as in [36].

Next, we compared how the attention weights obtained through the proposed AT-based technique agreed with the importance of words calculated by the gradients [24]. To evaluate the agreement, we compared the Pearson's correlations between the attention weights and the word importance of the gradient-based method. In [36], the Kendall tau, which represents rank correlation, was used to evaluate the relationship between attention and the word importance obtained by the gradients. Recently, however, it has been pointed out that rank correlations often misrepresent the relationship between the two due to the noise in the order of the low rankings [102]; we concurred with this, so we used Pearson's correlations. Refer to Appendix B.2.1 for details on the evaluation criteria.

Finally, we compared the effects of perturbation size ϵ of AT on the validation performance of the BC, QA, and NLI tasks with a fixed $\lambda = 1$. We randomly choose the value of ϵ in the 0–30 range and ran the training 100 times. The configurations in [75, 76] were $\epsilon = 5$ for Word AT and $\epsilon = 15$ for Word iAT.

[6] The F1 score is a metric that harmonizes precision and recall.. Therefore, this score takes both false positives and false negatives into account.

2.5 Results

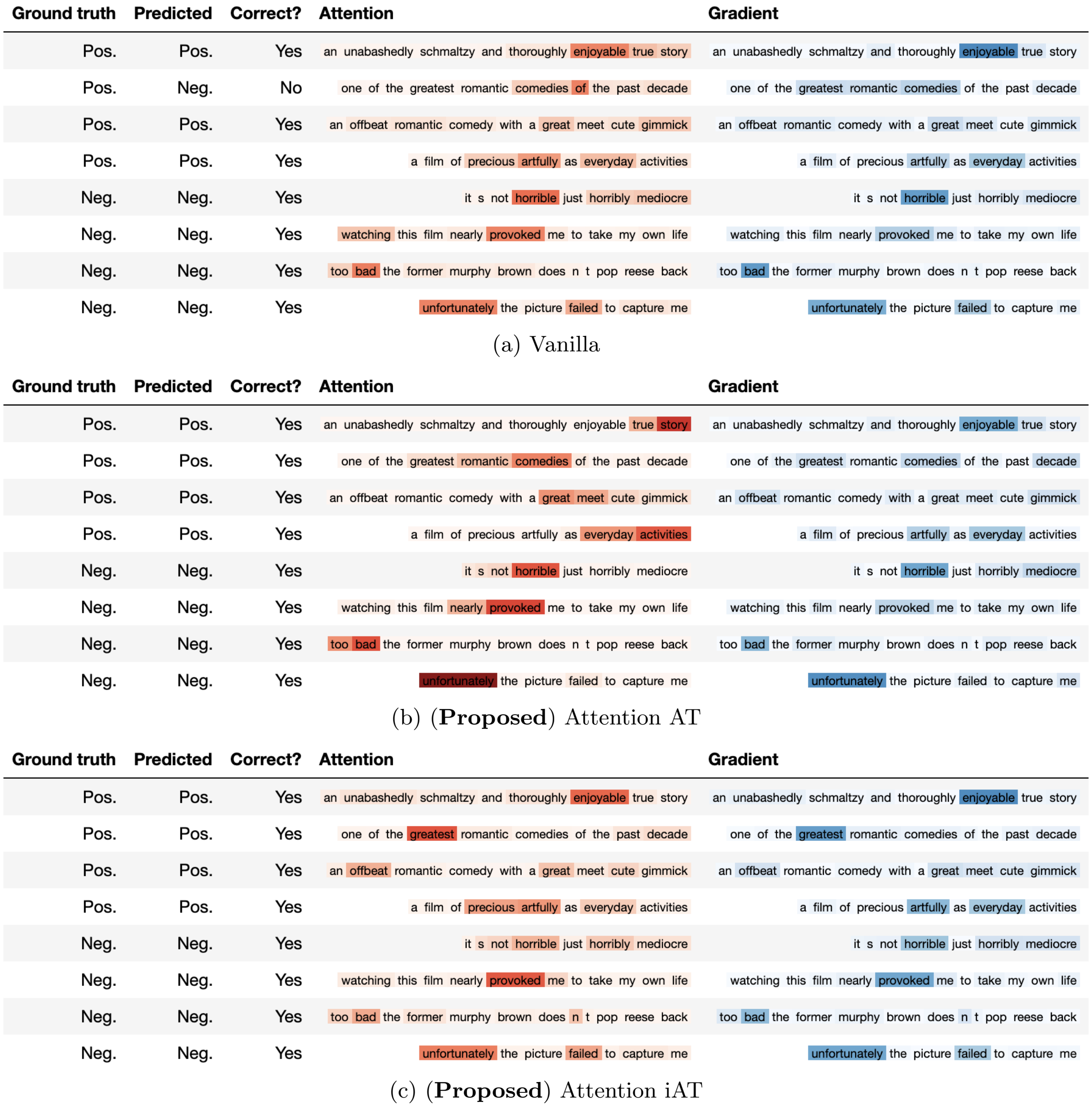
In this section, we share the results of the experiments. Table 2.2 presents the prediction performance and the Pearson's correlations between the attention weight for the words and word importance calculated from the model gradient. The most significant results are shown in bold.

2.5.1 Comparison of Prediction Performance

In terms of prediction performance, the model that applied the proposed Attention AT/iAT demonstrated a clear advantage over the model without AT (as shown in Vanilla [36]) as well as other AT-based techniques (Word AT [75] and Word iAT [76]). The proposed technique achieved the best results in almost all benchmarks. For 20News and AGNews in the BC and bAbI task 1 in QA, the conventional techniques, including the Vanilla model, were sufficiently accurate (the score was higher than 95%), so the performance improvement of the proposed techniques to the tasks was limited to some extent. Meanwhile, Attention AT/iAT contributed to solid performance improvements in other complicated tasks.

2.5.2 Comparison of Correlation between Attention Weights and Gradients on Word Importance

In terms of model interpretability, the attention to the words obtained with the Attention AT/iAT techniques notably correlated with the importance of the word as determined by the gradients. Attention iAT demonstrated the highest correlation among the techniques in all benchmarks. Figure 2.2 visualizes the attention weight for each word and gradient-based word importance in the SST test dataset. Attention AT yielded clearer attention compared to the Vanilla model or Attention iAT. Specifically, Attention AT tended to strongly focus attention on a few words. Regarding the correlation of word importance based on attention weights and gradient-based word importance, Attention iAT demonstrated higher similarities than the other models.



Ground truth	Predicted	Correct?	Attention	Gradient
Pos.	Pos.	Yes	an unabashedly schmaltzy and thoroughly enjoyable true story	an unabashedly schmaltzy and thoroughly enjoyable true story
Pos.	Neg.	No	one of the greatest romantic comedies of the past decade	one of the greatest romantic comedies of the past decade
Pos.	Pos.	Yes	an offbeat romantic comedy with a great meet cute gimmick	an offbeat romantic comedy with a great meet cute gimmick
Pos.	Pos.	Yes	a film of precious artfully as everyday activities	a film of precious artfully as everyday activities
Neg.	Neg.	Yes	it s not horrible just horribly mediocre	it s not horrible just horribly mediocre
Neg.	Neg.	Yes	watching this film nearly provoked me to take my own life	watching this film nearly provoked me to take my own life
Neg.	Neg.	Yes	too bad the former murphy brown does n t pop reese back	too bad the former murphy brown does n t pop reese back
Neg.	Neg.	Yes	unfortunately the picture failed to capture me	unfortunately the picture failed to capture me

(a) Vanilla

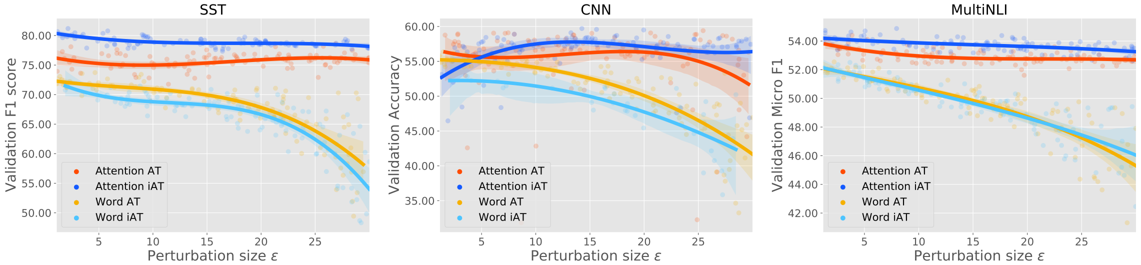
Ground truth	Predicted	Correct?	Attention	Gradient
Pos.	Pos.	Yes	an unabashedly schmaltzy and thoroughly enjoyable true story	an unabashedly schmaltzy and thoroughly enjoyable true story
Pos.	Pos.	Yes	one of the greatest romantic comedies of the past decade	one of the greatest romantic comedies of the past decade
Pos.	Pos.	Yes	an offbeat romantic comedy with a great meet cute gimmick	an offbeat romantic comedy with a great meet cute gimmick
Pos.	Pos.	Yes	a film of precious artfully as everyday activities	a film of precious artfully as everyday activities
Neg.	Neg.	Yes	it s not horrible just horribly mediocre	it s not horrible just horribly mediocre
Neg.	Neg.	Yes	watching this film nearly provoked me to take my own life	watching this film nearly provoked me to take my own life
Neg.	Neg.	Yes	too bad the former murphy brown does n t pop reese back	too bad the former murphy brown does n t pop reese back
Neg.	Neg.	Yes	unfortunately the picture failed to capture me	unfortunately the picture failed to capture me

(b) (**Proposed**) Attention AT

Ground truth	Predicted	Correct?	Attention	Gradient
Pos.	Pos.	Yes	an unabashedly schmaltzy and thoroughly enjoyable true story	an unabashedly schmaltzy and thoroughly enjoyable true story
Pos.	Pos.	Yes	one of the greatest romantic comedies of the past decade	one of the greatest romantic comedies of the past decade
Pos.	Pos.	Yes	an offbeat romantic comedy with a great meet cute gimmick	an offbeat romantic comedy with a great meet cute gimmick
Pos.	Pos.	Yes	a film of precious artfully as everyday activities	a film of precious artfully as everyday activities
Neg.	Neg.	Yes	it s not horrible just horribly mediocre	it s not horrible just horribly mediocre
Neg.	Neg.	Yes	watching this film nearly provoked me to take my own life	watching this film nearly provoked me to take my own life
Neg.	Neg.	Yes	too bad the former murphy brown does n t pop reese back	too bad the former murphy brown does n t pop reese back
Neg.	Neg.	Yes	unfortunately the picture failed to capture me	unfortunately the picture failed to capture me

(c) (**Proposed**) Attention iAT

Figure 2.2: Visualization of attention weight for each word and word importance calculated by gradients on the SST dataset. Best viewed in color. The models that apply the proposed Attention AT/iAT give clearer attention. In terms of word importance, Attention iAT has more similar attention-based and gradient-based results than the other methods.



(a) SST (BC) (b) CNN news (QA) (c) Multi NLI (NLI)

Figure 2.3: The effect of perturbation size ϵ on the validation performance. We observed that the model with Attention AT/iAT maintained an almost constant prediction performance even when the perturbation size increased. Compared to Word AT/iAT, The model with the proposed techniques, Attention AT/iAT, was more robust with a large ϵ.

2.5.3 Effects of Perturbation Size

Figure 2.3 shows the effect of the perturbation size ϵ on the validation performance of SST (BC), CNN news (QA), and Multi NLI (NLI) with a fixed $\lambda = 1$. We observed that the performances of the conventional Word AT/iAT techniques deteriorated according to the increase in the perturbation size; meanwhile, our Attention AT/iAT techniques maintained almost the same prediction performance. We observed similar trends in other datasets as described in Section 2.4.1.

2.6 Discussion

2.6.1 Comparison of Adversarial Training for Attention Mechanisms and Word Embedding

Attention AT/iAT is based on our hypothesis that attention is more important in finding significant words in document processing than the word embeddings themselves. Therefore, we sought to achieve prediction performance and model interpretability by introducing AT to the attention mechanisms. We confirmed that the application of AT to the attention mechanisms (Attention AT/iAT) was more effective than word embedding (Word AT/iAT) and supports the correctness of our hypothesis, as shown in Table 2.2. In particular, the Attention iAT technique was not only more accurate in its model than the Word AT/iAT techniques but also demonstrated a higher correlation with the importance of the words predicted based on the gradient.

As shown in Figure 2.2, Attention AT tended to display more attention to the sentence than the Vanilla model. The results showed that training with adversarial perturbations to the attention mechanism allowed for cleaner attention without changing word meanings or grammatical functions. Furthermore, we confirmed that the proposed Attention AT/iAT techniques were more robust regarding the variation of perturbation size ϵ than conventional Word AT/iAT, as shown in Figure 2.3. Although it is difficult to directly compare perturbations to attention and word embedding because of the difference in the range of the perturbation size to the part, the model that added perturbations to attention behaved robustly even when the perturbations were relatively large.

2.6.2 Comparison of Random Perturbations and Adversarial Perturbations

Attention RP demonstrated better prediction performance than Word AT/iAT. The results revealed that augmentation for the attention mechanism is very effective, even with simple random noise. In contrast, the correlations between the attention weight for the word and the gradient-based word importance were significantly reduced, as shown in Table 2.2. We consider that Attention RP is successful in learning robust discriminative boundaries through random perturbation and improving to the desired classification performance. However, as the gradient is smoothed out by the perturbation around the (supervised) data points, the correlation with the word importance by the gradient is considered to be degraded. In other words, Attention RP can achieve a certain level of classification performance, but it does not lead to which words are useful from their gradients.

2.6.3 Comparison of Attention AT, Attention iAT, and Word iAT

In the experiments, Attention iAT showed a better performance compared Attention AT in the prediction performance and the correlation with the gradient-based word importance. Attention iAT exploits the difference in the attention weight of each word in a sentence to determine adversarial perturbations. Because the norm of the difference in attention weight (as shown in Eq. 3.8) is normalized to one, adversarial perturbations in attention mechanisms will make these differences clear, especially in the case of sentences with a small difference in the attention to each word. That is, even in situations where there is little difference in attention between each element of $\boldsymbol{d}_t = (d_{t,1}, d_{t,2}, \cdots, d_{t,T})$, the difference is amplified by Eq. 3.8. Therefore, even for the same perturbation size $||\boldsymbol{\alpha}|| < \epsilon$, more effective perturbations $\boldsymbol{r}(\boldsymbol{\alpha})$ weighted by $\tilde{\boldsymbol{d}}_t$ were successfully obtained for each word. However, in the case of sentences where there was originally a difference in the clear attention to each word, the regularization of $\boldsymbol{d}_t$ in Eq. 3.8 had practically no effect because it did not change their ratio nearly as much. Thus, we posit that the Attention iAT technique enhances the effectiveness of AT applied to attention mechanisms by generating effective perturbations for each word.

The Attention iAT technique was inspired by Word iAT. Word iAT generates perturbations in the direction that maximizes the loss function while restricting the direction of the perturbation to become a linear combination of the direction of word embedding in the vocabulary. Word iAT indirectly improves the interpretability of the model by indicating which words in the vocabulary to which the perturbation is similar. However, we are confident that Attention iAT is a direct improvement in the interpretability of the model, because it can show more clearly which words to pay attention. This is owing to the attention difference enhancement process described in Eq. 3.8. Thus, the proposed techniques are highly effective in that they lead to a more substantive improvement in interpretability.

2.6.4 Limitations

Our proposal is a general-purpose robust training technique for DNN models, which are commonly used for NLP tasks. Therefore, we have chosen here an RNN with an attention mechanism that has been put to practical use [36]. For this reason, models such as BERT [7] that deal with self-attention were outside the scope of this study, and will be the subject of future work. We also did not deal with tasks (such as machine translation) that were not used in the literature [36] as baselines. Additionally, for the same reason in as [39], we did not consider the variants of attention mechanisms, such as bi-attentive architecture [66], multi-headed architecture [6], because they could have different interpretability properties.

As an extension of AT, virtual adversarial training (VAT), which is a semi-supervised training technique, was proposed in [49, 75]. Based on VAT, the proposed technique can be expected to improve accuracy by using unlabeled datasets.

2.7 Conclusion

We proposed robust and interpretable attention training techniques that exploit AT. In the experiments with various NLP tasks, we confirmed that AT for attention mechanisms achieves better performance than techniques using AT for word embedding in terms of the prediction performance and the interpretability of the model. Specifically, the Attention iAT technique introduced adversarial perturbations that emphasized differences in the importance of words in a sentence and combined high accuracy with interpretable attention, which was more strongly correlated with the gradient-based method of word importance. The proposed technique could be applied to various models and NLP tasks. This paper provides strong support and motivation for utilizing AT with attention mechanisms in NLP tasks. We describe the extension to semi-supervised learning that further increases the performance in the proposed technique in Chapter 3.

In the experiment, we demonstrated the effectiveness of the proposed techniques for RNN models that are reported to be vulnerable to attention mechanisms, but we will confirm the effectiveness of the proposed technique for large language models with attention mechanisms such as Transformer [6] or BERT [7] in the future. Because the proposed techniques are model-independent and general techniques for attention mechanisms, we can expect they will improve predictability and the interpretability for language models. The applicability to current mainstream models with attention mechanisms is discussed in Chapter 6.

Table 2.2: Comparison of prediction performance and the Pearson's correlation coefficients (Corr.) between the attention weight and the word importance of the gradient-based method. We used the same test metrics as in [36]: binary classification (BC), question answering (QA), and natural language inference (NLI). As an evaluation metrics of the prediction performance, we used the F1 score (F1), accuracy (Acc.), and the micro-F1 (Micro-F1) score for BC, QA, and NLI, respectively.

(a) Binary classification (BC)

Model	SST		IMDB		20News		AGNews	
	F1 [%]	Corr.	F1 [%]	Corr.	F1 [%]	Corr.	F1 [%]	Corr.
Vanilla [36]	79.27	0.652	88.77	0.788	95.05	0.891	95.27	0.822
Word AT [75]	79.61	0.647	89.65	0.838	95.11	0.892	95.59	0.813
Word iAT [76]	79.57	0.643	89.64	0.839	95.14	0.893	95.62	0.809
Attention RP	81.90	0.531	89.79	0.628	96.09	0.883	96.08	0.792
Attention AT (**Proposed**)	81.72	0.852	90.00	0.819	**96.69**	0.868	96.12	0.835
Attention iAT (**Proposed**)	**82.20**	**0.876**	**90.21**	**0.861**	96.58	**0.897**	**96.19**	**0.891**

(b) Question answering (QA)

Model	CNN news		bAbI					
	Acc. [%]	Corr.	Task 1		Task 2		Task 3	
			Acc. [%]	Corr.	Acc. [%]	Corr.	Acc. [%]	Corr.
Vanilla [36]	64.95	0.765	99.90	0.714	45.10	0.459	52.00	0.387
Word AT [75]	65.67	0.779	**100.00**	0.797	79.50	0.657	55.10	0.439
Word iAT [76]	65.66	0.776	99.90	0.798	79.80	0.658	54.90	0.437
Attention RP	65.78	0.614	**100.00**	0.592	80.60	0.584	55.35	0.373
Attention AT (**Proposed**)	65.93	0.771	**100.00**	0.807	82.30	0.632	56.00	0.514
Attention iAT (**Proposed**)	**66.17**	**0.784**	**100.00**	**0.821**	**85.40**	**0.710**	**57.10**	**0.589**

(c) Natural language inference (NLI)

Model	SNLI		Multi NLI			
	Micro-F1 [%]	Corr.	Micro-F1 [%]			Corr.
			Avg.	Matched	Mismatched	
Vanilla [36]	78.64	0.764	60.26	59.80	60.71	0.541
Word AT [75]	79.03	0.812	60.72	60.58	60.86	0.601
Word iAT [76]	79.12	0.815	60.73	60.59	60.87	0.603
Attention RP	79.23	0.569	60.97	61.02	60.91	0.547
Attention AT (**Proposed**)	79.19	0.792	61.17	61.20	**61.13**	0.626
Attention iAT (**Proposed**)	**79.32**	**0.818**	**61.34**	**61.75**	60.93	**0.668**

Chapter 3

Virtual Adversarial Training for Attention Mechanisms

In chapter 2, we proposed new techniques of adversarial training for attention mechanism. This chapter 3 aims to extend the technique to semi-supervised learning. We described the background in Section 3.1 that our supervised technique can be extended to semi-supervised learning in the framework of virtual adversarial training, which can expect to improve further the prediction performance and interpretability of deep learning models. We review the attention mechanism, adversarial training, and their relation to our proposal with Attention AT/iAT in Section 3.2. After that, we propose Attention VAT/iVAT, extensions of Attention AT/iAT based on the idea of virtual adversarial training in Section 3.3. We demonstrate in Section 3.4 and 3.5 that the proposed techniques outperform the conventional adversarial and virtual adversarial training methods for word embeddings, as well as the techniques proposed in Chapter 2, in terms of prediction performance and interpretability of the deep learning models. We discuss in Section 3.6 in terms of the comparison between supervised and semi-supervised learning based on the training techniques for attention mechanism, in terms of the amount of unlabeled data, and in terms of the agreement between human-annotated evidence labels and the prediction evidence presented by the proposed technique. Finally, we summarize this chapter in Section 3.7.

3.1 Background

Despite their significant success in various computer vision (CV) and natural language processing (NLP) tasks, deep neural networks (DNNs) are usually vulnerable to tiny input perturbations, which are often referred to as adversarial examples [46, 47, 103]. Goodfellow et al. [46] proposed adversarial training (AT) for protection against malicious perturbations, which improves robustness by smoothing the discrimination boundary through adding adversarial perturbations to the input. This technique has been reported as highly effective for solving common problems in the CV field. In NLP, Miyato et al. [75] proposed AT for word embeddings (Word AT) to improve model robustness. Moreover, Sato et al. [76] developed interpretable AT for word embeddings (Word iAT), which makes adversarial perturbations more interpretable by restricting the perturbation direction to existing words in the word embedding space.

At present, attention mechanisms, on which various studies have focused in recent years [6, 7, 33, 64], are among the most important components in DNN models. However, Jain and Wallace [36] reported a critical problem; attention mechanisms are vulnerable to malicious perturbations. To mitigate this issue, we recently proposed AT for attention mechanisms (Attention AT) and interpretable AT for attention mechanisms (Attention iAT) [104]. These proposals provide model robustness to the vulnerability of attention mechanisms, as well as a more interpretable attention heatmap; that is, our techniques provide clear attention in a sentence. Moreover, these techniques significantly improve performance.

Although AT successfully improves model robustness, the training technique is only available in supervised settings because labeled data are needed to calculate adversarial perturbation. However, in NLP, labeled data for training are often expensive, and it is generally not easy to apply data augmentation. Therefore, extending these techniques to semi-supervised learning, whereby unlabeled data can also be used for training, is desirable. Virtual AT (VAT) [49], which is a sophisticated semi-supervised learning technique, has been proposed to extend AT to semi-supervised settings. This technique improves model robustness by smoothing out discrimination boundaries by calculating the adversarial perturbations, which is known as "virtual adversarial perturbation," based on the current prediction using both labeled and unlabeled data. VAT is reportedly effective in both the CV [46, 49] and NLP [75, 76, 105] fields. Semi-supervised learning methods generally use unlabeled data based on the model's output (i.e., current parameters) at the time; therefore, to a certain extent, the unlabeled data should be similar to the labeled data.

In this study, we propose two new general training techniques: *VAT for attention mechanisms* (Attention VAT) and *interpretable VAT for attention mechanisms* (Attention iVAT), which extend AT for attention mechanisms to a semi-supervised setting, thereby making the model even more robust and interpretable. Attention VAT and iVAT improve model robustness by smoothing out the discriminative boundary even for unlabeled data points by calculating the direction of the virtual adversarial perturbation. In contrast to Attention VAT, Attention iVAT utilizes the difference in attention to each word in a sentence to construct adversarial perturbations, and can thus learn clear attention more efficiently, which improves the model interpretability.

We compared our proposed techniques with several other state-of-the-art AT-based techniques using six common NLP benchmarks. We evaluated the extent to which the attention weights that were obtained using our techniques agreed with the gradient-based word importance [24] and rationales provided by humans [106]. We investigated the effect on the performance of the unlabeled data being added to the training. This empirically confirmed that our techniques could improve both the prediction performance and model interpretability, even when trained on unlabeled data from a source different from that of the target task.

The advantages of our VAT techniques for attention mechanisms are summarized as follows:

- The VAT for attention mechanisms is an extension to semi-supervised learning of the effective AT for attention mechanisms, which provides improved prediction performance compared to conventional AT-based techniques and recent VAT-based techniques in semi-supervised settings.

- The VAT for attention mechanisms also favors model interpretability, exhibiting a stronger correlation with word importance based on gradients and higher agreement with the results of human annotations. This promising outcome has been particularly important for deep learning models in recent years.

- The benefit of the above extension to semi-supervised learning is almost independent of selecting the additional unlabeled data to be used. That is, a significant performance improvement can be expected even while reducing the difficulty of selecting unlabeled data. Our empirical experiments confirm that prediction performance improves by increasing the unlabeled data.

3.2 Related Work

In this section, we describe two topics that are closely related to our VAT for attention mechanisms: (1) AT and its extension, namely VAT, and (2) attention mechanisms.

3.2.1 AT and its Extension, VAT

AT [46, 47, 83, 107] is a powerful regularization technique in the CV field primarily explored to improve model's robustness to input perturbations. A recent proposal in the CV field, attention transfer based adversarial training (ATAT) [107], considers the class activation map (CAM) as an attention heatmap and applies adversarial training to the CAM. Furthermore, AT has been applied to various NLP tasks by extending the concept of adversarial perturbations, such as text classification [75, 76, 108], part-of-speech tagging [77], relation extraction [109], and machine reading comprehension [41]. The recently proposed stability fine-tuning framework (StaFF) [108] applies AT to defend against word-level adversarial attacks. Word AT [75] is a pioneering technique that applies AT for word embeddings in NLP. Word iAT [76] has subsequently been proposed to realize more "interpretable" AT for word embeddings. This technique restricts the perturbation direction to existing words in the word embeddings space. Each input with a perturbation can be interpreted as an actual sentence by considering the perturbation as a replacement for a word in the sentence. However, although this approach enhances the interpretability of the perturbations, it does not increase the interpretability of model prediction.

VAT [49] extends AT to semi-supervised settings, in which *virtual* adversarial perturbation, even for unlabeled data points, is defined. This technique provides smoother discriminative boundaries and improves model robustness compared to AT, in which only supervised data are used. VAT has achieved state-of-the-art performance in the CV field, e.g., in image classification tasks [49], as well as excellent results in the NLP fields, e.g., in text classification [75, 76] and sequence tagging [105]. In particular, Chen et al. [105] used an extremely large dataset, namely the One Billion Word Language Model Benchmark dataset [110], as an unlabeled data pool for semi-supervised learning. Their VAT-based technique outperformed state-of-the-art sequence labeling models. With a small number of labeled examples, An et al. [111] recently proposed a VAT-based semi-supervised question generation (QG) model, named virtual stroke recognition copy network (VSAC Net), for the Chinese QG tasks. These models that use VAT for NLP tasks are applied to input or hidden representations, which is different from our proposed concept of applying VAT to attention mechanisms.

Whereas previous AT and VAT techniques have been applied to the input space mainly for word embeddings, we proposed the Attention AT and Attention iAT techniques for the application of AT to attention mechanisms [104]. These techniques outperformed Word AT and Word iAT in various tasks, such as text classification, question answering (QA), and natural language inference (NLI) tasks. The results indicated that AT for attention mechanisms is more efficient than AT for word embeddings regarding both model accuracy and interpretability. Our proposed Attention VAT and Attention iVAT, are semi-supervised learning extensions of Attention AT and Attention iAT, respectively.

Table 3.1 summarizes the comparison between our Attention VAT/iVAT and those in related studies. Both of our proposals, Attention VAT and Attention iVAT, remedy the potential vulnerability to perturbations in the attention mechanisms and can handle unlabeled data. Moreover, Attention iVAT efficiently learns clearer attention and improves model interpretability without the selection of unlabeled data. VAT has often been employed for inputs and their embeddings [49, 76]; however, to our knowledge, no application to attention mechanisms has been reported. We have

Table 3.1: Comparison of the proposed Attention VAT/iVAT with that of related studies.

	Attention vulnerability	Unlabeled data	Model interpretability
Word AT [75]	✗	✗	✗
Word iAT [76]	✗	✗	✓
Word VAT [49]	✗	✓	✗
Word iVAT [76]	✗	✓	✓
Attention AT [104]	✓	✗	✗
Attention iAT [104]	✓	✗	✓
Attention VAT (ours)	✓	✓	✗
Attention iVAT (ours)	✓	✓	✓

empirically demonstrated that our VAT-based techniques are effective even when they are applied to unlabeled data from sources other than the labeled data.

3.2.2 Attention Mechanisms

Attention mechanisms for machine translation were first introduced by Bahdanau et al. [33]. These mechanisms contribute to improving the prediction performance of various NLP tasks, such as sentence-level classification [64], sentiment analysis [41, 112], QA [65], and NLI [66]. The recently proposed sentence-level attention transfer network (SentATN) [112] aims to solve the cross-domain sentence classification task with the advantages of the attention mechanism and sentence-level AT.

Attention weights are claimed to offer insights into the inner workings of DNNs [79]. However, learned attention weights are often uncorrelated with word importance, which is calculated using the gradient-based method [24], and perturbations to the attention mechanisms may interfere with the interpretation [36,113]. Our recently proposed AT for attention mechanisms [104], is a novel and practical training technique for solving these issues. We have demonstrated that this technique (1) improves the performance, (2) exhibits a stronger correlation with gradient-based word importance, and (3) is substantially less dependent on the scale of perturbation.

This study presents a new attempt to employ VAT in attention mechanisms. Our proposal addresses the potential problem with attention mechanisms, namely the low correlation between

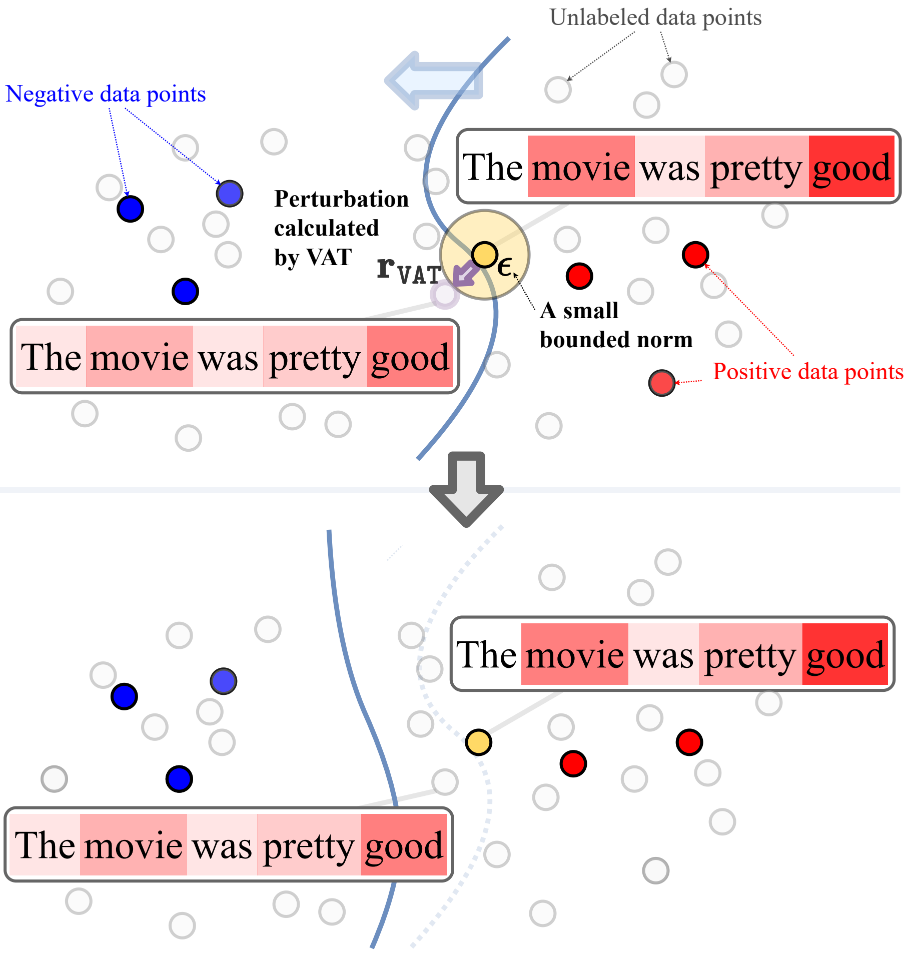

Unlabeled data points
Negative data points
The movie was pretty good
Perturbation calculated by VAT
$\mathbf{r}_{\mathtt{VAT}}$
ϵ
A small bounded norm
Positive data points
The movie was pretty good
The movie was pretty good
The movie was pretty good

Figure 3.1: Intuitive illustration of the proposed VAT for attention mechanisms. Our technique can learn clearer attention by overcoming adversarial perturbations $\boldsymbol{r}_{\mathtt{VAT}}$, thereby improving model interpretability

the attention weight and word importance. It is expected to increase the correlation (i.e., model interpretability) further by effectively using unlabeled data and learning to focus on the differences in the attention weights.

3.3 Virtual Adversarial Training for Attention Mechanism

Our proposal is an extension of training techniques that are effective by introducing adversarial perturbation to the attention mechanism to semi-supervised learning using VAT [49]. In this section, we introduce two new general training techniques for attention mechanisms based on VAT: (1) VAT for attention mechanisms (Attention VAT), and (2) interpretable VAT for attention mechanisms (Attention iVAT). These techniques follow the calculation of the perturbation direction in the original VAT, and thus, the formulas are similar although the implications are different, as we will explain in the following sections.

Fig. 3.1 provides an intuitive illustration of the proposed technique. Our proposals provide general-purpose robust training techniques that can be applied to any model that uses attention mechanisms and can also use unlabeled data. We apply the proposed techniques to the common model that is described in Appendix A.1.

Let $X_{\tilde{\boldsymbol{a}}}$ be an input sequence with an attention score $\tilde{\boldsymbol{a}}$. We model the conditional probability of class $\boldsymbol{y}$ as $p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents all model parameters. We minimize the following negative log-likelihood as a loss function for the model parameters to train the model:

$$\mathcal{L}(X_{\tilde{\boldsymbol{a}}},\boldsymbol{y};\boldsymbol{\theta}) = -\log p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}};\boldsymbol{\theta}). \tag{3.1}$$

Our proposed techniques introduce an adversarial perturbation to the attention score $\tilde{\boldsymbol{a}}$.

3.3.1 Attention VAT: Virtual Adversarial Training for Attention

The first proposal, Attention VAT is a natural extension of our previous Attention AT [104] for supporting semi-supervised learning. Whereas AT determines the adversarial perturbation in supervised setting based on label information, VAT computes the *virtual* adversarial perturbation even for unlabeled data, and is expected to provide more robust models. The virtual adversarial perturbation on the attention mechanisms is defined as the worst-case perturbation on the mechanisms of a small, bounded norm by maximizing the Kullback–Leibler (KL) divergence between the estimated label distribution of the original examples and the distribution of the perturbation that is added to the perturbation of the original example.

Suppose that $\mathcal{D}'$ denotes a set of N_l labeled and N_u unlabeled data. We introduce $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$, which denotes $X_{\tilde{\boldsymbol{a}}}$ with the additional perturbation $\boldsymbol{r}$. Attention VAT minimizes the following loss function for estimating the loss of the virtual adversarial perturbation $\boldsymbol{r}_{\mathtt{VAT}}$ on the input sequence with an attention score:

$$\mathcal{L}_{\mathtt{VAT}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathtt{VAT}}}; \hat{\boldsymbol{\theta}}) = \frac{1}{|\mathcal{D}'|} \sum_{X \in \mathcal{D}'} \mathcal{L}_{\mathtt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathtt{VAT}}}; \hat{\boldsymbol{\theta}}), \tag{3.2}$$

$$\boldsymbol{r}_{\mathtt{VAT}} = \underset{\boldsymbol{r}:\|\boldsymbol{r}\|_2 \leq \epsilon}{\operatorname{argmax}} \, \mathcal{L}_{\mathtt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}; \hat{\boldsymbol{\theta}}), \tag{3.3}$$

where ϵ is a hyperparameter that controls the norm of the perturbation and is determined using the validation set. $\hat{\boldsymbol{\theta}}$ represents the current model parameters. In this case, $\mathcal{L}_{\mathtt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}; \hat{\boldsymbol{\theta}})$ is the KL divergence, which can be calculated using $\mathrm{KL}(\cdot\|\cdot)$:

$$\mathcal{L}_{\mathtt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathtt{VAT}}}; \hat{\boldsymbol{\theta}}) = \mathrm{KL}(p(\cdot|X_{\tilde{\boldsymbol{a}}}, \hat{\boldsymbol{\theta}}) \parallel p(\cdot|X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathtt{VAT}}}, \hat{\boldsymbol{\theta}})). \tag{3.4}$$

VAT uses the current model parameters $\hat{\boldsymbol{\theta}}$ to find the worst perturbation, $\boldsymbol{r}_{\mathtt{VAT}}$, that most affects the output probability and learns, such that the prediction does not change when $\boldsymbol{r}_{\mathtt{VAT}}$ is added to each sample. Differently expressed, the effect is to make the model robust by smoothing the probability distribution of predictions for each data point, including the unlabeled data. We use an approximation method instead of solving the expensive optimization problem above [75]:

$$\boldsymbol{r}_{\mathtt{VAT}} = \epsilon \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \text{where } \boldsymbol{g} = \nabla_{\tilde{\boldsymbol{a}}} \mathcal{L}_{\mathtt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}; \hat{\boldsymbol{\theta}}). \tag{3.5}$$

In each training step, we obtain a virtual adversarial perturbation $\boldsymbol{r}_{\mathtt{VAT}}$ for the attention score $\tilde{\boldsymbol{a}}$ of each input data based on the current model $\hat{\boldsymbol{\theta}}$:

$$\tilde{\boldsymbol{a}}_{\mathtt{vadv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\mathtt{VAT}}. \tag{3.6}$$

3.3.2 Attention iVAT: Interpretable Virtual Adversarial Training for Attention

The other proposal, Attention iVAT, is also an extension of our previous Attention iAT [104] for semi-supervised learning. Our former Attention iAT generates adversarial perturbations by focusing on the difference in attention to each word, and thus, generates clearer differences in attention among words. That is, this technique significantly enhances model interpretability by providing an intuitive and clear indication of the words in a sentence that are important.

Another technique relating to our proposal, namely Word iVAT, extends interpretable AT for the word embedding of each word to semi-supervised learning. This technique may improve the interpretability of perturbations by restricting the direction of the added perturbation to the known direction of the word. Although the name of this technique is similar to that of our proposal, it does not provide explicit information regarding the important words in a sentence, and thus does not provide the high interpretability of the model as the name implies.

We formulate our Attention iVAT as follows. We define the word attentional difference vector $\boldsymbol{d}_t \in \mathbb{R}^T$, where T is the number of words in the sequence, as the attentional difference between the t-th word $\tilde{a}_t$ and any k-th word $\tilde{a}_k$ in a sentence:

$$\boldsymbol{d}_t = (d_{t,k})_{k=1}^T = (\tilde{a}_t - \tilde{a}_k)_{k=1}^T. \tag{3.7}$$

Note that the difference in the attention score to itself is 0; that is, $d_{t,t} = 0$. By normalizing the norm of the vector, we define a normalized word attentional difference vector of the attention for the t-th word, as follows:

$$\tilde{\boldsymbol{d}}_t = \frac{\boldsymbol{d}_t}{\|\boldsymbol{d}_t\|_2}. \tag{3.8}$$

We define the perturbation $r(\boldsymbol{w}_t)$ for the attention to the t-th word with the trainable parameters $\boldsymbol{w}_t = (w_{t,k})_{t=1}^T \in \mathbb{R}^T$ as follows:

$$r(\boldsymbol{w}_t) = \boldsymbol{w}_t^\top \tilde{\boldsymbol{d}}_t \tag{3.9}$$

By combining $\boldsymbol{w}_t$ for all t as $\boldsymbol{w}$, we can calculate the perturbation $\boldsymbol{r}(\boldsymbol{w})$ for the sentence:

$$\boldsymbol{r}(\boldsymbol{w}) = (r(\boldsymbol{w}_t))_{t=1}^{T}. \tag{3.10}$$

Subsequently, similar to $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$ in Eq. (3.3), we introduce $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{w})}$ and seek the worst-case direction of the difference vectors that maximize the loss function:

$$\boldsymbol{r}_{\texttt{iVAT}} = \underset{\boldsymbol{r}:\|\boldsymbol{r}\|_2 \leq \epsilon}{\operatorname{argmax}} \, \mathcal{L}_{\texttt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{w})}; \hat{\boldsymbol{\theta}}). \tag{3.11}$$

Using the same derivation method as in Eq. (3.5) to obtain the above approximation, we introduce the following equation to calculate $\boldsymbol{r}_{\texttt{iVAT}}$ as an extension to semi-supervised learning:

$$\boldsymbol{r}_{\texttt{iVAT}} = \epsilon \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \quad \boldsymbol{g} = \nabla_{\boldsymbol{w}} \mathcal{L}_{\texttt{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{w})}, \hat{\boldsymbol{\theta}}). \tag{3.12}$$

Similar to Eq. (3.6), we construct a perturbated example for the attention score $\tilde{\boldsymbol{a}}$:

$$\tilde{\boldsymbol{a}}_{\texttt{ivadv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\texttt{iVAT}}. \tag{3.13}$$

Unlike Attention VAT, Attention iVAT defines the perturbation $\boldsymbol{r}_{\texttt{iVAT}}$ as the product of a trainable parameter $\boldsymbol{w}_t$ and a normalized word attentional vector $\boldsymbol{d}_t$. In situations in which the difference in the attention among words is relatively small (i.e., the model does not know which words in the sentence are important), Eq. (3.8) enables the computation of meaningful virtual adversarial perturbations that increase the difference in the attention between words. Consequently, the model is expected to obtain clearer attention to overcome the perturbations. This process can be considered as attention difference enhancement and was described in our previous paper [104].

We describe the attention difference enhancement in Attention iVAT in terms of its formula. As the norm of the difference in the attention weight (as indicated in Eq. (3.8)) is normalized to one, virtual adversarial perturbations for attention mechanisms can clarify these differences, particularly in the case of sentences with a small difference in the attention to each word. Even in situations in which there is little difference in the attention between each element of $\boldsymbol{d}_t = (d_{t,1}, d_{t,2}, \cdots, d_{t,T})$, the difference is emphasized by Eq. (3.8). In the case of sentences in which there is originally a difference in the clear attention to each word, the regularization effect of $\hat{\boldsymbol{d}}_t$ in Eq. (3.8) has almost no effect because it hardly changes the ratio. As in our previous study [104], the Attention iVAT technique enhances the usefulness of VAT when it is employed in attention mechanisms by generating effective perturbations for each word. A notable advantage of our Attention iVAT is that this attention difference enhancement can be computed without label information.

3.3.3 Model Training with VAT

We generate a virtual adversarial perturbation based on the current model $\hat{\boldsymbol{\theta}}$ at each training step. To this end, we define the loss function for the VAT as follows:

$$\tilde{\mathcal{L}} = \underbrace{\mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \hat{\boldsymbol{\theta}})}_{\substack{\text{Loss from}\\ \text{unmodified examples}}} + \underbrace{\lambda \mathcal{L}_{\texttt{VAT}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}_{\texttt{VADV}}}; \hat{\boldsymbol{\theta}})}_{\substack{\text{Loss from}\\ \text{virtual adversarial examples}}} , \tag{3.14}$$

where λ is the coefficient that is used to control the two-loss functions. Note here that $X_{\tilde{\boldsymbol{a}}_{\texttt{VADV}}}$ may $X_{\tilde{\boldsymbol{a}}_{\texttt{vadv}}}$ for Attention VAT or $X_{\tilde{\boldsymbol{a}}_{\texttt{ivadv}}}$ for Attention iVAT.

3.4 Experiments

In this section, we describe the comparison models, datasets, evaluation tasks, and evaluation criteria that were used in the experiments.

3.4.1 Comparison Models

We compared Attention VAT and Attention iVAT, with conventional AT- and VAT-based training techniques. Following previous studies [36, 104, 114], we implemented the techniques using the same recurrent neural network (RNN)-based model architecture described in Appendix A.1. The proposed techniques effectively extend semi-supervised learning, a training technique that achieves high performance and generalizability by adding adversarial perturbation to the attention mechanism of the model. Therefore, we conducted comparative experiments that employed the widely used RNN-based model configurations as a baseline and evaluated them by how much performance improved. Our proposed technique is a general-purpose robust training method that can be applied to all models. Therefore, in this study, instead of comparing our techniques with models that aim to improve the individual performance, we used a model with a general configuration as a baseline and conducted an evaluation to compare it with other training methods that have been recently proposed.

We followed the configurations used in the literature to ensure a fair comparison of the training techniques [36, 104]. Further details are provided in Appendix D.1. We compared the following models in supervised and semi-supervised settings:

Supervised Models

We compared the following seven models:

- **Vanilla** [36]: A model with attention mechanisms that is trained without the use of AT- or VAT-based techniques, as described in Appendix A.1.
- **Word AT** [75]: Word embeddings trained with AT.
- **Word iAT** [76]: Word embeddings trained with iAT.
- **Attention AT** [104]: Attention to word embeddings trained with AT.
- **Attention iAT** [104]: Attention to word embeddings trained with iAT.
- **Attention VAT (ours)**: Attention to word embeddings trained with VAT.
- **Attention iVAT (ours)**: Attention to word embeddings trained with iVAT.

Table 3.2: Dataset statistics.

Task	Dataset	# classes	# train		# valid	# test	# vocab	Avg. # words
			# labeled (N_l)	# unlabeled (N_u)				
Single sequence task	SST [74]	2	6,920	-	872	1,821	13,723	18
	IMDB [91]	2	17.186	-	4,294	4,353	12,485	171
	AGNews [93]	2	51,000	-	9,000	3,800	13,713	35
Pair sequence task	CNN news [94]	584	190,149	190,149	3,924	3,198	70,192	773
	SNLI [96]	3	274,683	274,684	9,842	9,824	20,979	22
	MultiNLI [97]	3	157,080	157,081	78,541	19,647	53,112	34

Semi-supervised Models

We compared the following four models:

- **Word VAT** [49]: Word embeddings trained with VAT.
- **Word iVAT** [76]: Word embeddings trained with iVAT.
- **Attention VAT** (**ours**): Attention to word embeddings trained with VAT.
- **Attention iVAT** (**ours**): Attention to word embeddings trained with iVAT.

3.4.2 Datasets and Tasks

Table 3.2 presents the statistics for the labeled datasets, which include datasets for single-sequence tasks (e.g., text classification) and pair-sequence tasks (e.g., QA and NLI). We divided the dataset into training, validation, and test sets as done in [104]. Moreover, we preprocessed these datasets according to [36, 104]. Refer to Appendix B.1 for further details on the dataset and preprocessing.

Labeled Dataset

Following [36, 104], we evaluated the techniques using open datasets for three single-sequence and three pair-sequence tasks.

Labeled Dataset for Single-Sequence Tasks We used the following four datasets: Standard Sentiment Treebank (**SST**) [74], **IMDB**, a large movie reviews corpus [91], and the **AGNews** corpus [93]. The model was trained using the entire training set.

Labeled Dataset for Pair-Sequence Tasks We used the following three datasets: the **CNN** news article corpus [94], The Stanford Natural Language Inference (**SNLI**) [96], and the Multi-Genre NLI (**MultiNLI**) [97]. The model was trained using half of the randomly sampled training data. The other half was used as the unlabeled data, as described in Section 3.4.2.

Unlabeled Dataset

We describe the unlabeled data for the single-sequence and pair-sequence tasks.

Unlabeled Dataset for Single-Sequence Tasks We used the One Billion Word Language Model Benchmark dataset [110] as an unlabeled data pool for the semi-supervised learning step. This benchmark is used extensively to evaluate language modeling techniques and was recently employed by Chen et al. [105] for the model based on VAT. Thus, we considered that the dataset would also be useful in our VAT-based settings.

Unlabeled Dataset for Pair-Sequence Tasks Owing to the nature of the pair-sequence task, it is inherently difficult to use external sentences as unlabeled data, as in the case of single-sequence tasks. For training, we, therefore, decided to use half of the randomly sampled training data as labeled data and the other half as unlabeled data. The evaluation was performed using the unlabeled data. We adopted this evaluation method as it was used in the original VAT paper [75] and is a viable option.

3.4.3 Evaluation Criteria

We conducted four evaluations of model performance and interpretability when the proposed techniques were applied: (1) prediction performance following previous studies, (2) correlations based on word importance, (3) agreement with human annotations, and (4) the effects of the amount and selection of unlabeled data.

Prediction Performance

We used the F1 score, accuracy, and micro-F1 score for the single- and pair-sequence tasks, as in [36,104], to compare the performance of the model when the proposed and conventional VAT-based techniques were applied. The models were evaluated in supervised and semi-supervised settings.

Correlation with Word Importance

We compared how well the attention weights obtained using our VAT-based technique agreed with the importance of the words that were calculated using gradients [24]. We used Pearson's correlations between the attention weights and the word importance, as in the work of Mohankumar et al. [102] and our previous research [104], to evaluate the agreement. Refer to Appendix B.2.1 for details on the evaluation criteria.

Agreement with Human Annotation

We determined how strongly the attention weights that were obtained using our technique agreed with human annotations. DeYoung et al. [106] recently released a unified benchmark dataset including movie reviews, QA, and NLI, which featured human rationales for the decisions. They included both instance-level labels and the corresponding supporting snippets (i.e., rationales) from human annotators. We used this annotated dataset as an additional evaluation criterion for the movie review task in the IMDB dataset.

In accordance with [106], we used metrics that are appropriate for models that perform hard (discrete) rationale selection and soft (continuous) rationale selection. We used the intersection over union (IOU) and token-level F1 score as evaluation criteria for the hard rationale selection, and the area under the precision-recall curve (AUPRC), average precision (AP), and the area under the receiver operating characteristic curve (ROC-AUC) as evaluation criteria for the soft rationale selection, as in [106].

Effects of Amount and Selection on Unlabeled Data

To understand further the effects of unlabeled data in semi-supervised learning, we analyzed the impact of the amount of unlabeled data on the model's performance. We specifically focused on VAT-based techniques, such as Word VAT, Word iVAT, and our Attention VAT/iVAT.

Moreover, we investigated the effect of the selection strategy for the unlabeled data that were used in the semi-supervised learning in each task. We compared two strategies: the addition of unlabeled data without considering any of the contents, and the addition of data that are semantically close to the labeled data. Specifically, we applied random selection to the former and approximate nearest neighbor (ANN)-based selection to the latter.

Random Selection The random selection strategy randomly samples data from an unlabeled data pool. Unlabeled data can be prepared in large quantities compared to labeled data. Considering the relatively small size of the labeled dataset, the number of unlabeled datasets was adjusted accordingly for each task.

Approximate Nearest Neighbor (ANN) Selection The ANN selection strategy samples data from an unlabeled data pool in which the representations are close to the sentence representation that are contained in the labeled data. In this study, we applied ANN selection based on the following sentence representation:

- *Bag of words (BoW)-based representation*: The BoW representation was computed for each sentence in the pool.
- *Average word embedding-based representation*: The average of the word embeddings for each sentence in the pool was computed as the representation.

We selected the unlabeled data near $N_{\texttt{ANN}}$ (as the hyperparameter) for these representations using the ANN strategy and used these data for semi-supervised learning. We used Annoy[1] to search for

[1] spotify/annoy: `https://github.com/spotify/annoy`

neighbors and set $N_{\texttt{ANN}}$ to 10 as the neighborhood in the experiments. We designed the experiment for the ANN selection strategy to select the same amount of data as the random selection strategy.

3.5 Results

Tables 3.3 and 3.4 summarize the results of the prediction performance, and the correlations between the learned attention weights and word importance that was estimated using the gradient-based method, respectively. In both tables, (a) presents the results of the AT- and VAT-based techniques in which adversarial perturbations were added to the word or attention, whereas (b) displays the results of the iAT- and iVAT-based techniques in which improved interpretability was considered.

3.5.1 Prediction Performance

Table 3.3a indicates that our Attention VAT outperformed adversarial training for word embedding (Word AT), its semi-supervised extension (Word VAT), and adversarial training for attention mechanisms (Attention AT) in all performance measures. Moreover, Table 3.3b shows that our Attention iVAT exhibited the best prediction performance for all benchmarks. Thus, it can be confirmed that our Attention VAT/iVAT achieves better prediction performance than recent AT- and VAT-based techniques for the input space.

3.5.2 Correlation with Word Importance

According to Table 3.4a, the proposed Attention VAT yielded the strongest correlation between attention to words and word importance as determined by the gradients. In particular, VAT, which enables the use of unlabeled data, significantly improved the performance. As indicated in Table 3.4b, the proposed Attention iVAT achieved the highest correlation value in all datasets. Although our previous Attention iAT exhibited a significant increase in performance, Attention iVAT further increased the correlation by a substantial margin.

Table 3.3: Comparison of prediction performance. Unlabeled data were also used for the results of the VAT-based method, and the amounts of such data are listed below the table.

(a) AT- and VAT-based techniques

Model		Single-sequence task			Pair-sequence task	
		F1 [%]			Acc. [%]	Mic. F1 [%]	
	Unlabeled data	SST[1]	IMDB[2]	AgNews[3]	CNN[4]	SNLI[5]	MultiNLI[6]
Vanilla [36]	✗	79.27	88.77	95.27	56.25	73.52	56.59
Word AT [75]	✗	79.61	89.65	95.59	60.87	75.85	57.82
Word VAT [49]	✓	83.04	92.21	96.23	63.74	77.38	60.43
Attention AT [104]	✗	81.72	90.00	96.12	62.08	77.21	59.28
Attention VAT (**ours**)	✓	**83.18**	**92.48**	**96.35**	**64.93**	**77.87**	**61.07**

(b) iAT- and iVAT-based techniques

Model		Single-sequence task			Pair-sequence task	
		F1 [%]			Acc. [%]	Mic. F1 [%]	
	Unlabeled data	SST[1]	IMDB[2]	AgNews[3]	CNN[4]	SNLI[5]	MultiNLI[6]
Vanilla [36]	✗	79.27	88.77	95.27	56.25	73.52	56.59
Word iAT [76]	✗	79.57	89.64	95.62	61.21	75.91	57.91
Word iVAT [76]	✓	83.07	92.30	96.17	63.76	77.42	60.47
Attention iAT [104]	✗	82.20	90.21	96.19	62.15	77.23	59.42
Attention iVAT (**ours**)	✓	**83.22**	**92.56**	**96.48**	**65.08**	**78.13**	**61.18**

[1] N_u = 50,000; $N_u/N_l = 7.22$. [2] N_u = 150,000; $N_u/N_l = 8.23$. [3] N_u = 250,000; $N_u/N_l = 4.90$.
[4] N_u = 190,149; $N_u/N_l = 1.00$. [5] N_u = 270,683; $N_u/N_l = 1.00$. [6] N_u = 157,081; $N_u/N_l = 1.00$.

3.5.3 Agreement with Human Annotation

Table 3.5 compares the estimated important words (i.e., the attention) in the sentences with human annotations. The proposed techniques, especially our Attention iVAT, were more consistent with the human-provided rationales than existing AT techniques for the word as well as the baseline in both the hard and soft rationale selections. The BERT-to-BERT pipeline model [106][2] exhibited the strongest agreement with the human annotation. However, for rationale selection, it underwent supervised training with human annotations in a supervised manner for rationale selection, whereas

[2] They constructed a simple model in which they first trained the encoder to extract rationales, and then trained the decoder to perform prediction using only rationales based on the pipeline model [115]. The pipeline model adopts BERT for both the encoder and the decoder.

Table 3.4: Comparison of Pearson's correlation between attention to words and word importance estimated by gradients. The results are provided for the iVAT-based technique using unlabeled data.

(a) AT- and VAT-based techniques

Model		Single-sequence task			Pair-sequence task		
		Pearson's correlation					
	Unlabeled data	SST[1]	IMDB[2]	AgNews[3]	CNN[4]	SNLI[5]	MultiNLI[5]
Vanilla [36]	✗	0.852	0.788	0.822	0.638	0.741	0.517
Word AT [75]	✗	0.647	0.838	0.813	0.691	0.757	0.574
Word VAT [49]	✓	0.779	0.853	0.831	0.749	0.758	0.629
Attention AT [104]	✗	0.852	0.819	0.835	0.742	0.769	0.593
Attention VAT (**ours**)	✓	**0.898**	**0.879**	**0.903**	**0.766**	**0.784**	**0.638**

(b) iAT- and iVAT-based techniques

Model		Single-sequence task			Pair-sequence task		
		Pearson's correlation					
	Unlabeled dataset	SST[1]	IMDB[2]	AgNews[3]	CNN[4]	SNLI[5]	MultiNLI[6]
Vanilla [36]	✗	0.852	0.788	0.822	0.638	0.741	0.517
Word iAT [76]	✗	0.643	0.839	0.809	0.702	0.758	0.581
Word iVAT [76]	✓	0.781	0.859	0.893	0.751	0.761	0.631
Attention iAT [104]	✗	0.876	0.861	0.903	0.753	0.772	0.622
Attention iVAT (**ours**)	✓	**0.901**	**0.883**	**0.907**	**0.773**	**0.808**	**0.647**

[1] $N_u = 50{,}000$; $N_u/N_l = 7.22$. [2] $N_u = 150{,}000$; $N_u/N_l = 8.23$. [3] $N_u = 250{,}000$; $N_u/N_l = 4.90$.
[4] $N_u = 190{,}149$; $N_u/N_l = 1.00$. [5] $N_u = 270{,}683$; $N_u/N_l = 1.00$. [6] $N_u = 157{,}081$; $N_u/N_l = 1.00$.

the models that were applied in our technique did not use annotations, which means that the selection could be considered as unsupervised.

3.5.4 Amount and Selection of Unlabeled Data

Fig. 3.2 depicts the changes in the prediction performance with an increase in the amount of unlabeled data in the random selection strategy. It shows that the amount of unlabeled data was a crucial factor for this performance for both validation and test data. Here, because similar results were obtained in each of the three single- and pair-sequence tasks, two for each task are shown for reason

Table 3.5: Performance of vanilla and semi-supervised models for hard and soft rationale selection.

(a) Hard rationale selection

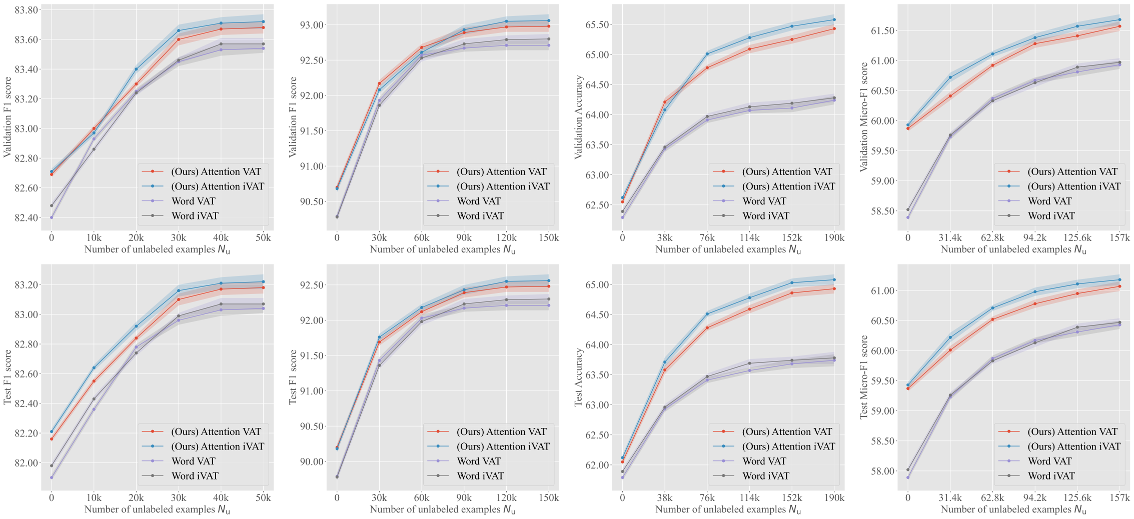
Model	IOU F1	Token F1
Vanilla [36]	0.011	0.097
Word VAT [49]	0.018	0.100
Word iVAT [76]	0.019	0.102
Attention VAT (**ours**)	0.030	0.126
Attention iVAT (**ours**)	**0.033**	**0.128**
BERT-to-BERT [106]	0.075	0.145

(b) Soft rationale selection

Model	AUPRC	AP	ROC-AUC
Vanilla [36]	0.326	0.395	0.563
Word VAT [49]	0.349	0.413	0.581
Word iVAT [76]	0.350	0.414	0.582
Attention VAT (**Proposed**)	0.403	0.477	0.646
Attention iVAT (**Proposed**)	**0.417**	**0.489**	**0.651**
BERT-to-BERT [106]	0.502	-	-

of space. The results of five experiments each with different seed values are shown with error bars. Specifically, in the single-sequence tasks, namely SST and IMDB, the performance improved as the amount of unlabeled data increased up to approximately seven times that of the labeled data. In the pair-sequence tasks, namely CNN news and MultiNLI, although we could only prepare up to twice the amount of unlabeled data because of the limitations of the experimental conditions as mentioned above, the accuracy also improved as the amount of data increased. This analysis demonstrates that VAT-based techniques provide significant benefits with a large amount of unlabeled data. Although we presented the results of four datasets (two for each dataset in the single- and pair-sequence tasks) that are representative of the six datasets used in the evaluation experiments, we found generally the same trends in the other datasets.

Table 3.6 compares the performance of the unlabeled data selection using the selection strategies in the semi-supervised learning process. Both the random and ANN strategies exhibited almost the same prediction performance in our techniques. This result suggests that the prediction performances



(a) SST (N_l = 6,920)

(b) IMDB (N_l = 17,186)

(c) CNN news (N_l = 190,149)

(d) MultiNLI (N_l = 157,081)

Figure 3.2: Model performance in validation and test score for single-sequence tasks (SST and IMDB) and pair-sequence tasks (CNN news and MultiNLI) with different amounts of unlabeled examples

of our techniques were almost the same, regardless of the selection strategy used. However, a remarkable improvement in the correlation with the word importance was observed. This finding is discussed in further detail in Section 3.6.2.

3.6 Discussion

3.6.1 AT and VAT for Attention Mechanisms

Our Attention VAT/iVAT techniques logically extend our Attention AT/iAT [104] techniques, according to our belief that attention is more important in identifying significant words in text/document processing than the actual word embeddings. Our proposed Attention VAT/iVAT significantly outperformed the Word AT/iAT in supervised settings, thereby reaffirming the above assumption. Moreover, our Attention VAT/iVAT exhibited equivalent performance to that of Attention AT/iAT, although our techniques did not use supervised labels for estimating the (virtual) adversarial perturbation. In such situations, AT, which can use supervised labels directly to compute the adversarial

Table 3.6: Comparison of unlabeled data selection performance in semi-supervised learning.

Model			SST		IMDB		AGNews	
	Selection strategy	Sentence representation	F1 [%]	Corr.	F1 [%]	Corr.	F1 [%]	Corr.
Attention VAT	Random		83.18	0.898	92.48	0.879	96.35	0.903
	ANN	BoW	83.20	0.903	92.51	0.884	96.38	0.908
		Avg. word embedding	83.21	0.927	92.50	0.902	96.39	0.928
Attention iVAT	Random		83.22	0.901	92.56	0.883	96.48	0.917
	ANN	BoW	83.26	0.909	92.59	0.891	96.51	0.924
		Avg. word embedding	83.26	0.941	92.60	0.914	96.52	0.957

direction, is advantageous. However, VAT, which relies on the current estimation for the computation and does not use labels, exhibits a fully equivalent performance, with no negative evidence observed.

In the semi-supervised settings, our VAT-based techniques further improved prediction performance and the correlation with the word importance in both the single- and pair-sequence tasks. Our techniques can use unlabeled data more efficiently than the conventional Word VAT, which, as well as its extension Word iVAT, have been validated extensively in semi-supervised settings. In summary, the application of VAT to the attention mechanism, rather than word embedding, resulted in improved model robustness (i.e., improved prediction performance) and interpretability (i.e., improved agreement with manually annotated rationales).

3.6.2 Amount and Selection of Unlabeled Data

The amount and selection of unlabeled data that are added for training the model are important factors for improving the performance in semi-supervised learning. First, we discuss the relationship between the amount of unlabeled data and the prediction performance. Our proposed techniques improved the performance until the unlabeled data were approximately seven times larger than the labeled data in the single-sequence tasks. Although the amount of unlabeled data was limited to the amount of training data, for the previously mentioned reasons, in the pair-sequence tasks an increase in the amount of unlabeled data sufficiently improved the prediction performance. This success is assumed to be owing to the VAT effectively utilizing unlabeled data, which contributed to smoothing the discriminative boundaries. It is noteworthy that the data added as unlabeled data produced a performance improvement even though they differed from the original dataset regarding type and quality.

Next, we discuss the selection strategy for unlabeled data. It is desirable for the input spaces of the unlabeled and labeled data to be similar when computing virtual adversarial perturbations using unlabeled data. Contrary to this empirical rule, our VAT for attention mechanisms worked effectively even when unlabeled data, from a domain different from that of the labeled data, were used. Our VAT for attention mechanisms essentially improved the model robustness. This confirms that our technique yields correct inference results for various inputs and can estimate the words that are important for the prediction as clear attention weights. As VAT computes the direction of the adversarial perturbation based on current model estimation, the addition of unlabeled data contributed to improving the performance, regardless of the selection strategy for the unlabeled data. This is possibly why our VAT for attention mechanisms need not be careful when selecting unlabeled data.

The training with more semantically similar unlabeled data from the pool (i.e., the ANN-based selection strategy) exhibited a stronger correlation between attention to words and the gradient-based estimates of the word importance than that without this consideration (i.e., the random selection strategy). The selection of similar unlabeled data is expected to offer certain advantages in terms of interpretability. As the ANN-based strategy selects similar data in a low-dimensional space, a higher correlation is anticipated. A performance improvement can be expected if a superior selection strategy is established in the future. Nevertheless, we emphasize that the proposed technique can improve model performance by simply adding a large unlabeled data pool to the training, without considering the similarity of the data.

3.6.3 Agreement with Human Annotation

The proposed Attention iVAT demonstrably exhibited substantially stronger agreement with human annotation than all other conventional techniques. In this evaluation, the BERT-to-BERT model [106] performed best, but this model was the only one that was trained to predict the textual portion of the rationale/evidence in a supervised manner. Hence, it was not possible to compare directly the models that applied our techniques. Our training technique can indirectly achieve highly efficient rationale selection for common and practical RNN-based models without using large models such as BERT and explicit learning for the rationale selection.

Similar to our previous proposal Attention iAT, Attention iVAT incorporates a process to emphasize the difference in attention to each word better, which may have contributed to the improved performance. Specifically, the norm of the weight of the difference in the attention of words (Eq. (3.8)) was normalized to 1. This process makes more effective use of the differences in attention to each word and calculates the virtual adversarial perturbations, which are determined based on more data than those in the former Attention iAT. Therefore, we believe that Attention iVAT can generate cleaner attention, and thus, achieve a higher hard/soft rationale selection score.

3.7 Conclusion

We have proposed novel, effective, and general-purpose semi-supervised training techniques for NLP tasks, namely Attention VAT and Attention iVAT. These methods are extensions to a semi-supervised setting using VAT of our previous techniques, namely Attention AT and Attention iAT respectively. Our training techniques significantly improve the prediction performance and model interpretability. In particular, our proposed techniques are highly effective when using unlabeled data and can outperform conventional VAT-based techniques. We confirmed that our techniques perform effectively even when using unlabeled data from a source other than the labeled data. That is, no careful selection (i.e., simple random selection) from the source is required. Our techniques can easily be applied to any DNN model with attention mechanisms.

The model used in the evaluation of the proposed technique in Chapters 2 and 3 is the RNN-based model. The application of the technique to the Transformer model, which is the current mainstream model, and the impact of our work on subsequent research is discussed in Chapter 6.

Chapter 4

Conversion Prediction Using Multi-task Conditional Attention Networks to Support the Creation of Effective Ad Creatives

Attention mechanisms have received a great deal of interest both in basic and applied research, but their evaluation is often limited to benchmark datasets in a laboratory environment. This chapter 4 aims to develop a framework that can effectively predict the effective ad creatives to serve and explain the trends of the effective ad creatives to us. The computational advertising field considers applications that are business-critical and operate with large amounts of real-world data, and support for the creation of more effective ad creatives for serving is one of the key subfields. Section 4.1 provides some background on the above. We describe in Section 4.2 the analysis and generation of ads with high impact, the CTR/CVR prediction task, and the background methods for the proposed framework in this chapter. We propose a new framework for supporting the creation of effective ad creatives that includes the following features in Section 4.3: (1) a new attention mechanism that considers the target user's preferences and (2) a new training method that simultaneously predicts the CTR and CVR. We confirm in Section 4.4 that the proposed framework is highly effective in evaluating operationally important metrics based on large-scale ad creative data obtained in the real world. Finally, we summarize this chapter in Section 4.5. We consider that the attention weights learned by the proposed framework can be applied to support the creation of effective ad creatives. In Chapter 6, we discuss the possibility that the word-level interpretations provided by the framework can practically assist in the operation, which relies mostly on human resources.

4.1 Background

In display advertisements, ad creatives, such as images and texts, play an important role in delivering product information to customers efficiently [116]. The performance of these advertisements is generally defined by *the revenue of conversions* per *the cost of the advertisement*. Conversions are user actions, such as the purchase of an item or the download of an application, and they represent a known metric that advertisers try to maximize through their ad creatives. The costs of advertisements are generally calculated by the cost per click (CPC), where an advertiser pays for the number of times their advertisement has been clicked. Therefore, the high performance of an ad is determined by minimizing the amount paid for the maximum number of conversions. Creating high-performing ad creatives is a difficult but crucial task for advertisers. The purpose of this study is to supporting the creation of ad creatives with many conversions, and we propose a new framework to support creating high-performing ad creatives, including accurate prediction of ad creative text conversions before delivery to the consumer[1]. If conversions of ad creatives can be predicted before delivery to consumers, advertisers can avoid the losses incurred by the high cost of ineffective advertisements. Moreover, because ad creatives with high click-through rates (CTRs), and low conversions have a tendency to deceive users, we also expect to improve the user experience on media displaying those ads. As a result, advertisers will be able to focus on improving the CTR

[1] We have also improved the CVR prediction using the result of conversion prediction.

of ad creatives.

Some attempts to support the creation of high-performing creatives by predicting ad creative conversions have been reported in the industry[2][3][4], but as far as we know, no academic research has been published in this area. Thomaidou et al. [117, 118] proposed a framework for generating ad creatives automatically. However, this framework focuses on search ads, and generates ad text according to set rules. Thus, this framework cannot be applied for our purpose. Some studies have reported that ad creatives affect the CTR of advertisements [119–121], but they do not predict the conversions. Prediction of a user's CTR or conversion rate (CVR) is a general task undertaken by many studies in this research area, but there are no studies that have predicted these rates for ad creatives. The prediction of an ad creative's performance is another important issue, but to the best of our knowledge, no study has examined this issue.

Although ad creatives are mainly image and text, we focus on the latter, and predicting its conversions. Because it is difficult to replace ad images, but easy to replace text, in this chapter, we propose a recurrent neural network (RNN)-based framework that predicts the performance of an ad creative text before delivery. The proposed framework includes three key ideas, namely, multi-task learning, conditional attention, and attention highlighting. Multi-task learning is an idea for improving the prediction accuracy of conversion, which predicts clicks and conversions simultaneously, to solve the difficulty of data imbalance. Conditional attention focuses on the feature representation of each creative based on its genre and target gender, thus improving conversion prediction accuracy. Attention highlighting visualizes important words and/or phrases based on conditional attention. We confirm that the proposed framework outperforms some baselines, and the proposed ideas are valid for conversion prediction. These ideas are expected to be useful for supporting the creation of ad creatives.

This research is motivated to support the creation of high performing creative text. The contributions are summarized as follows:

[2] https://www.facebook.com/business/m/facebook-dynamic-creative-ads

[3] https://www.adobe.com/en/advertising/creative-management.html

[4] https://support.google.com/google-ads/answer/2404190?hl=en

1. We propose a new framework that accurately predicts ad creative performance.

   To realize this, we propose two key strategies to improve the prediction performance of advertisement conversion.

   (a) Multi-task learning predicts conversion, together with previous click actions, by learning common feature representations.

   (b) The Conditional attention mechanism focuses attention on the feature representation of each creative text considering the target gender and genre.

2. We propose attention highlighting that offers important words and/or phrases using conditional attention.

A prototype implementation of the proposed framework with Chainer [122] has been released on GitHub[5].

4.2 Related Work

This study focuses on ad creatives. First, we describe existing studies that analyze high-performing ad creatives, and discuss how to generate them. Many studies on advertising creatives focus on images, and offer few results for texts. Furthermore, these studies focus on the CTR, rather than conversions. Second, we introduce studies on performance prediction for ads. In contrast to this study, which aims to predict the performance of new ads, these studies focus on images. Finally, highlighting studies related to our ideas, we introduce multi-task learning and RNN-based attention mechanisms.

4.2.1 Analysis and Generation of Effective Advertisements

Because ad creatives play an important role in the performance of ads, some studies analyzed ad creative performance [119–121]. For example, Azimi et al. [119] tried to predict some features of

[5] https://github.com/shunk031/Multi-task-Conditional-Attention-Networks

the CTR using ad creative images, and evaluated the effectiveness of visual features. The motivation of their study is similar to ours, but we focus on text instead of images in ad creatives and predict conversions rather than the CTR. Cheng et al. [120] proposed a model for predicting the CTR of new ads, and reported some knowledge using feature importance, but the text features of that study were based on fixed rules. With the development of deep learning, especially convolutional neural networks (CNNs) [4], visual features can be easily and effectively used for machine learning. Chen et al. [54] proposed Deep CTR, showing that using the features of ad images can significantly improve CTR prediction.

Thomaidou et al. [118] developed GrammaAds, which automatically generates keywords for search ads. In addition, they proposed an integrated framework for the automated development and optimization of search ads [117]. These studies support the creation of text ad creatives, but because these methods are rule-based, focusing only on search ads, the methods cannot be applied to display advertising.

4.2.2 CTR and Conversion Prediction in Display Advertising

CTR prediction of display advertising is important not only in the industry but also in academia. In [52, 53], a CTR prediction model was proposed using logistic regression (LR), and factorization machines (FMs) have been proposed to predict advertising performance [57–59]. In industry, LR and FMs are mainly used, because in display advertising, the prediction response time needs to be short to display an advertisement smoothly. In recent years, deep neural networks (DNNs) have been applied for predicting the advertisement CTR [54–56, 60, 123], and especially, some models combining DNNs with FMs have been proposed, and have improved predictions [56, 60, 123, 124]. The improvements achieved by these models show that explicit interaction between variables is important for advertisement performance prediction, so we adopted explicit interaction in our idea as a conditional attention mechanism.

There are several studies on CVR prediction [125–127], but there are not as many as the studies on CTR prediction. CVR prediction is difficult, because the number of conversions is imbalanced data that almost ad creative's conversions are zero. Existing studies tackled this difficulty. Yang et al. [128] adopted dynamic transfer learning for predicting the CVR, and demonstrating feature importance. Punjobi et al. [125] proposed robust FMs for overcoming user response noise. In this chapter, we tackle this difficulty using multi-task learning.

4.2.3 Background of the Proposed Strategies

In this chapter, we propose two key strategies for improving the prediction performance of advertisement conversion, namely, multi-task learning and a conditional attention mechanism. As the background of these strategies, we describe multi-task learning and the RNN-based attention mechanism.

Multi-task Learning. Multi-task learning [129] is a method that involves learning multiple related tasks. It improves the prediction performance by learning common feature representations. Recently, multi-task learning has been used in various research areas, especially natural language processing (NLP) [130, 131] and computer vision [132–134], and has achieved significant improvements. Conversions represent extremely imbalanced data, so conversion prediction is difficult. Because ad click actions represent a pre-action of conversion actions, click prediction may be related to conversion prediction. Therefore, we adopt multi-task learning, which predicts clicks and conversions simultaneously.

RNN-based Attention Mechanism. For supporting the creation of ad creative text, we use the knowledge of NLP. RNN-based models, such as long short-term memory (LSTM) [98], gated recurrent unit (GRU) [135], and attention mechanisms [33] have made breakthroughs in various NLP tasks, for example, machine translation [33], document classification [64, 128], and image captioning [34]. An RNN is a deep learning model for learning sequential data, and in NLP, this model can learn word order. Attention mechanisms compute an alignment score between two sources, and make significant improvements in some NLP tasks. Recently, self-attention [64], which computes alignment in a single source, was proposed. In addition, visual analysis using attention can highlight important phrases and/or words using the attention result, so the attention mechanism is also attractive for interpretability. In this chapter, we adopt a self-attention mechanism for improving conversion prediction performance and visualizing word importance.

4.3 Methodology

The outline of the proposed framework for evaluating ad creatives is shown in Figure 4.1. In the framework, we propose two strategies: multi-task learning, which simultaneously predicts conversions and clicks, and a conditional attention mechanism, which detects important representations in ad creative text according to the text's attributes.

Conversion prediction using ad creatives with an imbalanced number of conversions is a challenging task. Therefore, in multi-task learning, we expect to improve the model accuracy by predicting conversions along with clicks. The conditional attention mechanism makes it possible to dynamically compute attention according to the attributes of the ad creatives, its genre, and the target gender.

4.3.1 Framework Overview

The input of the proposed framework is ad creative text and ad creative attribute values. Ad creative consists mainly of two short pieces of text called the title and description. The ad attribute values are the gender of the delivery target and the genre of the ad creative, and they are related to the ad creatives.

Specifically, the input of the proposed framework is an ad creative text $S = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\}$ consisting of n word embeddings, where $\mathbf{w}_i \in \mathbb{R}^{d_w}$ represents the word vector at the i-th position in the ad creative text. Therefore, $S \in \mathbb{R}^{n \times d_w}$ is a two-dimensional matrix of the word sequence.

Incidentally, in the practical situation, a number of ad creative texts that have title and description texts are created for the target product. These texts often have different contexts for maximizing the amount of information empirically. Therefore, the proposed framework uses two *text encoders*, which learn the individual context from the title and the description.

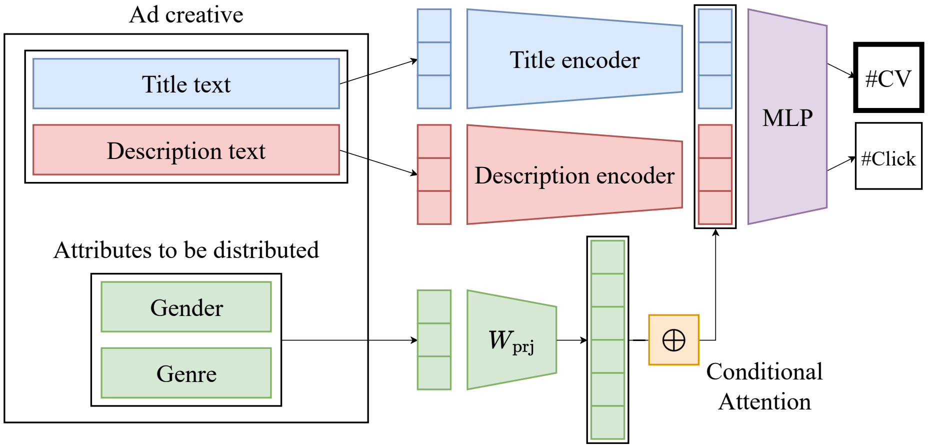


Figure 4.1: Outline of the proposed framework. In the framework, we propose two strategies: multi-task learning, which simultaneously predicts conversions and clicks, and a conditional attention mechanism, which detects important representations in ad creative text according to the text's attributes.

As a *text encoder*, we adopted the GRU, which can extract features from ad creative text considering word order. Specifically, title text $S^{\text{title}} = \{\mathbf{w}_1^{\text{title}}, \mathbf{w}_2^{\text{title}}, \cdots, \mathbf{w}_n^{\text{title}}\}$ and description text $S^{\text{desc}} = \{\mathbf{w}_1^{\text{desc}}, \mathbf{w}_2^{\text{desc}}, \cdots, \mathbf{w}_n^{\text{desc}}\}$ are input from the ad creative into title and description encoders, respectively, and are encoded into feature representations as $\mathbf{h}_t^{\text{title}} \in \mathbb{R}^{u_{\text{title}}}$ and $\mathbf{h}_t^{\text{desc}} \in \mathbb{R}^{u_{\text{desc}}}$; $t = 1, 2, \cdots, n$:

$$\begin{aligned} \mathbf{h}_t^{\text{title}} &= \text{title encoder}(\mathbf{w}_t^{\text{title}}, \mathbf{h}_{t-1}^{\text{title}}), \\ \mathbf{h}_t^{\text{desc}} &= \text{description encoder}(\mathbf{w}_t^{\text{desc}}, \mathbf{h}_{t-1}^{\text{desc}}). \end{aligned} \tag{4.1}$$

Let u_{title} and u_{desc} be the number of hidden units of the title and description encoders obtained here. The n hidden states can be expressed as $H^{\text{title}} = \{\mathbf{h}_1^{\text{title}}, \cdots, \mathbf{h}_n^{\text{title}}\}$ and $H^{\text{desc}} = \{\mathbf{h}_1^{\text{desc}}, \cdots, \mathbf{h}_n^{\text{desc}}\}$, respectively. Compute a vector $\mathbf{x}_{\text{feats}}$ that concatenates these hidden states, H^{title}, H^{desc}, one-hot vectors of gender features $\mathbf{x}_{\text{gender}} \in \mathbb{R}^{d_{\text{gender}}}$, and genre features $\mathbf{x}_{\text{genre}} \in \mathbb{R}^{d_{\text{genre}}}$:

$$\mathbf{x}_{\text{feats}} = \text{concat}(H^{\text{title}}, H^{\text{desc}}, \mathbf{x}_{\text{genre}}, \mathbf{x}_{\text{gender}}). \tag{4.2}$$

Note, $\mathbf{x}_{\text{feats}} \in \mathbb{R}^{d_{\text{feats}}}$;$d_{\text{feats}} = n \times (u_{\text{title}} + u_{\text{desc}}) + d_{\text{gender}} + d_{\text{genre}}$. These concatenated vectors are inputted in a multi-layer perceptron (MLP) which is an output layer of the proposed framework. To predict conversions $\hat{y}^{(\mathrm{cv})}$ and clicks $\hat{y}^{(\mathrm{click})}$, multi-task learning described later predicted $\hat{\mathbf{y}}_{\mathrm{multi}} = \{\hat{y}^{(\mathrm{cv})}, \hat{y}^{(\mathrm{click})}\}$ through the MLP:

$$\hat{\mathbf{y}}_{\mathrm{multi}} = \mathrm{MLP}(\mathbf{x}_{\mathrm{feats}}). \tag{4.3}$$

To improve the performance of the model robustness, we use wildcard training [90] with dropout [136] for the input word embeddings.

4.3.2 Multi-task Learning

Conversion prediction is difficult, due to the imbalanced data, so we use the strategy of multi-task learning. Multi-task learning is a method that solves multiple tasks related to each other, and that improves the prediction performance by learning common feature representations. We adapt multi-task learning, and predict clicks and conversions prediction simultaneously. Because click prediction may be related to conversion prediction, we expect to improve the prediction performance by learning common feature representations using multi-task learning.

In multi-task learning, the input is a feature vector of a training sample denoted by $\mathbf{x}$, and the ground truth is y. For training samples $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, a single model, f, learns to generate predictions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N\}$:

$$\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N). \tag{4.4}$$

We minimize the mean squared error (MSE) over all samples, N, in $l = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. In K supervised tasks, the multi-task model, $F = \{f_1, f_2, \cdots, f_K\}$, learns to generate predictions $\hat{\mathbf{y}} = \{\hat{y}^{(1)}, \hat{y}^{(2)}, \cdots, \hat{y}^{(K)}\}$:

$$\hat{\mathbf{y}} = F(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N). \tag{4.5}$$

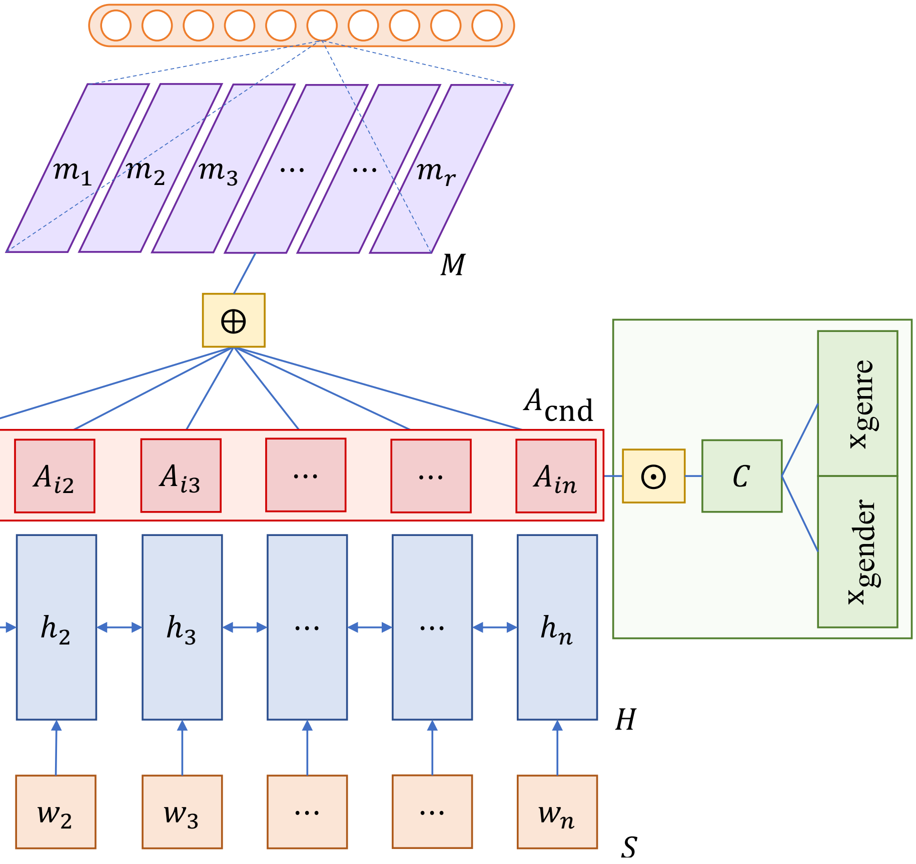


Figure 4.2: Example of the conditional attention mechanism. Conditional attention is calculated from the element-wise product of the attention matrix A and the feature vector $\mathbf{c}$ consisting of the gender and the genre.

The total loss is calculated from the sum of loss in each task,

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_i^{(k)} - \hat{y}_i^{(k)} \right)^2 . \tag{4.6}$$

In this task, for ground truth of $y^{(\mathrm{cv})}$ and $y^{(\mathrm{click})}$, we minimize losses for predicted conversions $\hat{y}^{(\mathrm{cv})}$ and clicks $\hat{y}^{(\mathrm{click})}$:

$$\mathcal{L}_{\mathrm{multi}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^{(\mathrm{cv})} - \hat{y}_i^{(\mathrm{cv})} \right)^2 + \lambda \frac{1}{N} \sum_{i=1}^{N} \left(y_i^{(\mathrm{click})} - \hat{y}_i^{(\mathrm{click})} \right)^2 , \tag{4.7}$$

where $\lambda > 0$ is the hyper-parameter to control the effect of the click loss.

4.3.3 Conditional Attention

We propose the strategy of the conditional attention mechanism. Supporting the creation of ad creatives by considering attribute values is useful, but the conventional attention mechanism learns keywords or key phrases, by calculating the alignment score using only the input sentence.

In this paper, we propose a conditional attention mechanism to calculate self-attention, using feature vectors obtained from the attribute values of the ad creative. Figure 4.2 illustrates the conditional attention mechanism. It can consider ad creative attributes against the conventional attention mechanism.

The conditional attention mechanism is calculated from the attention of the *text encoder* and the feature vector obtained from the attribute values of the ad creative text. Each word in the word sequence S is independent of the others. To capture these word order relations, we apply a *text encoder* to the text, to obtain the hidden state $\mathbf{h}_t \in \mathbb{R}^u$. The n hidden states of these $u \times n$ dimensions can be expressed as $H = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n\}$.

To consider ad attribute values, a *conditional* vector, $\mathbf{c} \in \mathbb{R}^n$, is calculated by performing a linear combination of $\mathbf{x}_{\text{feats}} \in \mathbb{R}^{d_{\text{feats}}}$ and trainable parameters $W_{\text{prj}} \in \mathbb{R}^{n \times d_{\text{feats}}}$:

$$\mathbf{c} = W_{\text{prj}} \mathbf{x}_{\text{feats}}. \tag{4.8}$$

Here, we use *self-attention* [64] for computing the linear combination. The attention mechanism takes the entire hidden state H of the *text encoder* as the input and outputs attention vector $\mathbf{a}$:

$$\mathbf{a} = \text{softmax}(\mathbf{w}_{s2}^T \text{tanh}(W_{s1} H)), \tag{4.9}$$

where $W_{s1} \in \mathbb{R}^{n \times u}$ and $\mathbf{w}_{s2} \in \mathbb{R}^n$ are trainable parameters. Because H is an $n \times u$ dimension, the size of attention vector $\mathbf{a}$ is n. The softmax$(\cdot)$ is calculated so that the sum of all the weight is 1.

Furthermore, we calculate the *conditional attention vector* using the attributes given to the ad creative. The *conditional attention* vector, $\mathbf{a}_{\text{cnd}}$, is calculated using conditional vector $\mathbf{c}$ and

Table 4.1: Features included in the ad creative dataset. It contains 1,694 campaigns, some of which were part of campaigns delivered by Gunosy. The average lengths of the title and description texts are about 15 and, 32 characters, respectively. The Campaign ID feature is not directly inputted in the model, because the ID is used for evaluations with cross-validation based on the ID.

	Features	Feature Description	Details
	Campaign ID	Campaign ID in Gunosy Ads	1,694 campaigns
Texts	Title	Title texts	Avg. 15.44±3.16 chars
	Description	Description texts	Avg. 32.69±5.43 chars
Attrs	Genre	Genre of the creatives	20 types
	Gender	Gender of delivery target	3 types

attention vector $\mathbf{a}$:

$$\mathbf{a}_{\text{cnd}} = \mathbf{a} \odot \mathbf{c}. \tag{4.10}$$

Here, $\odot$ is an element-wise product. We want r different parts to be extracted from the ad creative texts. Thus, the *conditional attention* vector $\mathbf{a}_{\text{cnd}}$ becomes *conditional attention* matrix $A_{\text{cnd}} \in \mathbb{R}^{n \times r}$. Therefore, sentence vector $\mathbf{m}$ with the embedded ad creative text becomes sentence matrix $M \in \mathbb{R}^{u \times r}$. The *conditional attention* matrix, A_{cnd}, is multiplied by hidden state H of the *text encoder*, and the r-weighted sentence matrices are calculated as follows:

$$M = HA_{\text{cnd}}. \tag{4.11}$$

In the proposed framework, the model makes predictions based on the calculated M and ad creative attributes, such as $\mathbf{x}_{\text{gender}}$ and $\mathbf{x}_{\text{genre}}$.

Table 4.2: Comparison of the prediction performance of CVs in mean squared error (MSE) criteria. The proposed multi-task learning and conditional attention reduced MSE in almost all the categories, especially estimating cases where the number of conversions (#CV) is one or more (#CV > 0). However, "All predicted as zero" showed sufficiently low MSE in this category, due to too many CV = 0 in this dataset. Therefore, we conclude using MSE as an evaluation metric is not suitable in this study.

Model		MSE			
		All		#CV >0	
		Single-task	**Multi-task**	Single-task	**Multi-task**
MLP		0.01712	0.01698	0.04735	0.03199
GRU	Vanilla	0.01696	0.01695	0.04657	0.04355
	Attention	0.01685	0.01688	0.04695	0.03105
	Conditional attention	**0.01683**	**0.01675**	**0.04641**	**0.02825**
All predicted as zero		**0.02148**		—	

4.4 Experiments

4.4.1 Dataset

We use real-world data from the Japanese digital advertising program Gunosy Ads[6], provided by Gunosy Inc.[7]. Gunosy Inc. is a provider of several news delivery applications, and Gunosy Ads delivers digital advertisements for these applications. Gunosy is a news delivery application that achieved more than 24 million downloads in January 2019.

For evaluation, we used 14,000 ad creatives delivered by Gunosy Ads from August 2017 to August 2018. In digital advertising, the cost of acquiring a conversion is called the cost per acquisition (CPA). Advertisers set target CPAs for a product, and manage its ad creatives to improve their performance. When the target CPAs for creatives are different, the trend of conversions may also vary, and for this reason, the dataset we selected comprises ad creatives where the target CPA was within a certain range. In addition, we removed creatives with a low number of impressions[8] from the dataset. As shown in Table 4.1, the title, description, and genre of the ad creative, as well as the gender to which the ad is delivered, are used as input features. Note that the Campaign ID

[6] `https://gunosy.co.jp/ad/`

[7] `https://gunosy.co.jp/en/`

[8] An occasion when a particular advertisement is seen by someone using the application.

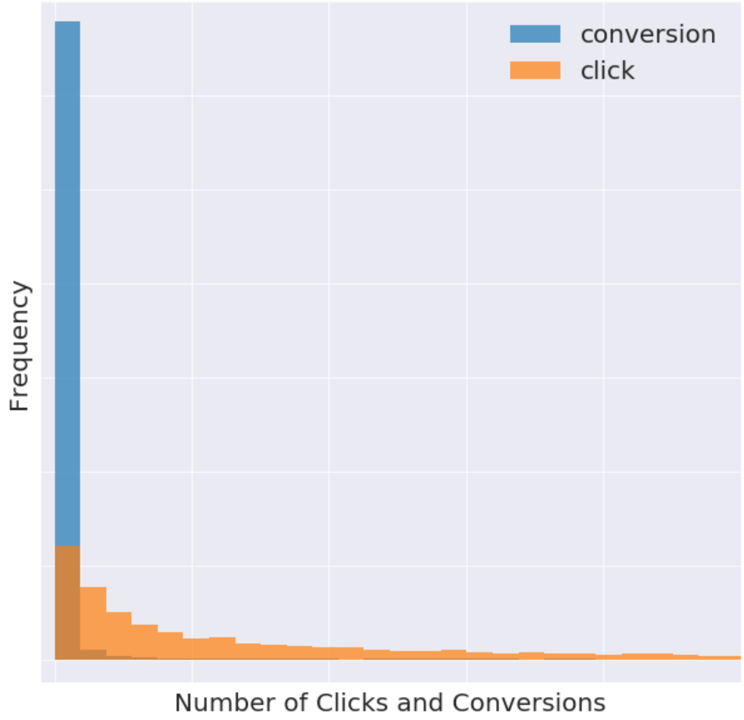


Figure 4.3: Distribution of clicks and conversions in the dataset. The number of conversions is concentrated on zero. Compared with the number of conversions, the number of clicks indicates a long-tail distribution.

is not a feature directly used as an input in the model, because the ID is used for evaluating with cross-validation based on the ID.

Creative texts written in Japanese are split into words using MeCab [137], a morphological analysis engine for Japanese texts, and mecab-ipadic-neologd [138], which is a customized system dictionary that includes many neologisms for MeCab. The number of clicks and conversions is log-normalized.

Figure 4.3 shows a histogram of the number of clicks and conversions. The number of conversions is concentrated on zero, and in relation, the number of clicks is a long-tailed distribution. Therefore, the ad creative dataset is definitely imbalanced. Figure 4.4 shows the distribution between the number of clicks and conversions in the dataset. The correlation coefficient between the number of clicks and conversions is 0.816, which is a strong correlation. As a reminder, we hide the number of

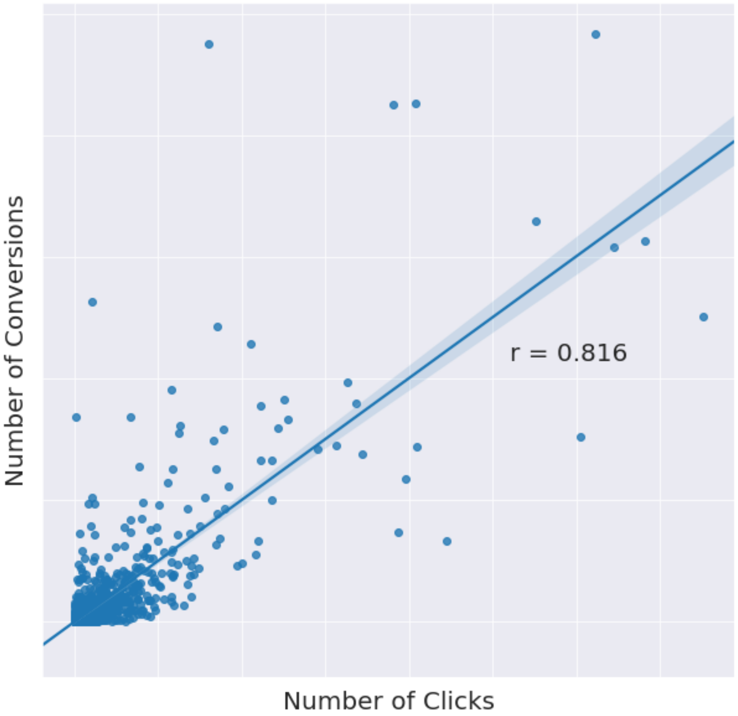


Figure 4.4: The linear relation between the clicks and conversions in the dataset (correlation coefficient $r = 0.816$).

clicks and conversions, also their frequencies, for confidentiality reasons.

4.4.2 Experimental Settings

In these experiments, support vector regression (SVR) and an MLP-based *text encoder* were used as a baseline model. When inputting creative text in the SVR model, we used average-pooled sentence representations computed from word representations, using pre-trained word2vec (w2v) [139]. The same pre-trained w2v was used as word embedding for the proposed model.

We compared and examined the following models: MLP (not considering word order) and GRU (considering word order) as the *text encoder* in the proposed framework. LSTM was also considered as a candidate for the baseline model; however, it showed no improvement in performance, so it was excluded from the experiment. In addition, CNNs are known to be capable of training at high speed, because they can perform parallel calculations, compared with LSTM and GRU, and their performances are also known to be equal. Nevertheless, these methods were excluded in these experiments, because we were targeting an RNN-based model that can apply attention for visualizing the contributions of words to ad creative evaluation.

We compared the proposed models used in the proposed framework. The following models were compared and examined, to confirm the effect of multi-task learning in conversion prediction:

Single-task: A commonly known model that predicts conversions only; and

Multi-task: A model that simultaneously predicts the number of clicks and the number of conversions.

To confirm the effect of the conditional attention mechanism, we compared the following models:

Vanilla: A simple *text encoder* without an attention mechanism. It is a baseline in the proposed model;

Attention: A mechanism that introduces self-attention to the *text encoder*. It makes it possible to visualize which word contributed to prediction during creative evaluation; and

Conditional Attention: A mechanism introduced to the *text encoder* of the proposed method. Conditional attention can be computed and visualized considering the attribute values of the ad creative. Different attentions can be visualized by changing the attribute value for the same creative text.

In addition, the hyper-parameter setting is described below. The mini-batch size was set to be 64, and the number of epochs was set to be 50. For multi-task learning, we used a fixed value of $\lambda = 1$. In the *text encoder*, the number of hidden units was set to be 200 for u_{title} and u_{desc}. For all models, we use Adam [140], with a weight decay of $1\mathrm{e}^{-4}$, for parameter optimization.

4.4.3 Evaluation Metrics

First, as evaluation metrics, we adopt not only MSE but also normalized discounted cumulative gain (NDCG) [141], which is evaluation metrics for ranking. MSE measures the average of the

Table 4.3: Comparison of the normalized discounted cumulative gain (NDCG) in the proposed model. When calculating NDCG scores, the results for all data and the scores restricted to the top 1% of conversions (#CV) were calculated.

Model		NDCG [%]			
		All		#CV top 1 %	
		single	**multi-task**	single	**multi-task**
SVM		96.72		83.73	
MLP		96.68	97.18	82.97	84.12
GRU	Vanilla	96.54	97.00	76.39	78.51
	Attention	96.76	97.11	83.00	85.49
	Conditional Attention	**96.77**	**97.20**	**87.11**	**87.14**

squares of the errors, which is the average squared difference between the estimated values and what is estimated. We adopted ranking evaluation metrics because the number of conversions is imbalanced. As shown in Figure 4.3, most ad creative conversions are zero and imbalanced. A high evaluation score can be achieved by an overfit model that predicts all outputs as zero when such metrics are used. For the creation of high-performing ad creatives, rather than predicting zero conversions, we would like to accurately predict high-conversion creatives as such.

NDCG is mainly used in the experiments. NDCG is the discounted cumulative gain (DCG) normalized score. In DCG, the score decreases as the evaluation of an advertisement declines, so a penalty is imposed if a low effect is predicted for highly effective creatives. At the time of the NDCG calculation, after obtaining the rank of the ground truth, and its predicted value, respectively, evaluation scores are calculated for all the evaluation data, as well as those restricted to the top 1% of conversions.

For ad creative evaluation, the metrics are computed with cross-validation. In most advertising systems, advertisements are delivered in units of campaigns. In a campaign, the target gender and its genre are set, and multiple ad creatives are developed.

In this paper, we predict the number of conversions for ad creative text in unknown campaigns, and confirm the generalization performance of the proposed framework. Therefore, at the time of the evaluation, five-fold cross-validation was performed in such a manner that the delivered campaigns did not overlap.

4.4.4 Experimental Results

For confirming the accuracy of the proposed framework compared with the baselines, we compared single-task and multi-task learning, and the results of the application of the conditional attention mechanism are described. Through almost all the results, the proposed framework applying multi-task learning and the conditional attention mechanism achieved a better performance than the other methods. Especially, when focusing on ad creatives with many conversions, the proposed framework achieved high prediction accuracy.

Table 4.2 shows the MSE score with all the evaluations in each model, and with one or more conversions in each model. Almost all the results show that the model applying the multi-task learning and conditional attention mechanism had a smaller MSE score than the other models did. Overall, the RNN-based GRU showed better performance than the baseline models. Therefore, the results suggest that it is important to properly capture word order when evaluating creative texts. Compared with *vanilla* and *attention*, in the proposed model, *conditional attention* showed a better performance.

Although the improvement of all datasets is weak, because as shown in Figure 4.3, the number of conversions of many ad creatives is zero, the MSE is small, even if the conversion of most ad creatives is predicted to be zero. Therefore, we evaluated data with conversions other than zero. As a result, we found that the proposed model exhibits much better performance than the baseline model for data with one or more conversions. The proposed model was able to predict creatives with more conversions than the baseline models.

To evaluate ad creatives with many conversions as such, we used the ranking algorithm NDCG. The NDCG result in the proposed model is shown in Table 4.3[9]. The NDCG score (regarded as *All* in Table 4.3) for all the datasets is shown for reference, because as noted above, most samples have zero conversions. The performance of the GRU model that considers word order compared with the baseline model improved by an average of approximately 3-5%, with many conversions.

[9] The same tendency was observed even when mean average precision (MAP) was used as an evaluation metric.

Table 4.4: Comparison of NDCG between the CVR directly predicted by the single-task model and the CVR (#conversions / #clicks) calculated from the multi-task GRU model's predicted conversions and clicks. The threshold value for calculating NDCG is assumed to be a CVR of 0.5 or more.

Model		NDCG [%]
Single-task	Vanilla	80.54
	Attention	82.58
	Conditional attention	83.89
Multi-task	Vanilla	82.63
	Attention	84.27
	Conditional attention	**85.61**

In the NDCG result (Table 4.3), the multi-task model realized higher prediction accuracy than the single-task model predicting only conversions did. A score improvement of approximately 1-2% was confirmed when compared with the baselines. Because clicks are highly correlated with target ad conversions, as shown in Figure 4.4, rather than predicting conversions alone, training the model to multi-task by predicting clicks simultaneously can improve prediction accuracy. By training clicks and conversions, the proposed model seems to implicitly learn features that contribute to conversion prediction.

Because several previous studies predicted the CVR directly, we also calculated it, using the prediction of the multi-task learning model, and compared the accuracy. In a multi-task model, the CVR can be calculated by dividing conversions by clicks. In Table 4.4, the multi-task model is compared with the single-task model by directly estimating the CVR. The prediction performance of the multi-task model is higher than that of the single-task model. Although the number of clicks and conversions predicted by multi-task learning may not always be close to the ground truth, the ratio of the number of clicks to the conversion number is captured properly.

In Table 4.3, the conditional attention mechanism achieved better results the NDCG metric. In particular, the conditional attention mechanism showed better results than the conventional attention mechanism did. In the conventional attention mechanism, the training was focused solely on the co-occurrence relation between words in the input text, but the conditional attention mechanism can predict conversion by using the attribute value.

Table 4.5: Comparison of GRU models for creative texts and their attribute value interactions. Performance is improved using conditional attention rather than giving attribute values directly to word vectors.

Model		NDCG [%]	
		Single-task	**Multi-task**
w2v + attributes	Vanilla	77.84	78.03
	Attention	80.39	83.52
w2v	**Conditional attention**	**87.11**	**87.14**

Table 4.5 shows the result comparing feature interaction between w2v-based embeddings and ad attribute values. In the proposed framework, this interaction is realized with the conditional attention mechanism, explicitly. Because attention is computed by the input variables, this interaction is implicitly expressed by inputting both variables in the *text encoder*. For confirming the effect of this explicit interaction in the conditional attention mechanism, we compared the model that inputted both variables in the *text encoder* with the conditional attention mechanism. The conditional attention mechanism showed the best performance in the single-task and multi-task model. Introducing the vanilla model and the conventional attention model to the word representation with ad attribute values resulted in a poor performance, mainly because the duplicate interactions were calculated excessively. It is suggested that it is better to introduce the explicit interaction of attribute values.

4.5 Conclusion

In this paper, we propose a new framework to support the creation of high-performing ad creative text. The proposed framework includes three key ideas, multi-task learning and conditional attention improve prediction performance of advertisement conversion, and attention highlighting offers important words and/or phrases in text creatives. We confirmed that the proposed framework realizes an excellent performance thanks to these ideas, through experiments with actual delivery history data.

In the future, we will build a framework that simultaneously uses images attached to ad creatives, and aim to improve the accuracy of conversion prediction.

Chapter 5

Ad Creative Discontinuation Prediction with Multi-task Hazard Networks

In chapter 4, we proposed a new framework focused on supporting the creation of more effective ad creatives. In contrast, support for discontinuing ad creatives that have become less effective in terms of serving is another important aspect of ad operation support. This chapter 5 aims to formulate the discontinuation of ad creatives. We outline in Section 5.1 that there have been various studies on support for ad creatives with high serving effectiveness, while there have been few studies on support for ad creatives with low serving effectiveness, and explain Section 5.2 based on specifically related studies. In Section 5.3, we analyze real-world data obtained from ad operations and confirm that there are both short- and long-term discontinuation trends for ad discontinuation problem, which have been previously well formulated. Against the background, we propose in Section 5.4 a new framework with a well-used attention mechanism for supporting ad operations that predicts ad discontinuation within the survival time analysis. We evaluate the proposed framework using large-scale datasets obtained through actual operation in Section 5.5, and conduct case studies in Section 5.6. Finally, we summarize this chapter in Section 5.7. The presentation of important words that serve as the basis for prediction by the proposed framework contributes greatly to the final decision of the operator. The interpretability of prediction by the framework is discussed in Chapter 6.

5.1 Background

With the increasing importance of digital advertising, the significance of ad operations has also increased. Ad operations consist of various tasks, such as preparing ad creatives, determining the bid price and the target audience, and discontinuing inefficient ad creatives. These operations are essential for business revenue; however, operators' abilities are dependent on their experience. Supporting ad operations is an important topic of study in the field of machine learning (ML) [61, 142–145]. Discontinuing ad creatives at the appropriate time is a crucial operation and can have a significant impact on sales. However, to the best of our knowledge, no studies have been conducted so far that support these operations.

To support ad operations, we address the discontinuation of ineffective ad creatives based on ML methods. Ad creatives are generally discontinued manually by ad operators when they become less effective. Considering that ad-serving algorithms (e.g., bid price optimization [143, 144]) are required to serve a large number of effective ad creatives, multiple studies have been conducted to develop algorithms that can provide accurate predictions [54, 146]. However, we posit that predicting ad performance for operational efficiency requires other mechanisms because the utilization of the results is different from efficient ad serving. In other words, in this chapter, we focus on predicting the effectiveness of the ad creatives so as to properly discontinue ineffective ones.

Empirically, discontinuation falls under two categories: (a) *cut-out* (i.e., for short-term discontinuation) and (b) *wear-out* (i.e., for long-term discontinuation). In cut-out discontinuation, ad creatives are discontinued after a short period of time due to incompatible serving effects, whereas in wear-out discontinuation, ad creatives are discontinued when ad creatives' performance deteriorates after a long time. Ad operators need to take different actions for each type of discontinuation. We clarified these empirical differences by analyzing real-world data (described in detail in Section 5.3).

We propose a framework based on a multi-modal deep neural network (DNN) for predicting the discontinuation of ad creatives that draws on the idea of survival prediction [147]. Survival analysis/prediction has been developed for use in the medical field [148] to model the time until events of interest; however, it is now commonly used in various other fields [149, 150] and is particularly effective in digital advertising [151–153]. Our framework integrates a hazard function, which is frequently used in survival prediction, into a loss function.[1] We believe that our loss function is more appropriate for predicting the timing of discontinuation than conventional classification or regression frameworks are.

For the two types of discontinuation identified with pre-data analysis, we employ a two-term estimation technique to predict the two terms through multi-task learning (MTL) [129]. Moreover, we employ a new click-through rate (CTR)-weighting technique for the loss function to identify the property of more valuable ad creatives. We used 1,000,000 real-world ad creatives, including 10 billion scale impressions, to evaluate our framework in offline experiments with a concordance index (CI) [154] and the F1 score, and case studies that simulate online experiments in two real-world practical cases. Currently, our framework is deployed in a production environment and provides predictions for our internal operators.[2]

The contributions of our study are summarized as follows:

- **Analyzing large-scale ad creative dataset**: We analyzed 1,000,000 real-world ad creatives, including 10 billion scale impressions, and identified two important aspects. (a) There are two reasons for the discontinuation of ad creatives: cut-out (for the short term) and wear-out (for the long term). (b) The business impact of discontinuation varies significantly for different ad creatives.

- **Proposing a new framework for predicting ad creative discontinuation**: We propose

[1] It is controversial to solve ad discontinuation prediction as a survival prediction. In this chapter, we do not elaborate on this topic but only discuss the idea of the hazard function. We intend to tackle this issue in our future studies.

[2] Based on the predictions, we have discussions with the operations team to further improve the operation efficiency.

Table 5.1: Classification of machine learning methods to support the process for ad creatives.

The process for ad creative operations	Machine learning methods to support the operation process
(1) Creating and submitting to the ad platform	Thomaidou et al. [155], Kitada et al. [142], Hughes et al. [156], Mishra et al. [61]
(2) Serving to users on the platform	Zhang et al. [157], Shen et al. [158], Maehara et al. [143], Doan et al. [145]
(3) Discontinuing when the ads become less effective	**Our study**

a multi-modal DNN-based framework for predicting ad discontinuation with a hazard function-based loss function. Our framework includes two effective techniques: (i) a two-term estimation technique with MTL and (ii) a CTR-weighting technique for the loss function. Our framework demonstrated significantly better performance (short: 0.896, long: 0.939, and overall: 0.792 in the CI) than the conventional method (0.531 in the CI) did. Additionally, our framework outperformed classification- and regression-based frameworks.

- **Evaluating the framework through practical case studies**: We evaluated our framework from two important perspectives of case studies and confirmed the following: For short-term cases, our framework demonstrated an equivalent or better discontinuation effect compared with manual operations. For long-term cases, our framework provided a high degree of similarity of order to manual discontinuation orders compared with other methods, based on other practical indicators; this is important for long-running ad creatives.

5.2 Related Work

5.2.1 Supporting Ad Operations

Table 3.1 shows a classification of ML methods that support the process for ad creatives. The operation process for ad creatives generally includes the following: (1) created and submitted to the ad platform, (2) served to users on the platform, and (3) discontinued when they become less effective. There are several studies to support (1) and (2) based on ML methods, such as supporting the creation of high-performing ad creatives [61, 142, 155, 156], determining the best bid price [143, 157, 159], and deciding to whom to serve the ad creatives [145, 158, 160]. However, to the best of our knowledge, no studies have been conducted on supporting the discontinuation process.

Thus, we focus on process (3) and develop a framework for predicting the appropriate timing for the discontinuation of a ad creative.

The serving performance of ads can be inefficient for a variety of reasons, one of which is wear-out. Wear-out is an event caused by decreasing ad performance due to repeated serving to users. To control for wear-out, frequency capping,[3] which restricts the number of times an ad creative is served to a user, is a traditional function, and the number of servings of an ad is useful in predicting the CTR and/or conversion rate (CVR) [146]. The event is commonly discussed in marketing science literature [161], and an ad-serving algorithm that incorporates wear-out as a feature has been previously proposed [162]. However, no studies have directly predicted the wear-out of ad creatives.

In related studies about ad creative discontinuation, a bandit algorithm has been commonly used for selecting effective ad creatives [163]. Using the algorithm, the efficiency of ad creatives can be estimated correctly with fewer impressions. However, ad operators must decide which ads to discontinue based on their performance; however, determining the appropriate timing of discontinuation demands much experience.

5.2.2 Survival Prediction and Hazard Function

Survival prediction is a branch of statistics that deals with time-to-event data, that is, data where the objective variable is the time until the event occurs. The effectiveness of the strategy has been confirmed across various domains, such as credit score analysis [150] and customer analysis [149]. In digital advertising, survival prediction helps to predict the conversion rate [164], user experiences after clicking on an ad [151], the winning bid price [152], and the failure rate of the reserve price [153].

The main feature of the survival prediction task is to correctly model censored data such that no event occurs during the observation period. However, in the ad discontinuation task, such data are scarce because the operator manually determines the time of discontinuation.[4] When such censoring data are scant or nonexistent, Zhong et al. [165] reported that the survival prediction task can be

[3] https://support.google.com/google-ads/answer/117579

[4] Although there are some cases where the serving volume is low, they are very few.

regarded as equivalent in terms of formulation to the classification task. Although it is debatable whether survival prediction is suitable for ad discontinuation tasks, we assume that the properties of ad discontinuation have some similarities with those of survival prediction tasks. Thus, we attempt to utilize the hazard function, which is an important idea in survival prediction.

The hazard function can model the main variable of interest as the time until an event. This function has been used to represent the decaying survival probability, and deep learning-based survival prediction models have been developed by training a loss function based on the function [147, 166, 167]. We considered that the function is suitable for modeling the discontinuation probability of exponentially decaying ad creatives (described in detail in Section 5.3). Through training based on the hazard function with a deep learning model, we expect to develop a better prediction model that will help ad operators with discontinuation.

5.3 Pre-Data Analysis

In this section, we clarify the characteristics of the two types of discontinuation (cut-out and wear-out) by analyzing a real-world dataset. First, we provide an overview of the dataset. Second, we analyze the dataset in terms of serving days and sales aspects. Then, we discuss the difference between the two types of discontinuation based on the analysis.

5.3.1 Dataset Overview

We used real-world data from the Japanese digital advertising program Gunosy Ads,[5] provided by Gunosy Inc.[6] Gunosy Inc. is a provider of several news delivery applications, and Gunosy Ads delivers digital advertisements for these applications. The applications have been downloaded more than 53 million times as of November 2019. We used approximately 1,000,000 ad creatives, including 10 billion scale impressions, served from 2018 to 2019 by Gunosy Ads. The ads were managed in units called *campaigns*, and each campaign was configured with a target gender, ad genre, and cost

[5] Gunosy Ads `https://gunosy.co.jp/ad/en/`

[6] Gunosy Inc. `https://gunosy.co.jp/en/`

Table 5.2: Examples included in the ad creative dataset. In addition to the Japanese title text and description text shown here, the dataset contains categorical information (e.g., gender of the serving target, genre of the ad creative), images (e.g., the image attached to the ad creative), and numerical information (e.g., the number of impressions, clicks, conversions, and the CPA for the ad creative).

Title text	Description text	Gender	Genre
1000万人が選ぶ！みんなが選んでいるゲーム10選 *Chosen by 10 million people! The 10 games played by everyone.*	スマホに入れておきたい無料ゲームを限定でご紹介 *Exclusively introducing free games* *that you will want to install on mobile phone.*	All	Game App.
-10kgのダイエットに成功！痩せる理由はこれ *Success in -10 kg weight loss! This is the reason for getting slim.*	女子に人気の方法で効果を実感 *Realizing the effects popular among girls.*	Female	Healthy Food
有名芸能人監修。簡単にできる料理レシピ本 *Supervised by a famous celebrity; easy cookbook.*	一人暮らしの男性にもおすすめ！ *Recommended for men living alone!*	Male	Books

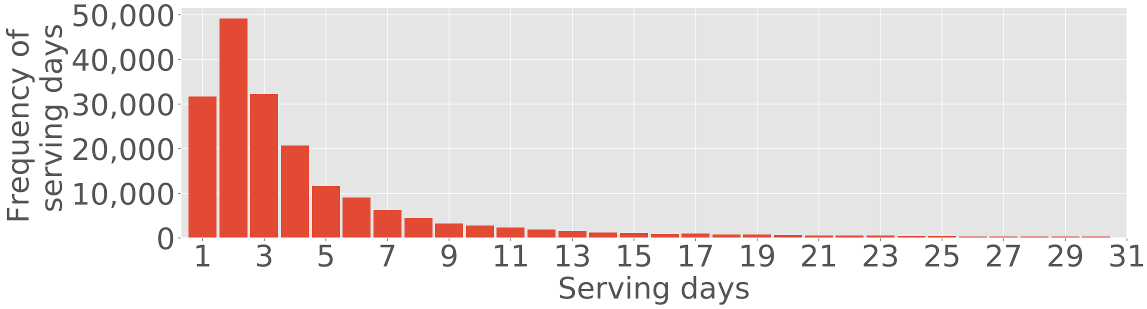


Figure 5.1: Frequency distribution of the number of serving days of the overall ad creatives in one month. Many creatives were discontinued within three days of serving.

per acquisition (CPA). Table 5.2 shows the examples included in the ad creative dataset. Each ad creative was associated with a campaign and consisted of an image, title, and description. Ad performance data, including the number of impressions, clicks, conversions, and sales, were tabulated for each ad creative daily. Because the data are confidential, the actual values cannot be disclosed. Thus, relative values are used when discussing the data.

5.3.2 Analysis of the Dataset

Variation of Serving Days

First, we analyzed the distribution of the serving days, which is the period of time since the ad creatives were served, as in Figure 5.1. Many ad creatives were discontinued within three days of serving, and the continuous serving days of the ad creatives were exponentially distributed.

Table 5.3: Percentage of the number of ads served and their sales on the serving date. While 80% of the ads were served within a week of serving, the 20% of the ads that were served for more than a week accounted for 80% of the sales.

Serving days	Number of Ads [%]	Sales [%]
$[0, 3)$	42.66	1.90
$[3, 7)$	38.72	16.34
$[7, +\infty)$	18.62	81.76

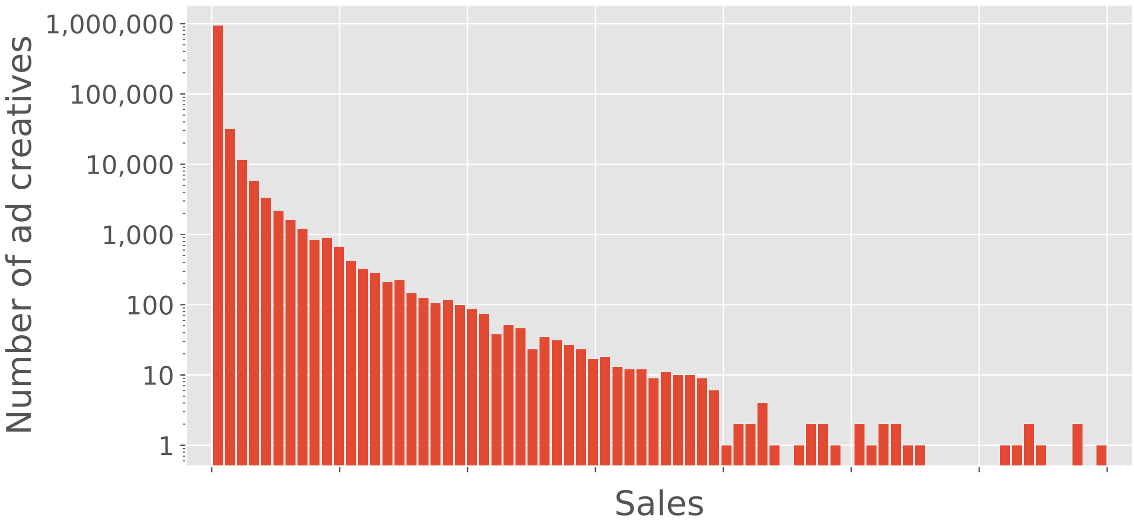


Figure 5.2: Histogram of ad creative sales in descending order. The y-axis is log-scaled. Most ad creatives are included in the first bin.

Table 5.3 shows the percentage of the number of ads served and their sales on the serving days. Although the number of ad creatives that were served for a long term is small, these creatives account for a large percentage of advertising sales. Specifically, comparing the serving days of $[0, 3)$ and $[7, +\infty)$, the former has the majority of the ad creatives, but the latter accounts for the highest percentage of sales. From this difference, we assume that each discontinuation is caused by different operations. We discuss the differences in detail in Section 5.3.3.

Variation in Sales

Figure 5.2 shows the histogram of ad creative sales in descending order. Sales vary considerably for each ad creative. Some are high-sales ad creatives, but most do not contribute to business revenue.

Thus, the impact of each ad creative on the revenue is very different.

Accordingly, we attempted to reflect the high-sales ad creatives for the prediction framework. If the framework is trained with equal use of all ad creatives, the framework would reflect the trends of low-sales ad creatives excessively. We assumed that the construction of the framework requires the weighting of learning depending on sales. Therefore, we employ a CTR-weighting technique for the loss function, as described in Section 5.4.4.

5.3.3 Two Types of Discontinuation

In this section, we discuss the differences between the two types of discontinuation based on the analysis in Section 5.3.2. We also describe how the differences can be handled effectively using our framework.

The short-term discontinuation is labeled *cut-out*. In digital advertising, a best practice is to create various patterns of ad creatives and look for effective ones by serving them. Clearly, it is not easy to find an effective pattern, and thus, many ad creatives are discontinued in the short-term. As Figure 5.1 shows, many ad creatives are discontinued early on, but this selection process is conducted manually. Thus, it is a burden on the operation loads. In fact, automation of the process has significant cost advantages (i.e., both time and money) from a business perspective.

In contrast, the discontinuation of an ad creative after it has been served for a long-term period is called *wear-out* [161]. Wear-out occurs when the ad creatives are overexposed to customers and response drops. The discontinuation of long-running ad creatives damages sales; however, the wear-out of ads is inevitable. Thus, ad operators are tasked with creating new compelling ad creatives for their replacement. Ad operators prioritize this task by estimating wear-out, which heavily depends on the operator's experience. Thus, by predicting wear-out, we expect that the sales loss could be minimized. In terms of ad operations, predicting wear-out has a different role from predicting cut-out.

In short, cut-out occurs when the user is not interested, and wear-out occurs when the user becomes fatigued. The operations involved in each type of discontinuation also differ. In the case of a cut-out, the operator discontinues many inefficient ad creatives rapidly, whereas in the case of a wear-out, the operator estimates when an ad creative's performance degrades. It is important that our framework predicts both types of discontinuation. Predicting cut-out (for the short term) can save human effort and enable efficient advertising operations. Predicting wear-out (for the long term) can have a significant impact on business revenue.

The properties of these two types of discontinuation are so different that they should be considered as two different tasks. Therefore, we employ a two-term estimation technique (as described in Section 5.4.3). Additionally, we analyze two real-world cases to evaluate the availability of each type of discontinuation (as described in Section 5.6).

5.4 Methodology

We propose a multi-modal DNN-based framework for predicting the timing of the discontinuation of ad creatives, based on the idea of survival prediction. First, we provide an overview of our framework. Then, we introduce a loss function based on the hazard function for training our framework. Additionally, we employ two techniques to enhance the performance for the prediction: (1) a two-term estimation technique with MTL and (2) a CTR-weighting technique for the loss function. We believe that it is important to make the DNN model simpler, especially for business problems, and we introduce these effective learning techniques for the simple DNN model.

5.4.1 Overview of the Proposed Framework

Figure 5.3 shows the outline of our framework. For predicting the timing of ad creative discontinuation, we adopt and integrate the hazard function, which is based on the discrete-time survival prediction strategy [147]. From the input ad creative X observed in time t, our framework outputs hazard probability $\boldsymbol{h}$ for each pre-defined L time interval $\mathcal{T}$ such that $(0, t_1]$, $(t_1, t_2]$, $\cdots$, $(t_{L-1}, t_L]$ through hazard function $f(\cdot|\cdot)$ with sigmoid function σ:

$$\boldsymbol{h} = \sigma(f(t|X)) \in \mathbb{R}^L. \tag{5.1}$$

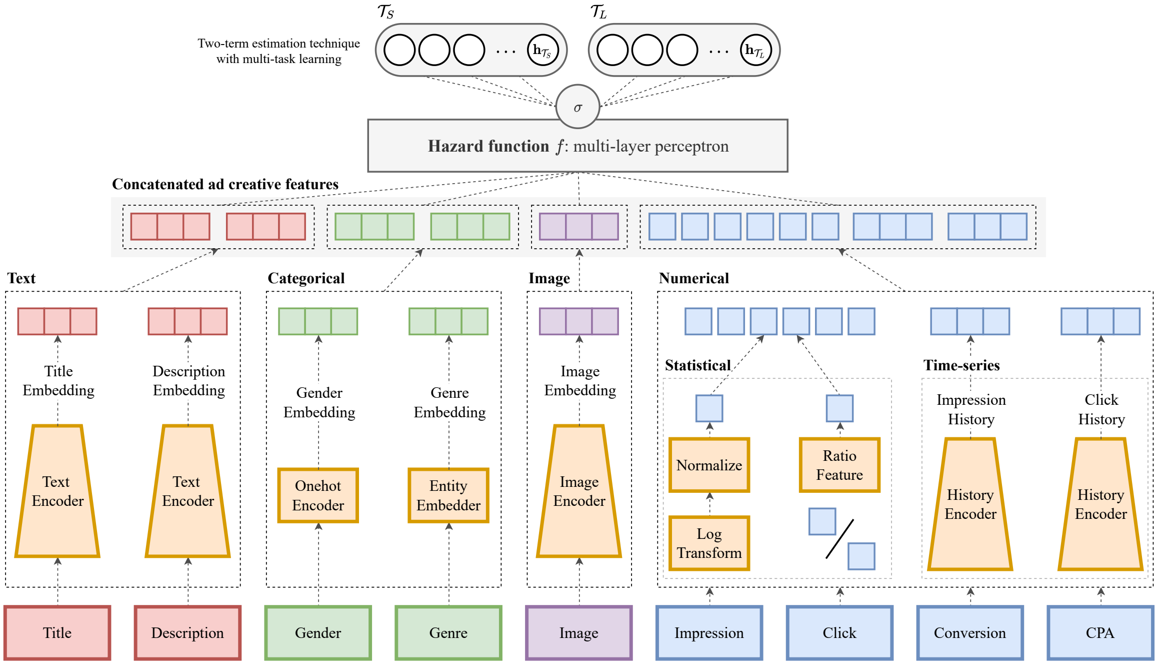


Figure 5.3: Outline of our framework that exploits a hazard function, which draws on the idea of survival prediction, to predict the discontinuation of ad creatives. The input includes the four types of features: text, categorical, image, and numerical features. The output is the hazard probability, which includes whether the target ad creative has been discontinued in each time interval.

The details of how the function is learned are in described in Section 5.4.2.

The input to our framework is ad creative X observed in time t; the framework uses the four types of features that constitute an ad creative: text features $\boldsymbol{x}_1$ (e.g., title, description), categorical features $\boldsymbol{x}_2$ (e.g., gender of the serving target, genre of the ad creative), image features $\boldsymbol{x}_3$ (e.g., the image attached to the ad creative), and numerical features $\boldsymbol{x}_4(t)$ (e.g., the number of impressions, clicks, conversions, and the CPA for the ad creative), which include statistical features and time-series features. $\boldsymbol{x}_1$ to $\boldsymbol{x}_3$ are invariant with time interval t; $\boldsymbol{x_4}$ varies with t due to the serving performance.

For the input features, each type of feature is encoded according to the type, such as text embeddings $\boldsymbol{e}_1$, categorical embeddings $\boldsymbol{e}_2$, image embeddings $\boldsymbol{e}_3$, and numerical embeddings $\boldsymbol{e}_4(t)$. For all features, we performed conventional encoding and aggregation as well as encoding based on DNN techniques. Due to the space limitation, the details of these input features and encoding processes are presented in Appendix E.1. Additionally, we briefly mention the properties of the features obtained by these encoding processes in the section.

For predicting the hazard probability, the encoded features are finally integrated using a multi-layer perceptron (MLP) to estimate the serving days of an ad creative in the pre-defined time intervals:

$$\boldsymbol{h} = \sigma(f(t|X)) = \sigma(\mathrm{MLP}([\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4(t)])) \in \mathbb{R}^L. \tag{5.2}$$

5.4.2 Learning the Hazard Function for Ad Creative Discontinuation

We describe the problem settings of the hazard function in discrete-time survival prediction and how it can be learned for ad discontinuation. The hazard function can model the main variable of interest as the time to an event. To predict the timing of discontinuation, we are interested in discrete problems because currently discontinuation is done manually in real-world scenarios.

In discrete-time survival prediction, the follow-up time is broken up into time intervals, which are often right-closed and left-open. Let t' denote the time at which the event occurs, $l = 1, 2, \cdots, L$ index the time intervals, and t_{l-1} and t_l denote the lower and upper bounds of the l-th time interval, respectively. The hazard probability within the l-th time interval is defined as follows:

$$h_l = Pr(t' \in (t_{l-1}, t_l] | t' > t_{l-1}). \tag{5.3}$$

Here, the time interval is the same setting defined in Section 5.4.1.

The hazard probability can be modeled as a function of covariates to assess the effect of covariates on the time to the event of interest. It is our belief that this property is appropriate for the discontinuation of ad creatives. The hazard probability of the observation surviving the l-th time interval h_l is equal to the l-th output of the hazard function, which is equivalent to the following:

$$h_l = \sigma(f_l(t_l|X)). \tag{5.4}$$

where f_l is the l-th output of hazard function $f(\cdot|\cdot)$.

The parameters are estimated through the maximization of the likelihood function for the discrete survival prediction strategy [147]. The likelihood is written as follows:

$$\ell(\mathcal{T}, X) = \prod_{l=1}^{l'} h_l^{\delta_l} (1 - h_l)^{(1-\delta_l)}. \tag{5.5}$$

where l' is the number of time intervals that was observed, and δ_l is the event indicator within the l-th time interval, which is equal to 1 if the event occurs within that interval, and 0 otherwise. To maximize the likelihood function, we minimized the following negative log-likelihood:

$$\mathcal{L}(\mathcal{T}, X) = \sum_{l=1}^{l'} \delta_k \log h_l + (1 - \delta_l) \log(1 - h_l). \tag{5.6}$$

The full log-likelihood is the sum of the log-likelihood for each data and we minimize the log-likelihood by mini-batch gradient decent.

5.4.3 Two-term Estimation Technique: For Short- and Long-term Discontinuations

We employ a two-term estimation technique to improve the performance for the two types of discontinuations (cut-out and wear-out). To account for the difference between the two types of discontinuations, we train two prediction models by setting different time intervals for each model. We call these two models the *short-term model* and the *long-term model*, respectively.

The short-term model learns the cut-out features, and the long-term model learns the wear-out features. In the short-term model, for time interval $\mathcal{T}_S$, which has $L_{\mathcal{T}_S}$ intervals, the target log-likelihood is defined as $\mathcal{L}(\mathcal{T}_S, X)$. In the long-term model, for time interval $\mathcal{T}_L$, which has $L_{\mathcal{T}_L}$ intervals, the target log-likelihood is defined as $\mathcal{L}(\mathcal{T}_L, X)$. Training with these different time intervals for each model is expected to enable learning of different properties of short- and long-term discontinuation, which will provide better predictions.

The MTL allows us to train short-term and long-term models in a unified single model. In applying this technique to our framework to train the model, we used the following two likelihoods:

$$\mathcal{L}_{\text{multi}}(\mathcal{T}_S, \mathcal{T}_L, X) = \lambda\mathcal{L}(\mathcal{T}_S, X) + (1-\lambda)\mathcal{L}(\mathcal{T}_L, X), \tag{5.7}$$

where λ is a regularization coefficient that controls the balance between two loss functions. In this chapter, we set λ to 0.5.[7] The two hazard probabilities for the short- and long-term time intervals outputted from the MTL technique can be merged, multiplying by the last interval of the short-term time interval and the beginning of the long-term time intervals, to the last time interval of the long-term time intervals. We used this result as the overall model of MTL with short- and long-term time intervals.

In particular, $\mathcal{T}_S$ and $\mathcal{T}_L$ were set to be as follows:

$$\mathcal{T}_S = (1,3],(3,5],(5,7],(7,10], \quad \mathcal{T}_L = (1,10],(10,30],(30,60],(60,90],(90,120]. \tag{5.8}$$

The time intervals for the short- and the long-term model were defined based on observations of a dataset in which most ads were discontinued in about 10 days or 120 days, respectively, under the assumption [147] that it was effective to make the interval width exponential.

5.4.4 CTR-weighting Technique for the Hazard-based Loss Function

We employ a CTR-weighting technique for the hazard function-based loss function. It is important to capture the features of ad creatives that have a high CTR, because this is known to contribute to business revenues, as stated in Section 5.3.2. Thus, we utilize the weighting technique to the loss function with the CTR of ad creatives:

$$\mathcal{L}_{\text{CTR}}(\mathcal{T}, X) = (r_{\text{CTR}} + 1) \cdot \mathcal{L}(\mathcal{T}, X), \tag{5.9}$$

[7] As the fraction of short-term discontinuation is much larger than that of long-term discontinuation, tuning hyperparameter λ would yield better performance.

where $r_{\mathrm{CTR}} = \#\texttt{click} \ / \ \#\texttt{impression}$ and generally takes the value $0 \leq r_{\mathrm{CTR}} \leq 1$. Thus, we added one to this value to prevent the loss value from becoming too small for data with a low CTR. For comparison, we used the impression-weighting technique $\mathcal{L}_{\mathrm{imp}}$ with the same settings as the CTR-weighting technique.[8]

In the MTL, to apply our CTR-weighting technique, we used the following likelihood:

$$\mathcal{L}_{\mathrm{CTR}}^{\mathrm{multi}}(\mathcal{T}_S, \mathcal{T}_L, X) = \lambda \mathcal{L}_{\mathrm{CTR}}(\mathcal{T}_S, X) + (1-\lambda)\mathcal{L}_{\mathrm{CTR}}(\mathcal{T}_L, X). \tag{5.10}$$

5.5 Offline Experiments

5.5.1 Motivation and Tasks of Offline Experiments

The purpose of the offline experiments is to evaluate whether our framework can correctly predict ad discontinuation using observed data. To this end, we used the real-world dataset of ad creatives shown in Section 5.3. The discontinuation date of the ad creative was defined as the date on which the ad operator discontinued it manually. The dataset was divided into a training set, a validation set, and a test set in a campaign-based and stratified fashion[9], comprising 600,000, 200,000, and 200,000 ad creatives, respectively.

5.5.2 Evaluation Criteria of Offline Experiments

We evaluated our framework on the following aspects: (1) comparison with the conventional survival prediction-based methods and its suitable metrics and (2) comparison with the classification and regression frameworks. Regarding the former, considering our framework draws on the idea from

[8] Using sales directly as a weighting should be avoided in terms of business aspects because it would be too significantly affected by an advertiser's business.

[9] In most advertising systems, advertisements are served in units of campaigns, and multiple creatives with similar trends are developed in a campaign. Thus, to avoid data leakage due to potential similarity, we divided the dataset depending on the campaigns.

Table 5.4: Comparison of concordance indexes with different time intervals and different features. The time interval was compared for the short term, long term, and the whole period. We calculated the concordance indexes using the prediction for all test data and the prediction for the top 25% of sales data in the test data.

Model		Feature				Concordance Index					
						Short term		**Long term**		Overall	
		Stat.	Text	Image	Time.	All	Top 25% of sales	All	Top 25% of sales	All	Top 25% of sales
Cox time-varying PH [168]		✓	✓			0.6098	0.7293	0.6574	0.6932	0.5287	0.5320
Random Survival Forest [169]		✓	✓			0.6213	0.7578	0.6832	0.7294	0.5311	0.5487
Our framework	Single-task	✓	✓			0.7958	0.8262	0.7541	0.8001	0.5874	0.6001
		✓		✓		0.7962	0.8232	0.7536	0.8045	0.5536	0.6045
		✓			✓	0.7931	0.8346	0.7880	0.8397	0.5935	0.6101
		✓	✓	✓		0.7962	0.8313	0.7575	0.8113	0.5908	0.6113
		✓	✓	✓	✓	**0.8289**	**0.8640**	**0.7892**	**0.8456**	**0.6225**	**0.6456**
	Multi-task	✓	✓	✓	✓	**0.8700**	**0.8712**	**0.9228**	**0.9293**	**0.7915**	**0.8049**

survival prediction, we follow the evaluation in a survival prediction manner. For the latter, we evaluate the prediction made by the hazard function in our framework.

Survival Prediction-based Models with CI

For the performance evaluation, we employed a CI [154], which is the standard measure for model assessment in survival prediction strategies. A CI evaluates, for each random pair, the probability that the two predicted survival times will be in the same order as the actual survival time. For context, CI = 0.5 is the average CI of a random prediction, whereas CI = 1.0 is the perfect ranking of a prediction. We believe that this measure is sufficient to evaluate the framework based on our survival prediction.

From a business perspective, it is important to accurately predict the duration of top-selling creatives. The differences between the sales of the ad creatives were found to be substantial, as shown in Section 5.3. To determine whether the high-value creatives were accurately predicted, we evaluated only the top 25% of the ad creatives in the dataset based on their sales values.

For comparison with the survival prediction technique, we used the conventional methods of the Cox time-varying proportional-hazards model (Cox time-varying PH) [168] and random survival forest [169]. The hyperparameters of these methods tuned the validation set, and we report the prediction results for the test set. To confirm the effects of our short-term and long-term models, we constructed an overall model that includes the time interval used in the short- and long-term models. The hyperparameter settings of our framework are described in Appendix E.2.

Classification and Regression Frameworks with F1 score

To validate our framework, we compared it with binary classification and regression frameworks that simulate direct discontinuation. This leads to an intuitive evaluation with other frameworks for which the CI cannot be directly applied. The details of the classification and regression models are described in Appendix E.3.

We used the F1 score to compare our framework with the classification/regression frameworks. The only difference between our framework and the others is the output objective; our framework predicts the hazard probability at each time interval in a single-task model, whereas the classification/regression framework predicts whether to discontinue at the time interval in each model.

5.5.3 Evaluation Results of Offline Experiments

Comparison of Survival Prediction-based Models

First, we compared the conventional method of the Cox time-varying PH and random survival forest in terms of the statistical and text features. Table 5.4 shows in terms of the CI. We can regard a model with the same features (i.e., the statistical and text features) under a single-task setting as a baseline based on our DNN. Compared the DNN baseline with the conventional methods, our DNN-based framework showed better performance, with an approximate 10-point gain. By using our framework, not only can the performance be improved, but various features, such as image and time-series features, can be accounted for as well. Furthermore, comparing our DNN baseline with the full model of our framework (i.e., the model with all features under multi-task setting), we can confirm an average improvement of 3 points in prediction performance.

For the short- and long-term models, the models that incorporated all features, such as image features, text features, and time-series features, had the best scores by an average of approximately

Table 5.5: Comparison of prediction performance by proposed CTR-weighted loss in the multi-task model. We compared the vanilla model (without any special loss), $\mathcal{L}_{\text{imp}}$, and $\mathcal{L}_{\text{CTR}}$.

Model	Concordance Index		
	without	$\mathcal{L}_{\text{imp}}$	$\mathcal{L}_{\text{CTR}}$
Short-term	0.8700	0.8884	**0.8958**
Long-term	0.9228	0.9297	**0.9390**
Overall	0.7915	0.8060	**0.8127**

4 points. In the short-term model, the text and image features contributed more compared with the long-term model. However, in the long-term model, the time-series features contributed more than in the short-term model. These differences are in line with the trends of the two types of discontinuation. The effect of time-series data is described in Appendix E.4.

Compared with the models where the short- and long-term models were trained individually, we observed even better performance in the model that was trained simultaneously with MTL. The performance improved significantly by approximately 5 points and 13 points in the short- and long-term models, respectively. MTL further boosted the overall performance, with a clear improvement of about 17 points.

Table 5.5 shows the MTL model performance comparison to confirm the effect of the CTR- and impression-weighting techniques. By introducing the CTR-weighting technique, the prediction performance was improved in the short-term, long-term, and overall models. The impression-weighting technique performed better than the vanilla model (e.g., without any weighting for the loss) but did not show a better performance compared with the CTR-weighting technique.

In the short-term model with CTR weighting, the CI showed an improvement of approximately 3 points, and in the long-term model, the improvement was approximately 2 points. The improvement in performance was observed even when impression weighting was introduced; however, the performance improved more when the CTR-weighting technique was introduced. We assume that the large variance in the number of impressions is the reason for the difference between these scores. Thus, weighting with CTR enables training for more accurate prediction.

Table 5.6: Comparison of the F1 score in discontinuation prediction after N days using the binary classification framework and regression framework.

	Short-term		Long-term	
	3 days	7 days	30 days	90 days
Classification framework	0.3208	0.3106	0.1098	0.0980
Regression framework	0.3813	0.2852	0.0969	0.0735
Our framework	**0.7988**	**0.7796**	**0.7287**	**0.7004**

Comparison of Classification and Regression Frameworks

Table 5.6 compares our framework and classification/regression frameworks. Compared with other frameworks, our framework performed better by approximately 40 points in the short-term model and by 60 points in the long-term model. In the short-term model, our framework achieved superior performance, about twice that of the other frameworks. In the long-term model, the classification/regression frameworks did not appear to perform well. Based on these results, our hazard function-based approach represents an effective method for predicting ad creative discontinuation problems.

5.5.4 Discussion of Offline Experiments

The Effect of Input Features on Prediction Performance

In our framework, the effective feature of the prediction of short- and long-term ad discontinuation is different. In the short-term model under the single-task setting, the text and image features that make up the ad creative contribute to the prediction. The short-term discontinuation (i.e., cut-out) may be caused by many users who are not immediately interested in the ad creatives. Since the text and image affect the user's impression of the ad creative, this result reflects characteristics of the short-term discontinuation. In the long-term model, the time-series feature contribute to the prediction. The long-term discontinuation (i.e., wear-out) is caused by the user getting fatigued, as described in Section 5.3. This is due to the fact that users have been repeatedly interacting with similar or the same ads, and thus the CPA has been decreasing. Since the time-series feature captures the user's fatigue by change of the ad performance, this result reflects characteristics of the long-term discontinuation. As a result, we confirmed that each input feature contributes to the prediction of both cut-out and wear-out appropriately.

The Effect of Two-term Estimation with MTL and CTR-weighting Technique

Ad creative discontinuations are of two types with different characteristics, as described in Section 5.3. To accurately predict the appropriate timing for the discontinuations, we introduce MTL into our two-term estimation technique. Additionally, we introduce the CTR-weighting technique to learn more valuable ad creative features. We confirmed that these techniques contribute to the prediction of ad creative discontinuation.

For ad creative discontinuation that has two different types of characteristics, we demonstrated that the model with our two-term estimation technique is superior to the single (overall) model. The short- and long-term models trained by the two-term estimation technique achieved much higher prediction performance than the single overall model, which included both short- and long-term time intervals. Furthermore, we found that training a unified model of these two-term models by MTL achieved better performance than the short- and long-term models alone. We believe that the unified model with MTL can learn the fine-grained property of short- and long-term discontinuation than the single overall model.

Our CTR-weighting technique is expected to learn more valuable ad creative features accurately. The results showed that our model was a more accurate predictor with CTR-weighting loss than without it. It improved the accuracy of the prediction of discontinuation of top-selling ad creatives and the overall accuracy. Consequently, it can be assumed that accurate training of effective ad creatives will have a positive effect on the overall prediction.

Our Hazard function-based Framework vs. Classification/Regression Framework

Our framework achieves much higher prediction performance for the ad discontinuation prediction problem by introducing the hazard function instead of the conventional classification/regression framework. We assume that there are two reasons due to the property of the discontinuation problem: (1) the data imbalance and (2) the time dependence of the discontinuation. Regarding the data imbalance, most of the ad creatives were discontinued early. If the classification/regression framework is trained on such data, it may output only biased predictions. Regarding the time dependence of the discontinuation, the probability of the ad creative discontinuation increases as time goes on. The classification/regression framework does not have the assumption of time dependency, while the hazard function-based framework is known to have the assumption. We consider that the hazard function appropriately captures the property of the two reasons derived from the ad creative discontinuation problem. In contrast, the classification/regression framework does not seem to sufficiently handle the problems of the data imbalance and the time dependence. In particular, the classification- and regression-based long-term models may output only discontinuation, which may be why the F1 score is low. Furthermore, as ad creatives tend to be discontinued due to their decreasing effectiveness of serving over time, our framework, which can consider the discontinuation probability at the previous time interval, is considered to function effectively.

5.6 Case Studies

5.6.1 Motivation and Tasks of the Case Studies

As our goal is to support ad operations (specifically ad discontinuation), we needed to verify how effective our framework is in practical cases. In Section 5.5, we confirmed that our framework performs well for predicting the ad creative discontinuation. However, we should evaluate with other criteria in addition to the traditional performance criteria in survival prediction such as CI.

In this section, we evaluate the availability of our framework on two real-world cases according to the following research questions using as yet unobserved data:

- RQ1: How does the efficiency of our framework for discontinued ad creatives compare to that of human operation?
- RQ2: How close is the order of the ad discontinuation based on the framework's predictions

to the actual manual order?

In the first case, we evaluate the cut-out (i.e., short-term) performance of our framework. Our purpose in the short-term is to reduce the operation loads on the operators from having to discontinue many ad creatives manually. We confirm how the efficiency changes when the ad is discontinued according to our framework, compared with that under manual operation. In the second case, we evaluate the wear-out (i.e., long-term) performance. Ad operators must prioritize tasks that create new compelling ad creatives to replace of long-running ads, as described in Section 5.3.3. Therefore, it is crucial to identify an appropriate order for ad discontinuation in the ad business.

For the evaluation, we use the best model identified in Section 5.5. When the hazard probability from the model exceeds 0.9 for the first time, we consider that the ad creative has been discontinued in the time interval. The two practical cases are analyzed for another period to confirm the time robustness with 400,000 ad creatives for three months, unlike in the offline experiments.[10]

5.6.2 Evaluation Criteria of the Case Studies

To evaluate our framework more closely against real-world ad operations, we conducted case studies in terms of CPA[11] and the order in which the ad creative is discontinued. The evaluation results with the criteria respond to the aforementioned research questions. We describe the details of these criteria below.

CPA Ratio: The Ratio of the Actual CPA to the Target CPA

We introduce practical metrics to assess the performance of ad creatives. One of the major metrics for deciding discontinuation in ad operation is the *CPA ratio*, which measures the efficiency of the CPA. We use this metric as the baseline for the experiments. The CPA ratio is defined as follows:

$$\texttt{CPA ratio} = \texttt{Actual CPA} \; / \; \texttt{Target CPA}. \tag{5.11}$$

[10] These data were sampled randomly to ensure confidentiality.

[11] While we understand the importance of the volume metric of the ads, we keep it private for business reasons because the metric is closely linked to real-world commercial systems.

Table 5.7: Comparison of the CPA ratios for predicting cut-out in the practical evaluation experiment. The CPA ratio is calculated as `Actual CPA/Acceptable CPA`. The closer the CPA ratio is to 1, the more profitable it is for the advertiser and the media. Human performance is calculated on the day the ad operator discontinues the ad creative.

Short-term	Single-task	Multi-task	Human performance
CPA ratio	1.13 ± 0.9	**1.12 ± 0.7**	1.18 ± 0.4

The actual CPA is calculated based on the number of sales and conversions and is the indicator of how much the ad creative costs compared with the conversions. The target CPA is the CPA that the advertiser wants. This metric measures CPA efficiency, which represents both the costs and volume. We should avoid simply evaluating a low CPA because it would reduce the exposure of the ad when the CPA is low. Advertisers are expected to minimize the volume within their budgets. We believe that the CPA ratio can be used to evaluate both the cost and volume.

The CPA ratio should not be high but not low either. When the metric is low, the setting of the cost can be satisfied, but the volume can be increased by allowing a higher actual CPA. The closer the CPA ratio is to 1, the more profitable it is for the advertiser and the media. This metric is actually used by ad operators to decide whether discontinue ad creatives. We assume that evaluation based on this metric is similar to professional operator decisions.

The Order of Ad Discontinuation

We evaluated the order of discontinuation for RQ2. To prepare for the sudden discontinuation of long-running ad creatives, it is necessary to ascertain to evaluate whether our framework can accurately predict the order in which ads are discontinued. To evaluate the order of ad discontinuation, we compared the normalized discounted cumulative gain (NDCG) [141]. The NDCG is calculated for ad creatives in the order in which they are discontinued by the ad operator and in the order in which our framework predicts discontinuations.

5.6.3 Evaluation Results of the Case Studies

Comparison of CPA Ratios

To answer RQ1, we evaluated the CPA efficiency if the ad operators decide to discontinue ad creatives according to our framework. Our framework predicted the discontinuation date based on only one day of data from serving. The CPA ratios of our framework are calculated based on the predicted discontinuation date, and the ratios for human performance are calculated based on the actual discontinuation date. We note that the discontinuations by human decision used the performance after the fist day, contrary to our framework. This experiment assumed the cut-out case, and thus, we evaluated the output of the short-term model, as described in Section 5.6.1.

Table 5.7 shows the comparison of the CPA ratios. Compared to the best model under the single-task setting, the best model under the multi-task setting successfully predicted the discontinuation with a CPA ratio close to one. Furthermore, the multi-task model shows better performance of discontinuation in terms of the CPA ratio than the human operator's decision. Overall, we confirmed that the single- and multi-task models have the same level of human performance.

Comparison of the Discontinuation-order Evaluation

To answer RQ2, we determined how close the discontinuation order predicted by our framework was to the actual discontinuation order. Correctly predicting which ads are likely to discontinue allows ad operators to set operational priorities. This experiment assumed the wear-out case; hence we evaluated the output of the long-term model, as described in Section 5.6.1.

To decide on a discontinuation order using our framework, we utilized the predicted discontinuation date and the hazard probability. To predict the wear-out case, our framework used only the first 10 days of data from serving. In each time interval, we sorted ad creatives based on the probability, and we concatenated the ordered ad creatives in each time interval.

To compare the evaluation results, we introduced two practical rule-based indicators, which are similar to practical operations. One was to sort by sales, namely, *sales order*. The business is seriously affected if an ad with high sales is discontinued. The other was to sort by the CPA ratio,

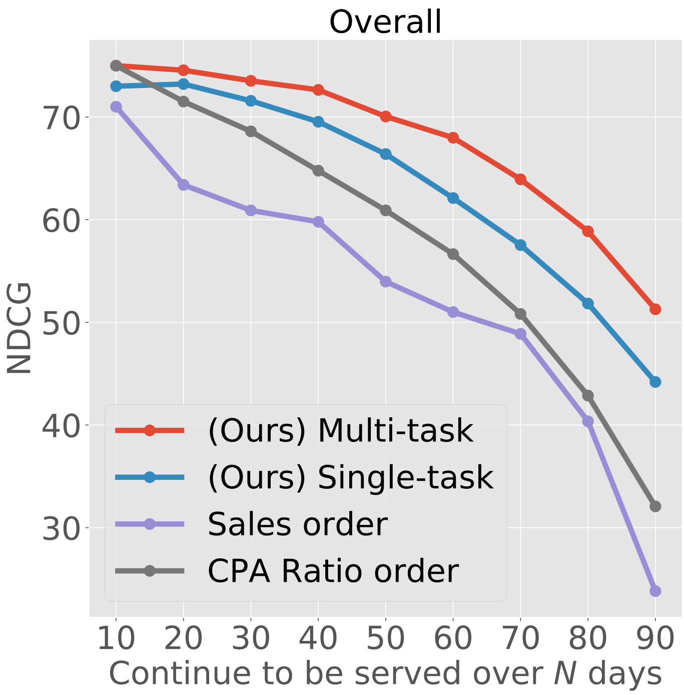


Figure 5.4: Comparison of the NDCG for the following prediction and actual discontinuation orders for each continuous serving days: by our framework (i.e., single-task and multi-task) and by the rule-based on the sales and the CPA ratio.

namely, the *CPA ratio order*, because ads with a high CPA ratio are easy to discontinue. These indicators are considered to be correlated with human decision-making.[12]

Figure 5.4 shows the NDCG of the prediction of our framework and the actual discontinuation orders for each continuous serving day. This experiment targeted ad creatives that have been served for more than 10 days. Each evaluation was conducted every 10 days for ad creatives within continuous serving days. For example, when the x-axis value was 20, the y-axis showed the NDCG value with ad creatives with 20 or more days of continuous serving. This evaluation determined how accurately we could predict future discontinuation.

[12] Ad operators usually monitor these indicators daily.

5.6.4 Discussion of the Case Studies

RQ1: Perspective of CPA ratio

In Table 5.7, we confirmed that both single- and multi-task models outperform the performance of the human operator. Because the performance of our models does not use unseen future data, which are used in the human performance, we posit that the same level of the performance is sufficient. This result suggests that, even if we decide the discontinuation using the prediction of our framework, the CPA efficiency does not get worse.

Clearly, the performance after ad discontinuation cannot be known, and thus, when the prediction result comes after the actual discontinuation date, the CPA ratio is the same as the actual CPA ratio. We posit that this problem does not affect our purpose, as our proposal is not to outperform but rather support human decision-making.

> *To answer* **RQ1***, when our framework discontinues an ad creative, the CPA efficiency is almost equal to human operation.*

RQ2: Perspective of the Order of Ad Discontinuation

Overall, our framework provided appropriate predictions of the discontinuation order compared with sales- and CPA ratio-based metrics. The CPA ratio and our framework exhibited the same performance for the first 10 days. However, when the number of serving days was increased, our framework performed better. In other words, the CPA ratio is suitable indicator for predicting which ad will be discontinued in the near future but cannot predict the discontinuation days in the long term, unlike our framework. Because it is important to plan priorities using long-term predictions, our framework plays an important role in prioritizing ad operations.

> *To answer* **RQ2***, our framework achieves a closer prediction of the order of the actual discontinuation than practical indicators.*

5.7 Conclusion

Ad creative discontinuation has a major role in ad operations. Although ad operation efficiency using ML methods is an active research topic, there have been few practical studies regarding how to predict the appropriate timing of ad creative discontinuation. To achieve this goal in a real-world environment, we proposed a multi-modal DNN-based framework for predicting the timing of the discontinuation of ad creatives that draws on the idea of survival prediction. Our framework has a loss function based on the hazard function and contains the following two simple but novel techniques: (1) a two-term estimation technique with multi-task learning and (2) a CTR-weighting technique for the loss function.

Evaluations using 1,000,000 real-world ad creatives confirmed that our two-term estimation technique significantly improves the prediction performance, and CTR-weighting technique further improved the performance. Compared with classification/regression frameworks, we observed that our framework performed notably better for predicting discontinuation. Additionally, for the two practical cases in the case studies, our framework achieved equivalent CPA efficiency with manual operation for short-term prediction and a higher NDCG with manual operation compared with other indicators using actual sales metrics (sales order, CPA ratio order) in long-term predictions. While the DNN-based architecture in our framework is not entirely novel, we addressed a crucial problem encountered in real-world ad operations and confirmed its feasibility in practical cases.

Our framework has been deployed in our production environment, and we engage in regular discussions with the operations team to figure out ways to further improve the operation efficiency. In the future, by analyzing the prediction results, we intend to establish the best practice to utilize the result and publish our framework for customers. Additionally, we intend to train a model that rewards the operator's discontinuation behavior based on reinforcement learning.

Chapter 6

Discussion

The goal of the research described in this dissertation is to improve the prediction performance and model interpretability through attention mechanisms from the basic and applied research perspectives. First, we discuss the scope of application of the research results. Next, we discuss the interpretability of the DL model, which is important in the future collaboration between humans and DL models, and the interpretability of the proposed methods. Finally, we discuss the impact of the proposed methods on subsequent studies published in response to our study.

6.1 Scope of Application of the Research Results

Throughout this dissertation, we used RNN-based models for the parts of the research results that we worked on, especially for NLP task. Such models were already widely used in practical applications when we started the research, and could be trained with relatively small computational resources. In light of the above, we considered that RNN-based evaluation was appropriate from the viewpoints of both basic and applied research at that time.

In the basic research, we evaluated the effectiveness of the proposed techniques in NLP tasks against RNNs with attention mechanisms. On the other hand, the proposed techniques are applicable to all models with attention mechanisms (including Transformer [6]). Although the mainstream has moved to Transformer in recent years, the original articles in this dissertation [104, 170] only stated that it was applicable.

The main reasons for not evaluating Transformer [6] or its follow-up models [70, 171] are: (1) the significant increase in computational complexity, (2) the unconfirmed vulnerability of the Transformer to perturbations in the attention mechanism, and (3) the unknown model interpretability properties of the Transformer. Regarding (1), Transformer was difficult to train at the time because it was a very deep model with a large number of parameters compared to conventional RNN-based models. For (2), we focused on the vulnerability in RNNs because the vulnerability was confirmed in Jain and Wallace [36], which we referred to, and was unknown in Transformer. Furthermore, for (3), since the RNN with attention mechanism has the mechanism before the final layer, we considered that the features that directly contribute to prediction are reflected in the attention weights. This setup allows for easier interpretation of prediction results compared to the Transformer model, which has a very complex attention mechanism.

While limited to evaluation by RNN for the above reasons, it is also considered to be an effective training technique for the Transformer. The proposed technique is expected to have the same model regularization effect as the original AT technique [49, 75]. For Transformer-based pre-trained models such as BERT [7], the proposed technique can be applied to the frozen encoder part. Then, the predictions can be made using the obtained the `CLS` tokens, which can be trained with a relatively small amount of computation. In fact, the ideas for our technique have been evaluated in the Transformer-based model in a subsequent study [172, 173] and are discussed in detail in Section 6.3.

In applied research, we developed frameworks for ad operations using a similar RNN model with the attention mechanism. Since our framework is generic, it is possible to change the text encoder responsible for NLP to a Transformer. On the other hand, the study related to interpretability in attention-based RNN models is more proven than in the Transformer model, which may have advantages, especially in businesses where mature technology is preferred.

6.2 Discussion of Interpretability in This Study

Following Arrieta et al. [16]'s definition of explainable AI, we proposed some techniques/frameworks to achieve the goal of "*building explainable AI that provides reasons and details that make understanding easier.*" Throughout this paper, we refer to "interpretability" in NLP tasks as methods that focus on each word in the input text and give the user an interpretation of the prediction.

6.2.1 Basic Research Perspectives

Recall that we focused on (1) trustworthiness and (5) confidence among the properties required of explainable AI. Methods that have been widely known to provide interpretations of model predictions, utilizing the attention and gradient, have often shown different results, and further study is needed to achieve more reliable model interpretations. We first examine the metrics for interpretability used in prior studies, and then discuss interpretability in Chapters 2 and 3.

Let us first discuss the evaluation using rank correlations, which was used in Jain and Wallace [36] to evaluate interpretability. Consider a situation where a relatively long sentence is input into the model. Intuitively, few words are considered important for solving the task for the entire input sentence. Thus, the majority of words are not necessarily required for prediction. Recall here that the gradient-based interpretation method may potentially present a noisy rationale. When the relationship between attention and the noisy gradient is evaluated by the rank correlation, the results are computed close to uncorrelated, even if the attention does provide a sufficient interpretation. From the above, we have examined the relationship between attention and gradient importance by Pearson's correlation coefficient in Chapters 2 and 3.

In Chapter 2, we showed that applying adversarial training to attention resulted in a higher correlation between attention and importance by gradient. Even in the baseline model, which was noted as "attention is not explanation," the Pearson's correlation between attention and gradient tended to be positive. The proposed technique enables the model to learn to pay more attention to words that are important for predicting the task. As a result, the noisy gradient becomes a clean gradient so that this kind of training occurs; attention is also clean as described in Chapter 2.

We believe that the interpretations resulting from such learning are more likely to be selected from words that are easier for humans to interpret, since both attention and gradient are activated for words that are more important for solving the task.

In Chapter 3, we showed that the correlation was further increased by utilizing unlabeled data to ensure diversity in the input data. Even in unlabeled data, there should be characteristic words that may contribute to predicting the task to be solved. The proposed technique enabled training that showed similar interpretations of both the words contributing to the prediction and the attention and gradient, even with unlabeled data that may be of a different nature than the task to be solved. The proposed technique enables training to present similar interpretations of both words and attention/gradients that contribute to predictions, even for unlabeled data of a different source than the problem to be solved. Additionally, we assessed the agreement with the manually annotated evidence [106]. The parts labeled as evidence by humans are sufficient information to solve the task and are highly interpretable in terms of *faithfulness*. Lipton [174] describes *faithfulness* as follows: "an explanation provided by a model is faithful if it reflects the information actually used by said model to come to a disposition." Although the model to which the proposed technique is applied has a lower apparent score because it does not have a mechanism for directly estimating evidence, it has a higher score than the baseline model and provides a more faithful explanation.

We argue for a different conclusion about interpretability for attention mechanisms, which previous studies have concluded that "attention is not an explanation [36]." The measure of interpretability by rank correlation used in [36] does not always adequately assess it. Our proposed technique also allows different interpretation methods to show similar results, which may give a more reliable interpretation. On the other hand, it is possible that attention and gradient may simultaneously represent interpretations that are not on target. It is still not clear whether the interpretations presented by these two methods are necessarily beneficial to humans. We believe that this point will require further investigation through error analysis.

6.2.2 Applied Research Perspectives

Among the properties required of explainable AI, we focused on (4) informativeness in the applied research. In actual decision-making in advertising operations, it is important to provide information so that appropriate decisions can be made. Most of the tasks in the computational advertising field addressed in this study rely on human labor, and many of them have not been formulated in the first place. We believe that this research has made a significant contribution to the research field, even if only by formulating and providing a preliminary interpretation of ad operations.

In Chapter 4, we provided the creator with keywords of interest to the user in an interpretable form to assist in the creation of effective ad creatives. The ad creator can check the words that contribute to conversion prediction from multiple ad creative proposals prepared in advance, based on the ad attributes, using our conditional attention mechanism weights. We can find keywords that cannot be found by simple aggregation (e.g., the correlation between keywords and the number of conversions). This may be because the proposed framework combines these elements, as attention is given to words in the context of the text, including the attributes of the advertisement. Our proposed framework allows the creators to provide important interpretations in order to create attractive ad creatives that directly contribute to sales.

In Chapter 5, for the ad discontinuation prediction, which had not been formulated before, the use of a survival prediction framework that can predict the discontinuation probabilities in detail and keywords that contribute to discontinue are provided in an interpretable form. Note that the challenge of using ML models to solve the problem of predicting ad discontinuation in the first place is itself new to academia. Therefore, decision support using conventional ML models (e.g., logistic regression, decision trees, SVM, etc.) for ad discontinuation prediction is in itself novel and has a significant impact on business sales as well as the work of ad operators. In contrast to the prediction of discontinuation through the classification and regression frameworks, this research allows for more detailed discontinuation probabilities to be provided in the framework of survival analysis so that ad operators can extract the information they need to make decisions when discontinuing ads. Since the operators can make discontinuation decisions based on more detailed fine-grained predictions, the proposed framework is considered to provide useful operational support.

In addition to the above, we also provided interpretable keywords that contribute to the predicted discontinuation of ineffective ad creatives. While survival prediction has provided information on the probability of discontinuing at each time interval, it is still often unclear what factors contribute to discontinuing in the ad creative. Since the decision to discontinue an ad creative that has become less effective is difficult to make, it is possible to identify the words that contribute to the prediction by checking the weights by the learned attention mechanism within the proposed framework. While a variety of factors determine the discontinuation of an ad creative, we expect to see different trends in short- and long-term discontinuation in the ad text. Specifically, it is possible to check word expressions that the user was not interested in during the short-term discontinuation. Such keywords are useful in determining the decision, but they can also be interpreted as keywords that would immediately discontinue an ad creative in terms of creating an ad creative. At long-term discontinuation, the proposed framework can present word expressions that are no longer in season. If served over a long period of time, they contain keywords that are no longer in line with current events and are thought to increase wear-out. Unfortunately, due to disclosure constraints, quantitative evaluation of interpretability by specific attention within the dissertation has not been directly possible. On the other hand, as described in Chapter 5, it is true that the proposed framework provides daily predictions to ad operators, and decisions are made based on these predictions. While discontinuing ineffective ads has as much business impact as creating highly effective ad creatives, we believe that the interpretations provided by the attention mechanism solve a part of the real-world problems in the computational advertising field.

6.3 The Impact of the Proposed Methods on Subsequent Research

As described above, we evaluated our proposals using the RNN-based model, however, the outcomes of our basic and applied research are applicable to models with attention mechanisms in general.

We describe below how our proposals has had an academic impact on papers written after our paper was published.

6.3.1 Basic Research Perspectives

The idea of introducing adversarial training to attention mechanism proposed in the basic research has greatly influenced subsequent research. Similar ideas were applied to RoBERTa [70] to demonstrate its effectiveness in temporal reasoning tasks, although our research was limited to evaluate based on the RNNs. In temporal reasoning tasks, words such as *before* and *after* are important to predict, however, Ribeiro et al. [175] has shown that even these simple temporal segments often fail, resulting in poor performance. As in the examples, similar but different words are usually placed close together in the word embedding space. The reason for this is that DL models including the word embedding learn co-occurrence patterns between words using a large corpus of text. As a result, words that are commonly used in similar contexts tend to be placed close together in the embedding space, even if they have different meanings. They agree with our idea and aim to improve generalization performance by adversarial training on such words and their embeddings.

Inspired by our ideas, Hu et al. [176] proposed a new attention mechanism that makes the perturbation-vulnerable attention mechanism more stable and explainable. Their proposed Stable and Explainable Attention (SEAT) is trained to have the following properties: (1) its prediction distribution is enforced to be close to the distribution based on the vanilla attention; (2) its top-k indices have large overlaps with those of the vanilla attention; (3) it is robust w.r.t perturbation. Their SEAT, and extension of our idea, has been evaluated in comparison with BERT in addition to RNN, and achieves even better prediction performance and interpretability than our methods.

Our ideas are also powerful in real-world applications. It plays a role in predictive and cognitive methods, especially regarding mental health [177, 178] and fake news [48], where metaphorical expressions are often used. The former is working on health mention classification based on our ideas, which is difficult to recognize because of its various metaphorical expressions [177, 178]. The latter constructs a prediction model with reference to our idea for fake news that is carefully written to deceive people by incorporating metaphorical expressions [48].

Our idea of extending to semi-supervised learning has been proven effective in the medical and biological fields [179, 180]. Due to low annotation costs, the number of data is generally limited in these fields. In the medical field, our idea if part of a new framework for predicting ASD (autism spectrum disorder) subjects and typical controls based on fMRI data [179]. Additionally, in the biological field, where the number of data is limited and annotation is difficult in many cases, our idea is being considered for application with the expectation of a kind of data augmentation effect [180].

On the other hand, the idea of AT for attention has certain limitations When calculating adversarial perturbations, data are first forwarded to the model, and then the direction of the adversarial perturbation is calculated based on the obtained gradient, which generally increases the computational complexity. This is a drawback of using our method in the pre-training phase of recent large-scale language models, as it increases the complexity. We believe that using our method in the fine-tuning phase will further improve the generalization performance of the model by acting a data augmentation role with AT for attention, as well as the effect of the pre-trained model on a small number of data.

6.3.2 Applied Research Perspectives

The computational advertising research focused on in this study is a relatively new field and is still in its early stages of development. In particular, while there have been some attempts at analytical aspects of ML/DL-related research focused on ad creatives [118, 155], few studies have addressed problems solving in actual business operations. Since around 2019, there has been an increase in efforts related to the operation support of ad creatives [61,62,156,181], including our ideas [142]. We believe that our efforts have created a certain research trend in the sub-domain of the computational advertising field: the application of ML/DL on ad creative.

While there are a variety of operations related to ad creative, research has started to focus on operational support related to highly effective ads. Specifically, there is support for creating ad creatives with high serving effectiveness [142, 181–183], automatic generation of descriptions that lead to product purchases [156, 184, 185], and recommendation of keywords and designs that users find attractive [61, 62]. In search engine advertising, a system that can automatically generate advertisement text using reinforcement learning has been proposed [156], and the generation of ad creatives personalized to each user is realized [184]. Our research was ahead of these efforts, providing advance predictions of effective ad creatives and word-by-word interpretations that could assist in their creation.

Although ad operations related in ineffective one can have a significant business impact, we are the first to focus on supporting the discontinuation of ad creatives that have become ineffective [186]. Conventionally, the decision to discontinue an ad creative was partially supported by using a bandit algorithm to predict the CTR and CVR of the ads [143,145,157,158], however, it was difficult to set a threshold for discontinuing ads based on these metrics, as described in Chapter 5. The proposed framework has constructed an operationally useful system that directly supports the decision to discontinue an ad creative. Our major academic contribution is that we formulated these studies in the framework of survival time analysis, rather than in the framework of classification and regression, which is generally conceivable. It is meaningful that we formulated the business problem in the framework of survival analysis, rather than in the framework of classification and regression, which is generally conceivable. We believe this effort to be a major academic contribution.

Chapter 7

Conclusion

This dissertation discussed both basic and applied research on how attention mechanisms improve the performance and interpretability of machine learning models, especially in NLP-related tasks. Focusing on the black-box nature of deep learning models, which have recently achieved significant results in various fields, we attempted to develop new training techniques and models that help interpret the prediction results for the input words. For the further development of applications of deep learning models, this black box nature is a serious barrier to analyzing the behavior of the models and using them in situations where prediction failures are not tolerated. To the best of our knowledge, where there is a trade-off between prediction performance and interpretability, we have pioneered a new technique that aims to improve both.

From the basic research perspective, we proposed new training techniques to improve both prediction performance and model interpretability by employing adversarial perturbations to the attention mechanisms. The attention mechanisms are currently an indispensable key component of deep learning models, and the vulnerability of these mechanisms to noise and perturbation can seriously affect the prediction performance and interpretability of the model. Our experiments demonstrated that our proposed AT for attention mechanism achieves a preferable performance gain than conventional models in terms of the prediction performance and interpretability of the model. Furthermore, to effectively utilize unlabeled data, our proposed VAT for attention mechanisms significantly improved both the metrics and performed effectively even when using unlabeled data from a source other than the labeled data. These training techniques are model-independent with attention mechanisms and are scalable to any model.

From the applied research perspective, we proposed interpretable prediction frameworks to support ad operations with large-scale ad data for real-world applications. Because ad data is generally much larger, more diverse, and noisier than the toy data used in basic research, the effectiveness of models with attention mechanisms for such data remains unclear. We attempted to construct models that can be interpreted from the aspects of both predicting the serving effect of ad creatives that have not yet been served in advance and predicting the appropriate timing for discontinuing ad creatives whose effect is declining, which have a particular impact on business. The former model for predicting more effective ad creatives can learn attention that appropriately reflects user interest while predicting highly imbalanced CTRs and CVRs, suggesting that it is possible to support the creation of effective ad creatives based on such attention weight. In the latter model for predicting less effective ad creatives, we first formulated the problem of discontinuing ad creatives and constructed a multimodal DNNs to predict the discontinuation with high accuracy. Furthermore, we empirically confirmed that the performance of discontinuation is comparable to that of humans.

Although the proposals within this dissertation are mainly validated using RNN models, we believe that the proposed methods are generic and can be applied to all models with attention mechanisms that will emerge in the future. In interpreting the prediction evidence for each input word, we discussed the application of the proposals entail a certain degree of interpretability. Our ideas are providing a positive effect on those that follow, and are informing the emergence of further basic research into the development of the ideas, as well as efforts to support real-world operations.

Because there are several other options to provide explanations to DNNs besides attention mechanisms, we will confirm the applicability of the ideas in this method to these options in the future. Additionally, we will discuss the interpretability of the models in business applications, such as recommendation systems, to see if our method can be applied to such systems.

Appendix A

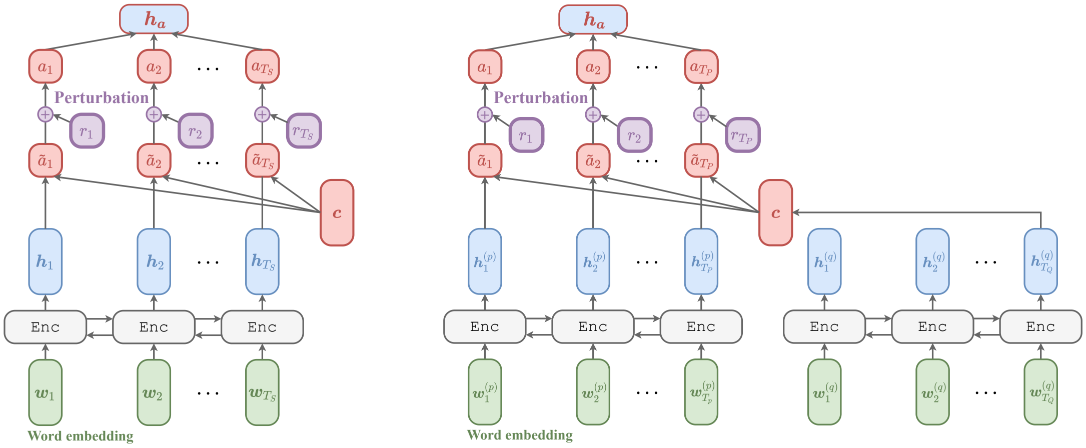
Appendix for Common Model Architecture

A.1 Common Model Architecture

In this appendix, we introduce the common model architecture to which our training techniques were applied. The common model is a practical and widely used RNN-based model, and its performance has been compared in extensive experiments focusing on attention mechanisms [36,104,114]. As the settings differ for single- and pair-sequence tasks, we defined a model for each task, as illustrated in Fig. A.1, and the details are described below.

A.1.1 Model for Single-Sequence Tasks

Fig. A.1a presents the model for the single-sequence tasks, such as text classification. The input of the model was a word sequence of one-hot encoding $X_S = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{T_S}) \in \mathbb{R}^{|V| \times T_S}$, where $|V|$ and T_S are the vocabulary size and the number of words in the sequence, respectively. Let $\boldsymbol{w}_t \in \mathbb{R}^d$ be a d-dimensional word embedding corresponding to $\boldsymbol{x}_t$. Each word was represented with the word embeddings to obtain $(\boldsymbol{w}_t)_{t=1}^{T_S} \in \mathbb{R}^{d \times T_S}$. The word embeddings were encoded with a bidirectional



(a) Single-sequence model

(b) Pair-sequence model

Figure A.1: Common models for applying the proposed training technique: (a) single-sequence model for the text classification task, and (b) pair-sequence model for QA and NLI tasks In (a), the input of the model was the word embeddings, $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{T_S}\}$ that were associated with the input sentence X_S. In (b), the inputs were the word embeddings $\{\boldsymbol{w}_1^{(p)}, \cdots, \boldsymbol{w}_{T_P}^{(p)}\}$ and $\{\boldsymbol{w}_1^{(q)}, \cdots, \boldsymbol{w}_{T_Q}^{(q)}\}$ from the two input sequences, X_P and X_Q, respectively. These inputs were encoded into hidden states through a bidirectional encoder (**Enc**). In conventional models, the worst-case perturbation $\boldsymbol{r}$ is added to the word embeddings. In our Attention VAT and iVAT, $\boldsymbol{r}$ is computed and added to the attention score $\tilde{\boldsymbol{a}}$ to improve the prediction performance and model interpretability, even with unlabeled data.

RNN (BiRNN)-based encoder **Enc** to obtain the m-dimensional hidden state:

$$\boldsymbol{h}_t = \mathbf{Enc}(\boldsymbol{w}_t, \boldsymbol{h}_{t-1}), \tag{A.1}$$

where $\boldsymbol{h}_0$ is the initial hidden state and it is regarded as a zero vector. Following [36] and [104], we used the additive formulation of attention mechanisms [33] to compute the attention score for the t-th word $\tilde{a}_t$, which is defined as:

$$\tilde{a}_t = \boldsymbol{c}^\top \tanh(W\boldsymbol{h}_t + \boldsymbol{b}), \tag{A.2}$$

where $W \in \mathbb{R}^{d' \times m}$ and $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{d'}$ are the model parameters. Subsequently, the attention weights $\boldsymbol{a} \in \mathbb{R}^T$ for all words were computed from the attention scores $\tilde{\boldsymbol{a}} = (\tilde{a}_t)_{t=1}^{T_S}$, as follows:

$$\boldsymbol{a} = (a_t)_{t=1}^{T_S} = \mathrm{softmax}(\tilde{\boldsymbol{a}}). \tag{A.3}$$

The weighted instance representation $\boldsymbol{h_a}$ was calculated using the attention weights $\boldsymbol{a}$ and hidden state $\boldsymbol{h}_t$, as follows:

$$\boldsymbol{h_a} = \sum_{t=1}^{T_S} a_t \boldsymbol{h}_t. \tag{A.4}$$

Finally, $\boldsymbol{h_a}$ was fed to a dense layer **Dec**, and the output activation function σ was used to obtain the following predictions:

$$\hat{\boldsymbol{y}} = \sigma(\mathbf{Dec}(\boldsymbol{h_a})) \in \mathbb{R}^{|\boldsymbol{y}|}, \tag{A.5}$$

where σ is a sigmoid function and $|\boldsymbol{y}|$ is the class label set size.

A.1.2 Model for Pair-Sequence Tasks

Fig. A.1b presents the model for pair-sequence tasks, such as QA and NLI. The input of the model was $X_P = (\boldsymbol{x}_t^{(p)})_{t=1}^{T_P} \in \mathbb{R}^{|V| \times T_P}$ and $X_Q = (\boldsymbol{x}_t^{(q)})_{t=1}^{T_Q} \in \mathbb{R}^{|V| \times T_Q}$, where T_P and T_Q are the number of words in each sentence. Furthermore, X_P and X_Q represent the paragraph and question in QA tasks and the hypothesis and premise in NLI tasks, respectively. We used two separate BiRNN encoders ($\mathbf{Enc}_P$ and $\mathbf{Enc}_Q$) to obtain the hidden states $\boldsymbol{h}_t^{(p)} \in \mathbb{R}^m$ and $\boldsymbol{h}_t^{(q)} \in \mathbb{R}^m$:

$$\boldsymbol{h}_t^{(p)} = \mathbf{Enc}_P(\boldsymbol{w}_t^{(p)}, \boldsymbol{h}_{t-1}^{(p)}); \quad \boldsymbol{h}_t^{(q)} = \mathbf{Enc}_Q(\boldsymbol{w}_t^{(q)}, \boldsymbol{h}_{t-1}^{(q)}), \tag{A.6}$$

where $\boldsymbol{h}_0^{(p)}$ and $\boldsymbol{h}_0^{(q)}$ are the initial hidden states, and they are regarded as zero vectors. Subsequently, we computed the attention weight $\tilde{a}_t$ of each word of X_P as follows:

$$\tilde{a}_t = \boldsymbol{c}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_t^{(p)} + \boldsymbol{W}_2 \boldsymbol{h}_{T_Q}^{(q)} + \boldsymbol{b}), \tag{A.7}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d' \times m}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d' \times m}$ denote the projection matrices, and $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{d'}$ are the parameter vectors. Similar to Eq. (A.3), the attention weight a_t could be calculated from $\tilde{a}_t$. The presentation was obtained from the sum of the words in X_P.

$$\boldsymbol{h_a} = \sum_{t=1}^{T_P} a_t \boldsymbol{h}_t^{(p)} \tag{A.8}$$

was fed to a **Dec**, following which a softmax function was used as σ to obtain the prediction (in the same manner as in Eq. (A.5)).

A.1.3 Model Training with Attention Mechanisms

Let $X_{\tilde{\boldsymbol{a}}}$ be an input sequence with an attention score $\tilde{\boldsymbol{a}}$, where $\tilde{\boldsymbol{a}}$ is a concatenated attention score for all t. The conditional probability of the class $\boldsymbol{y}$ was modeled as $p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents all model parameters. We minimized the following negative log-likelihood as a loss function for the model parameters to train the model:

$$\mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \boldsymbol{\theta}) = -\log p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}};\boldsymbol{\theta}). \tag{A.9}$$

Appendix B

Appendix for Tasks and Datasets

B.1 Tasks and Dataset

B.1.1 Binary Classification

The following datasets were used for evaluation. The Stanford Sentiment Treebank (SST) [74][1] was used to ascertain positive or negative sentiment from a sentence. IMDB Large Movie Reviews (IMDB) [91][2,3] was used to identify positive or negative sentiment from movie reviews. 20 Newsgroups (20News) [92][4] were used to ascertain the topic of news articles as either baseball (set as a negative label) or hockey (set as a positive label). The AG News (AGNews) [93][5] was used to identify the topic of news articles as either world (set as a negative label) or business (set as a positive label).

[1] https://nlp.stanford.edu/sentiment/trainDevTestTrees_PTB.zip

[2] https://s3.amazonaws.com/text-datasets/imdb_full.pkl

[3] https://s3.amazonaws.com/text-datasets/imdb_word_index.json

[4] https://ndownloader.figshare.com/files/5975967

[5] The dataset can be found on Xiang Zhang's Google Drive.

B.1.2 Question Answering

The following datasets were used for evaluation. The CNN news article corpus (CNN news) [94][6] was used to identify answer entities from a paragraph. The bAbI dataset (bAbI) [95][7] contains 20 different question-answer tasks, and we considered three tasks: (task 1) basic factoid question answered with a single supporting fact, (task 2) factoid question answered with two supporting facts, and (task 3) factoid question answered with three supporting facts. The model was trained for each task.

B.1.3 Natural Language Inference

The following datasets were used for evaluation. The Stanford Natural Language Inference (SNLI) [96][8] is used to identify whether a hypothesis sentence entails, contradicts, or is neutral concerning a given premise sentence. Multi-Genre NLI (MultiNLI) [97][9] uses the same format as SNLI and is comparable in size, but it includes a more diverse range of text, as well as an auxiliary test set for cross-genre transfer evaluation.

B.2 Details of Evaluation Criteria

B.2.1 Correlation between Attention and Gradient-based Word Importance

We computed how the attention weighted obtained through our VAT-based technique agree with the importance of words calculated by gradients [24]. This evaluation follows Jain and Wallace [36]. The correlation τ_g is defined from the attention $\boldsymbol{a} \in \mathbb{R}^T$ and the gradient-based word importance $\boldsymbol{g} \in \mathbb{R}^T$ as follows:

$$\tau_g = \text{PearsonCorr}(\boldsymbol{a}, \boldsymbol{g}). \tag{B.1}$$

[6] The dataset can be found on Deep Mind Q&A Google Drive.

[7] https://research.fb.com/downloads/babi/

[8] https://nlp.stanford.edu/projects/snli/snli_1.0.zip

[9] https://www.nyu.edu/projects/bowman/multinli/multinli_1.0.zip

The gradient-based word importance $\boldsymbol{g} = (g_t)_{t=1}^{T}$ is calculated as follows:

$$g_t = |\sum_{i=1}^{|V|} \mathbb{1}[X_{it} = 1] \frac{\partial \boldsymbol{y}}{\partial X_{it}}|, \forall t \in [1, T], \tag{B.2}$$

where X_{it} is the t-th one-hot encoded word for the i-th vocabulary in $X \in R^{|V| \times T}$, and T is the number of words in the sequence.

Appendix C

Appendix for Adversarial Training for Attention Mechanism

C.1 Implementation Detail

For all datasets, we either used pretrained GloVe [187] or fastText [188] word embedding with 300 dimensions except the bAbI dataset. For the bAbI dataset, we trained 50 dimensional word embeddings from scratch during training. We used a one-layer LSTM as the encoder with a hidden size of 64 for the bAbI dataset and 256 for the other datasets. All models were regularized using L_2 regularization (10^{-5}) applied to all parameters. We trained the model using the maximum likelihood loss utilizing the Adam [140] optimizer with a learning rate of 0.001.

Appendix D

Appendix for Virtual Adversarial Training for Attention Mechanism

D.1 Implementation Details

We implemented all of the training techniques using the AllenNLP library with Interpret [99, 100]. We evaluated the test set only once in all experiments. The experiments were conducted on an Ubuntu PC with a GeForce GTX 1080 Ti GPU. Our implementation is based on the one published by [104].[1] Note that the model that was used in our experiment had a small number of parameters compared to recent models; therefore, the execution speed of the model was also fast.

D.1.1 Supervised Classification Task

We used pretrained fastText [188] word embedding with 300 dimensions and a one-layer bi-directional long short-term memory (LSTM) [98] as the encoder with a hidden size of 256 for the supervised settings. We used a sigmoid function as the output activation function. All models were regularized using L_2 regularization (10^{-5}) that was applied to all of the parameters. We trained the model

[1] https://github.com/shunk031/attention-meets-perturbation

using the maximum likelihood loss and the Adam optimizer [140] with a learning rate of 0.001. All of the experiments were conducted at $\lambda = 1$.

We searched for the best hyperparameter ϵ from $[0.01, 30.00]$ following [104]. The Allentune library [101] was used to adjust ϵ, and we decided on the value of the hyperparameter ϵ based on the validation score.

D.1.2 Semi-supervised Classification Task

We used the same pretrained fastText and encoder for the semi-supervised settings. We again used the Adam optimizer [140] with the same learning rate as that in the supervised classification task. The same hyperparameter search was performed as in the supervised settings. All of the experiments were conducted at $\lambda = 1$ in the semi-supervised settings.

We also performed the same preprocessing according to Chen et al. [105] using the same unlabeled data. We determined the amount of unlabeled data N_{ul} based on the validation score for each benchmark dataset. We reported the test score of the model with the highest validation score.

Appendix E

Appendix for Ad Discontinuation Prediction

E.1 Detail of Features

Our framework used the four feature areas that constitute the ad creative: text features, categorical features, image features, and numerical features. The details of these input features and encoding processes are described as follows.

E.1.1 Text features

Text features were extracted from the title and description of an ad creative. Specifically, words were embedded from the title and the description and encoded by a text encoder using a recurrent neural network (RNN)-based model. For the RNN-based model, we used bi-directional long short-term memory (LSTM) [98] in our implementation. We can expect the extracted text features to consider the context of the ad creative text. Subsequently, title embedding and description embedding were calculated as the last hidden state of each text encoder as: $\boldsymbol{h}^{\text{title}} \in \mathbb{R}^{d_{\text{title}}}$ and $\boldsymbol{h}^{\text{desc}} \in \mathbb{R}^{d_{\text{desc}}}$. Finally, we concatenated these embeddings to text embeddings:

$$\boldsymbol{e}_1 = [\boldsymbol{h}^{\text{title}}; \boldsymbol{h}^{\text{desc}}] \in \mathbb{R}^{d_{\text{title}}+d_{\text{text}}}. \tag{E.1}$$

E.1.2 Categorical features

Categorical features were extracted from the gender of the serving target and the genre of the ad creative. Specifically, the gender of the distribution target was extracted as $\boldsymbol{x}^{\text{gender}} \in \mathbb{R}^{d_{\text{gender}}}$, which was one-hot encoded. Additionally, the genre of the ad creative, which was embedded by entity embedding method [189], was extracted as $\boldsymbol{x}^{\text{genre}} \in \mathbb{R}^{d_{\text{genre}}}$. We can expect the extracted categorical features to consider the demographic information of the target users. Finally, we combined these embeddings in categorical embeddings:

$$\boldsymbol{e}_2 = [\boldsymbol{x}^{\text{gender}}; \boldsymbol{x}^{\text{genre}}] \in \mathbb{R}^{d_{\text{gender}}+d_{\text{genre}}}. \tag{E.2}$$

E.1.3 Image features

Image features were extracted from an image of an ad creative. Specifically, image embedding $\boldsymbol{e}_3$ was obtained by encoding $d_h \times d_w$ sized ad image $\boldsymbol{x}_3$ using the ResNet34 [5] pretrained by ImageNet:

$$\boldsymbol{e}_3 = \text{ResNet34}(\boldsymbol{x}_3). \tag{E.3}$$

We can expect the extracted image feature to consider the characteristics of the ad image.

E.1.4 Numerical features

Numerical features include the number of impressions, clicks, and conversions and the CPA for an ad creative. The numerical features consist of statistical and time-series features.

The statistical feature is composed of statistical information about target date t. The CTR was obtained from the number of impressions and the number of clicks, and the CVR was obtained from the number of clicks and the number of conversions. We used the following features as the statistical feature: the number of impressions x_{s1}, clicks $x_{s2}(t)$, conversions $x_{s3}(t)$, the CTR $x_{s4}(t)$, the CVR $x_{s5}(t)$, and the CPA $x_{s6}(t)$. These features were normalized through logarithmic transformation. Finally, we concatenated these features as statistical embeddings:

$$\boldsymbol{e}_4^{\text{Stat}}(t) = [x_{s1}(t);\ x_{s2}(t);\ x_{s3}(t);\ x_{s4}(t);\ x_{s5}(t);\ x_{s6}(t)]. \tag{E.4}$$

Time-series features were extracted from impressions and clicks accumulated from the first day of serving to the target date t based on statistical features. Specifically, from the accumulated impressions and clicks, those encoded by the history encoder with different LSTM models were obtained as an impression history embedding $\boldsymbol{h}^{\text{imp}}$ and a click history embedding $\boldsymbol{h}^{\text{click}}$, respectively. These embeddings are the last hidden states encoded by the history encoder of the serving performance data up to time t. Then, we combined these features into time-series embeddings:

$$\boldsymbol{e}_4^{\text{Time}}(t) = [\boldsymbol{h}^{\text{imp}}; \boldsymbol{h}^{\text{click}}] \in \mathbb{R}^{d_{\text{imp}}+d_{\text{click}}}. \tag{E.5}$$

Finally, we combined statistical embedding $\boldsymbol{e}_4^{\text{Stat}}(t)$ and time-series embedding $\boldsymbol{e}_4^{\text{Time}}(t)$ into numerical embeddings:

$$\boldsymbol{e}_4(t) = [\boldsymbol{e}_4^{\text{Stat}}(t);\ \boldsymbol{e}_4^{\text{Time}}(t)]. \tag{E.6}$$

We can expect the extracted numerical features to consider the both statistical and time-series properties of the ad creative.

E.2 Implementation Details

The texts of the ad creatives, written in Japanese, were split into words using MeCab [137], which is a type of morphological analysis for Japanese texts. The custom dictionary mecab-ipadic-neologd[1], which includes various neologisms, was used as the dictionary.

[1] `https://github.com/neologd/mecab-ipadic-neologd`

We used pre-trained word2vec [139] to initialize the word embedding layer in our framework. For the text encoder, d_{title} and d_{desc} were set to 16. For the history encoder, d_{imp} and d_{click} were set to 10. For an image encoder, we extracted 512 dimensional feature vectors. The feature vectors were then converted using a shallow MLP to $d_{\text{img}} = 16$ dimensional image features.

We used pycox[2] to implement the base of our survival prediction framework. The mini-batch size was set to be 32, and the number of epochs was set to be 50. Adam [140] was used for the model optimization.

In all experiments, we performed the evaluation on the test set only once. The experiments were conducted on an Ubuntu PC with one GeForce GTX 1080 Ti GPU. As our framework is relatively small, we ran the case studies on the CPU.

Regarding the deployment of our framework, we dockerized the implementation of our framework to work on Amazon Web Service (AWS). The prediction from our framework is output daily to a Google spreadsheet for discussion with the ad operation team.

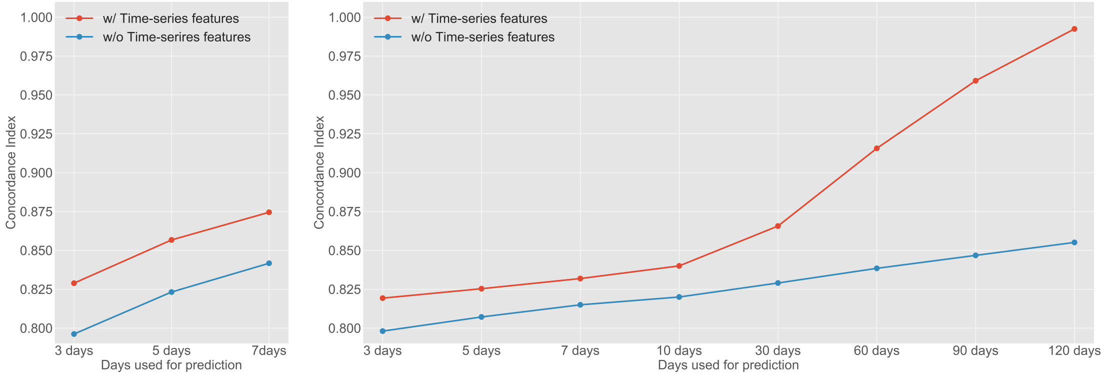
E.3 Classification and Regression Frameworks

As binary classification models, we built models that predicted discontinuation after 3, 7, 30, and 90 days. We evaluated the predictions of the model specific to each time interval. In training the binary classification model, we used the binary cross entropy loss to train the model. For a fair comparison, we applied a CTR-weighting technique based on our idea to the comparison models.

Regarding the regression model, we built models for short-term and long-term. The regression models were trained on data with time intervals adjusted for short- and long-term intervals. For training the regression model, we used mean squared error loss to train the model. We used the time interval that was closest to the prediction made by the regression model as the final prediction. Similar to binary classification, we applied the CTR-weighting technique for a fair comparison.

The same architecture as our single-task model (as described in Section 5.5) was used in these models, and the same features were used as inputs. Note here that we considered an evaluation based

[2] `https://github.com/havakv/pycox`



(a) Short-term

(b) Long-term

Figure E.1: Comparison of the effects of time-series features with daily changes in ad discontinuation prediction task. For the short- and the long-term models, the more days that can be used for prediction, the higher the performance. In the long-term model, when used for prediction for 30 days or more, the concordance index is dramatically improved compared with the case where the time-series feature is not used.

on the mean squared error, but we believe that the metrics and results presented in Section 5.5 are more suitable for comparison.

E.4 Effects of Time-series Data

Figure E.1 shows the effect on prediction performance of the time-series features, changing the number of days used for prediction. Figure E.1a shows the change in the CI in the short-term model. As described above, the model to which the time-series feature is input achieves a higher score compared with the model to which no time-series feature is input. Additionally, as the number of days used for prediction increased, the prediction performance of each model improved. Figure E.1b shows the change in the CI in the long-term model. Similar to the short-term model, in the long-term setting, the model with time-series features has a higher score compared with the model without time-series features. Therefore, it appears that the effectiveness of the time-series features for prediction increases with time.

Bibliography

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and regression trees. belmont, ca: Wadsworth,” International Group, vol. 432, no. 151-166, p. 9, 1984.

[3] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18–28, 1998.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. of CVPR, 2016, pp. 770–778.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. of the 30th International Conference on Neural Information Processing Systems, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), ser.

Association for Computational Linguistics (ACL), 2019, pp. 4171–4186. [Online]. Available: http://dx.doi.org/10.18653/v1/N19-1423

[8] D. Castelvecchi, "Can we open the black box of ai?" Nature News, vol. 538, no. 7623, p. 20, 2016.

[9] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe, "Explanation capabilities of production-based consultation systems," STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, Tech. Rep., 1977.

[10] W. R. Swartout, "Explaining and justifying expert consulting programs," in Computer-assisted medical decision making. Springer, 1985, pp. 254–271.

[11] M. Sharp, R. Ak, and T. Hedberg Jr, "A survey of the advancing use and development of machine learning in smart manufacturing," Journal of manufacturing systems, vol. 48, pp. 170–179, 2018.

[12] Y. Zhang, X. Chen et al., "Explainable recommendation: A survey and new perspectives," Foundations and Trends® in Information Retrieval, vol. 14, no. 1, pp. 1–101, 2020.

[13] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas, "Deep learning in robotics: Survey on model structures and training strategies," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 1, pp. 266–279, 2020.

[14] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," Journal of Imaging, vol. 6, no. 6, p. 52, 2020.

[15] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," IEEE Transactions on Intelligent Transportation Systems, 2021.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," Information fusion, vol. 58, pp. 82–115, 2020.

[17] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," IEEE access, vol. 6, pp. 52 138–52 160, 2018.

[18] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM computing surveys (CSUR), vol. 51, no. 5, pp. 1–42, 2018.

[19] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," Nature machine intelligence, vol. 2, no. 1, pp. 56–67, 2020.

[20] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," SN Applied Sciences, vol. 3, no. 2, pp. 1–12, 2021.

[21] O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis, "Convex optimization techniques for fitting sparse gaussian graphical models," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 89–96.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. of the 34th International Conference on Machine Learning, ser. JMLR. org, vol. 70, 2017, pp. 3319–3328.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in 2nd International Conference on Learning Representations, ICLR, Workshop Track Proceedings, 2013. [Online]. Available: https://arxiv.org/abs/1312.6034

[25] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies," Artificial Intelligence, vol. 294, p. 103459, 2021.

[26] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in International conference on machine learning. PMLR, 2017, pp. 3145–3153.

[27] M. Aubakirova and M. Bansal, "Interpreting neural networks to improve politeness comprehension," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2035–2041.

[28] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer' s dementia by interpreting neural models," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 701–707.

[29] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," ICML Workshop on Visualization for Deep Learning, 2017.

[30] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in International conference on machine learning. PMLR, 2017, pp. 1885–1894.

[31] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in 6th International Conference on Learning Representations, ICLR, Conference Track Proceedings, 2017.

[32] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3429–3437.

[33] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd International Conference on Learning Representations, ICLR 2015, 2015.

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.

[35] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4190–4197.

[36] S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), ser. Association for Computational Linguistics (ACL), 2019, pp. 3543–3556. [Online]. Available: http://dx.doi.org/10.18653/v1/N19-1357

[37] M. G. Kendall, "A new measure of rank correlation," Biometrika, vol. 30, no. 1/2, pp. 81–93, 1938.

[38] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), ser. Association for Computational Linguistics (ACL), 2019, pp. 11–20.

[39] S. Serrano and N. A. Smith, "Is attention interpretable?" in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, ser. Association for Computational Linguistics (ACL), 2019, pp. 2931–2951. [Online]. Available: http://dx.doi.org/10.18653/v1/P19-1282

[40] C. Grimsley, E. Mayfield, and J. R. Bursten, “Why attention is not explanation: Surgical intervention and causal reasoning about neural models,” in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 1780–1790.

[41] Y. Wang, M. Huang, and L. Zhao, “Attention-based LSTM for aspect-level sentiment classification,” in Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2016, pp. 606–615. [Online]. Available: http://dx.doi.org/10.18653/v1/D16-1058

[42] J. Wang, J. Tuyls, E. Wallace, and S. Singh, “Gradient-based analysis of nlp models is manipulable,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 247–258.

[43] K. Mrini, F. Dernoncourt, Q. H. Tran, T. Bui, W. Chang, and N. Nakashole, “Rethinking self-attention: Towards interpretability in neural parsing,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 731–742.

[44] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro, “Timeshap: Explaining recurrent models through sequence perturbations,” in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2565–2573.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2020.

[46] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in 3rd International Conference on Learning Representations, ICLR, Conference Track Proceedings, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

[47] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in 2nd International Conference on

Learning Representations, ICLR, Conference Track Proceedings, 2013. [Online]. Available: https://arxiv.org/abs/1312.6199

[48] A. Tariq, A. Mehmood, M. Elhadef, and M. U. G. Khan, "Adversarial training for fake news classification," IEEE Access, vol. 10, pp. 82 706–82 715, 2022.

[49] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 8, pp. 1979–1993, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2858821

[50] K. Dave, V. Varma et al., "Computational advertising: Techniques for targeting relevant ads," Foundations and Trends® in Information Retrieval, vol. 8, no. 4–5, pp. 263–418, 2014.

[51] I. IAB, "Internet advertising revenue report: Full-year 2021 results, april 2022," Internet Advertising Revenue Report, 2022. [Online]. Available: https://www.iab.com/wp-content/uploads/2022/04/IAB_Internet_Advertising_Revenue_Report_Full_Year_2021.pdf

[52] D. Chakrabarti, D. Agarwal, and V. Josifovski, "Contextual advertising by combining relevance with click feedback," in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 417–426.

[53] M. Richardson, E. Dominowska, and R. Ragno, "Predicting Clicks: Estimating the Click-Through Rate for New Ads," in Proc. of the 16th International World Wide Web Conference(WWW-2007), 2007.

[54] J. Chen, B. Sun, H. Li, H. Lu, and X.-S. Hua, "Deep ctr prediction in display advertising," in Proceedings of the 2016 ACM on Multimedia Conference, 2016, pp. 811–820.

[55] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in Proc. of the 10th ACM Conference on Recommender Systems, 2016, pp. 191–198.

[56] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir et al., “Wide & deep learning for recommender systems,” in Proc. of the 1st Workshop on Deep Learning for Recommender Systems, 2016, pp. 7–10.

[57] S. Rendle, “Factorization machines,” in Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010, pp. 995–1000.

[58] Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin, “Field-aware factorization machines for CTR prediction,” in Proc. of the 10th ACM Conference on Recommender Systems, 2016, pp. 43–50.

[59] Y. Juan, D. Lefortier, and O. Chapelle, “Field-aware factorization machines in a real-world online advertising system,” in Proc. of the 26th International Conference on World Wide Web Companion, 2017, pp. 680–688.

[60] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, “Deepfm: a factorization-machine based neural network for ctr prediction,” CoRR arXiv:1703.04247, 2014.

[61] S. Mishra, M. Verma, and J. Gligorijevic, “Guiding creative design in online advertising,” in Proc. of RecSys, 2019, pp. 418–422.

[62] Y. Zhou, S. Mishra, M. Verma, N. Bhamidipati, and W. Wang, “Recommending themes for ad creative design via visual-linguistic representations,” in Proceedings of The Web Conference 2020, 2020, pp. 2521–2527.

[63] J. Gope and S. K. Jain, “A survey on solving cold start problem in recommender systems,” in 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017, pp. 133–138.

[64] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” in Proc. of the 5th International Conference on Learning Representations, ICLR, Conference Track Proceedings, 2017. [Online]. Available: https://openreview.net/forum?id=BJC_jUqxe&noteId=BJC_jUqxe

[65] X. He and D. Golub, "Character-level question answering with attention," in Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2016, pp. 1598–1607. [Online]. Available: http://dx.doi.org/10.18653/v1/D16-1166

[66] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2016, pp. 2249–2255. [Online]. Available: http://dx.doi.org/10.18653/v1/D16-1244

[67] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2015, pp. 1412–1421. [Online]. Available: http://dx.doi.org/10.18653/v1/D15-1166

[68] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2015, pp. 379–389. [Online]. Available: http://dx.doi.org/10.18653/v1/D15-1044

[69] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," CoRR preprint arXiv:1611.02770, 2016. [Online]. Available: https://arxiv.org/abs/1611.02770

[70] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," CoRR preprint arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[71] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[72] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," CoRR preprint arXiv:2009.06732, 2020. [Online]. Available: https://arxiv.org/abs/2009.06732

[73] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," Neurocomputing, vol. 307, pp. 195–204, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.04.027

[74] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2013, pp. 1631–1642. [Online]. Available: https://www.aclweb.org/anthology/D13-1170/

[75] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in 5th International Conference on Learning Representations, ICLR, Conference Track Proceedings, 2016. [Online]. Available: https://openreview.net/forum?id=r1X3g2_xl

[76] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, "Interpretable adversarial perturbation in input embedding space for text," in Proc. of the 27th International Joint Conference on Artificial Intelligence, ser. AAAI Press, 2018, pp. 4323–4330. [Online]. Available: https://dl.acm.org/doi/10.5555/3304222.3304371

[77] M. Yasunaga, J. Kasai, and D. Radev, "Robust multilingual part-of-speech tagging via adversarial training," in Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), ser. Association for Computational Linguistics (ACL), 2018, pp. 976–986. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-1089

[78] Y. Wang and M. Bansal, "Robust machine comprehension models via adversarial training," in Proc. of the 2018 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), ser. Association for Computational Linguistics (ACL), 2018, pp. 575–581. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-2091

[79] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," CoRR preprint arXiv:1612.08220, 2016. [Online]. Available: https://arxiv.org/abs/1612.08220

[80] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in International Conference on Learning Representations, no. 2019, 2019. [Online]. Available: https://openreview.net/forum?id=SyxAb30cY7

[81] T. Itazuri, Y. Fukuhara, H. Kataoka, and S. Morishima, "What do adversarially robust models look at?" CoRR preprint arXiv:1905.07666, 2019. [Online]. Available: https://arxiv.org/abs/1905.07666

[82] T. Zhang and Z. Zhu, "Interpreting adversarially trained convolutional neural networks," in International Conference on Machine Learning. PMLR, 2019, pp. 7502–7511. [Online]. Available: http://proceedings.mlr.press/v97/zhang19s.html

[83] B. Wang, J. Gao, and Y. Qi, "A theoretical framework for robustness of (deep) classifiers against adversarial examples," CoRR preprint arXiv:1612.00334, 2016. [Online]. Available: https://arxiv.org/abs/1612.00334

[84] K. Liu, X. Liu, A. Yang, J. Liu, J. Su, S. Li, and Q. She, "A robust adversarial training approach to machine reading comprehension," in Proc. of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 8392–8400. [Online]. Available: https://doi.org/10.1609/aaai.v34i05.6357

[85] S. Barham and S. Feizi, "Interpretable adversarial training for text," CoRR preprint arXiv:1905.12864, 2019. [Online]. Available: https://arxiv.org/abs/1905.12864

[86] Q. Zhu, L. Cui, W.-N. Zhang, F. Wei, and T. Liu, "Retrieval-enhanced adversarial training for neural response generation," in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019, pp. 3763–3773. [Online]. Available: http://dx.doi.org/10.18653/v1/P19-1366

[87] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: https://jmlr.org/papers/v15/srivastava14a.html

[88] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning, 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[89] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1681–1691. [Online]. Available: http://dx.doi.org/10.3115/v1/P15-1162

[90] D. Shimada, R. Kotani, and H. Iyatomi, "Document classification through image-based character embedding and wildcard training," in Proc. of IEEE International Conference on Big Data. IEEE, 2016, pp. 3922–3927. [Online]. Available: https://doi.org/10.1109/BigData.2016.7841067

[91] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ser. Association for Computational Linguistics (ACL), vol. 1, 2011, pp. 142–150. [Online]. Available: https://www.aclweb.org/anthology/P11-1015/

[92] K. Lang, “Newsweeder: Learning to filter netnews,” in Machine Learning Proceedings. Elsevier, 1995, pp. 331–339. [Online]. Available: https://doi.org/10.1016/B978-1-55860-377-6.50048-7

[93] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in Proc. of the 28th International Conference on Neural Information Processing Systems, ser. MIT Press, vol. 1, 2015, pp. 649–657. [Online]. Available: https://dl.acm.org/doi/10.5555/2969239.2969312

[94] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” in Proc. of the 28th International Conference on Neural Information Processing Systems, ser. MIT Press, vol. 1, 2015, pp. 1693–1701. [Online]. Available: https://dl.acm.org/doi/10.5555/2969239.2969428

[95] J. Weston, A. Bordes, S. Chopra, and T. Mikolov, “Towards AI-complete question answering: A set of prerequisite toy tasks,” in 4th International Conference on Learning Representations, ICLR, Conference Track Proceedings, 2016. [Online]. Available: https://arxiv.org/abs/1502.05698

[96] S. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2015, pp. 632–642. [Online]. Available: http://dx.doi.org/10.18653/v1/D15-1075

[97] A. Williams, N. Nangia, and S. R. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” in Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers), ser. Association for Computational Linguistics (ACL), 2017. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-1101

[98] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[99] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," CoRR preprint arXiv:1803.07640, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/W18-2501

[100] E. Wallace, J. Tuyls, J. Wang, S. Subramanian, M. Gardner, and S. Singh, "AllenNLP Interpret: A framework for explaining predictions of NLP models," in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, ser. Association for Computational Linguistics (ACL), 2019, pp. 7–12. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-3002

[101] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, "Show your work: Improved reporting of experimental results," in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), ser. Association for Computational Linguistics (ACL), 2019, pp. 2185–2194. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-1224

[102] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran, "Towards transparent and explainable attention models," in Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, ser. Association for Computational Linguistics (ACL), 2020, pp. 4206–4216. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.acl-main.387

[103] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, "Did the model understand the question?" in Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ser. Association for Computational Linguistics (ACL), 2018, pp. 1896–1906.

[104] S. Kitada and H. Iyatomi, "Attention meets perturbations: Robust and interpretable attention with adversarial training," IEEE Access, vol. 9, pp. 92 974–92 985, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3093456

[105] L. Chen, W. Ruan, X. Liu, and J. Lu, "Seqvat: Virtual adversarial training for semi-supervised sequence labeling," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ser. Association for Computational Linguistics (ACL), 2020, pp. 8801–8811. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.acl-main.777

[106] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ser. Association for Computational Linguistics (ACL), 2019, pp. 4443–4458. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.acl-main.408

[107] Z. Li, C. Feng, M. Wu, H. Yu, J. Zheng, and F. Zhu, "Adversarial robustness via attention transfer," Pattern Recognition Letters, vol. 146, pp. 172–178, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865521000982

[108] Z. Yi, J. Yu, Y. Tan, and Q. Wu, "Fine-tuning more stable neural text classifiers for defending word level adversarial attacks," Applied Intelligence, pp. 1–18, 2022.

[109] Y. Wu, D. Bamman, and S. Russell, "Adversarial training for relation extraction," in Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2017, pp. 1778–1783.

[110] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," in Proceedings of the 15th Annual Conference of the International Speech Communication Association, ser. International Speech Communication Association (ISCA), 2014, pp. 2635–2639. [Online]. Available: https://arxiv.org/abs/1312.3005

[111] J. An, K. Wang, H. Sun, C. Cui, W. Li, and C. Ma, "Attention virtual adversarial based semi-supervised question generation," Concurrency and Computation: Practice and Experience, vol. 34, no. 10, p. e6797, 2022.

[112] K. Dai, X. Li, X. Huang, and Y. Ye, "Sentatn: learning sentence transferable embeddings for cross-domain sentiment classification," Applied Intelligence, pp. 1–14, 2022.

[113] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," CoRR preprint arXiv:1909.07913, 2019.

[114] C. Meister, S. Lazov, I. Augenstein, and R. Cotterell, "Is sparse attention more interpretable?" in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 122–129.

[115] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace, "Inferring which medical treatments work from reports of clinical trials," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 3705–3717. [Online]. Available: http://dx.doi.org/10.18653/v1/N19-1371

[116] O. Chapelle, E. Manavoglu, and R. Rosales, "Simple and scalable response prediction for display advertising," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 4, p. 61, 2015.

[117] S. Thomaidou, K. Liakopoulos, and M. Vazirgiannis, "Toward an integrated framework for automated development and optimization of online advertising campaigns," Intelligent Data Analysis, vol. 18, no. 6, pp. 1199–1227, 2014.

[118] S. Thomaidou, K. Leymonis, and M. Vazirgiannis, "GrammAds: Keyword and ad creative generator for online advertising campaigns," in Digital Enterprise Design and Management 2013, 2013, pp. 33–44.

[119] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern, “Visual Appearance of Display Ads and Its Effect on Click Through Rate,” in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 495–504.

[120] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam, “Multimedia features for click prediction of new ads in display advertising,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 777–785.

[121] N. I. Bruce, B. Murthi, and R. C. Rao, “A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement,” Journal of marketing research, vol. 54, no. 2, pp. 202–218, 2017.

[122] S. Tokui, K. Oono, S. Hido, and J. Clayton, “Chainer: a next-generation open source framework for deep learning,” in Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015. [Online]. Available: http://learningsys.org/papers/LearningSys_2015_paper_33.pdf

[123] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, “xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems,” in Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018.

[124] W. Liu, R. Tang, J. Li, J. Yu, H. Guo, X. He, and S. Zhang, “Field-aware Probabilistic Embedding Neural Network for CTR Prediction,” in Proc. of the 12th ACM Conference on Recommender Systems, 2018, pp. 412–416.

[125] S. Punjabi and P. Bhatt, “Robust Factorization Machines for User Response Prediction,” in Proc. of the 2018 World Wide Web Conference, 2018, pp. 669–678.

[126] H. Yang, Q. Lu, A. X. Qiu, and C. Han, “Large Scale CVR Prediction through Dynamic Transfer Learning of Global and Local Features,” in Proc. of the 5th International Workshop

on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications at KDD 2016, 2016, pp. 103–119.

[127] Q. Lu, S. Pan, L. Wang, J. Pan, F. Wan, and H. Yang, "A Practical Framework of Conversion Rate Prediction for Online Display Advertising," in Proc. of the ADKDD'17, 2017, pp. 9:1–9:9.

[128] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ser. Association for Computational Linguistics (ACL), 2016, pp. 1480–1489.

[129] R. Caruana, "Multitask learning," Machine learning, vol. 28, no. 1, pp. 41–75, 1997.

[130] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol. 12, no. 8, pp. 2493–2537, 2011.

[131] T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in International Conference on Learning Representations, 2016.

[132] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in European Conference on Computer Vision, 2014, pp. 94–108.

[133] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3707–3715.

[134] X. Chu, W. Ouyang, W. Yang, and X. Wang, "Multi-task recurrent neural network for immediacy prediction," in Proc. of the IEEE international conference on computer vision, 2015, pp. 3352–3360.

[135] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in Advances in neural information processing systems Workshop, 2014.

[136] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," CoRR arXiv:1207.0580, 2012.

[137] T. Kudo, "Mecab: Yet another part-of-speech and morphological analyzer," 2006.

[138] S. Toshinori, "Neologism dictionary based on the language resources on the web for mecab," 2015. [Online]. Available: https://github.com/neologd/mecab-ipadic-neologd

[139] M. Suzuki, K. Matsuda, S. Sekine, N. Okazaki, and K. Inui, "A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles," IEICE Transactions on Information and Systems, vol. E101.D, no. 1, pp. 73–81, 2018.

[140] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR preprint arXiv:1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[141] K. Järvelin and J. Kekäläinen, "Cumulated Gain-based Evaluation of IR Techniques," ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, 2002.

[142] S. Kitada, H. Iyatomi, and Y. Seki, "Conversion prediction using multi-task conditional attention networks to support the creation of effective ad creatives," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2069–2077.

[143] T. Maehara, A. Narita, J. Baba, and T. Kawabata, "Optimal bidding strategy for brand advertising," in Proc. of IJCAI, 2018, p. 424–432.

[144] X. Yang, Y. Li, H. Wang, D. Wu, Q. Tan, J. Xu, and K. Gai, "Bid optimization by multivariable control in display advertising," in Proc. of KDD, 2019, p. 1966–1974.

[145] K. D. Doan, P. Yadav, and C. K. Reddy, "Adversarial factorization autoencoder for look-alike modeling," in Proc. of CIKM, 2019, p. 2803–2812.

[146] M. Shatnawi and N. Mohamed, “Statistical techniques for online personalized advertising: A survey,” in Proc. of SAC, 2012, p. 680–687.

[147] M. F. Gensheimer and B. Narasimhan, “A scalable discrete-time survival model for neural networks,” PeerJ, vol. 7, 2019.

[148] D. R. Cox, “Regression models and life-tables,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pp. 187–202, 1972.

[149] D. Van den Poel and B. Lariviere, “Customer attrition analysis for financial services using proportional hazard models,” European journal of operational research, vol. 157, no. 1, pp. 196–217, 2004.

[150] L. Dirick, G. Claeskens, and B. Baesens, “Time to default in credit scoring using survival analysis: a benchmark study,” Journal of the Operational Research Society, vol. 68, no. 6, pp. 652–665, 2017.

[151] N. Barbieri, F. Silvestri, and M. Lalmas, “Improving post-click user engagement on native ads via survival analysis,” in Proc. of WWW, 2016, pp. 761–770.

[152] W. C.-H. Wu, M.-Y. Yeh, and M.-S. Chen, “Predicting winning price in real time bidding with censored data,” in Proc. of KDD, 2015, pp. 1305–1314.

[153] A. Kalra, C. Wang, C. Borcea, and Y. Chen, “Reserve price failure rate prediction with header bidding in display advertising,” in Proc. of KDD, 2019, pp. 2819–2827.

[154] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, “Evaluating the yield of medical tests,” Jama, vol. 247, no. 18, pp. 2543–2546, 1982.

[155] S. Thomaidou, I. Lourentzou, P. Katsivelis-Perakis, and M. Vazirgiannis, “Automated snippet generation for online advertising,” in Proc. of CIKM, 2013, p. 1841–1844.

[156] J. W. Hughes, K.-h. Chang, and R. Zhang, “Generating better search engine text advertisements with deep reinforcement learning,” in Proc. of KDD, 2019, pp. 2269–2277.

[157] W. Zhang, Y. Zhang, B. Gao, Y. Yu, X. Yuan, and T.-Y. Liu, “Joint optimization of bid and budget allocation in sponsored search,” in Proc. of KDD, 2012, p. 1177–1185.

[158] J. Shen, S. C. Geyik, and A. Dasdan, “Effective audience extension in online advertising,” in Proc. of KDD, 2015, p. 2099–2108.

[159] K. Ren, W. Zhang, K. Chang, Y. Rong, Y. Yu, and J. Wang, “Bidding machine: Learning to bid for directly optimizing profits in display advertising,” IEEE Trans. on Knowledge and Data Engineering, vol. 30, no. 4, pp. 645–659, 2018.

[160] H. Liu, D. Pardoe, K. Liu, M. Thakur, F. Cao, and C. Li, “Audience expansion for online social network advertising,” in Proc. of KDD, 2016, p. 165–174.

[161] C. Pechmann and D. W. Stewart, “Advertising repetition: A critical review of wearin and wearout,” Current Issues and Research in Advertising, vol. 11, no. 1-2, pp. 285–329, 1988.

[162] D. Moriwaki, K. Fujita, S. Yasui, and T. Hoshino, “Fatigue-aware ad creative selection,” in Proc. of WSDM workshop SUM, 2020.

[163] L. Tang, R. Rosales, A. Singh, and D. Agarwal, “Automatic ad format selection via contextual bandits,” in Proc. of CIKM, 2013, p. 1587–1594.

[164] O. Chapelle, “Modeling delayed feedback in display advertising,” in Proc. of KDD, 2014, pp. 1097–1105.

[165] C. Zhong and R. Tibshirani, “Survival analysis as a classification problem,” CoRR preprint arXiv:1909.11171, 2019.

[166] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network,” BMC medical research methodology, vol. 18, no. 1, p. 24, 2018.

[167] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, “Deephit: A deep learning approach to survival analysis with competing risks,” in Proc. of AAAI, vol. 32, no. 1, 2018.

[168] L. D. Fisher and D. Y. Lin, "Time-dependent covariates in the cox proportional-hazards regression model," Annual review of public health, vol. 20, no. 1, pp. 145–157, 1999.

[169] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer et al., "Random survival forests," The annals of applied statistics, vol. 2, no. 3, pp. 841–860, 2008.

[170] S. Kitada and H. Iyatomi, "Making attention mechanisms more robust and interpretable with virtual adversarial training," Applied Intelligence, pp. 1–16, 2022.

[171] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.

[172] L. Pereira, F. Cheng, M. Asahara, and I. Kobayashi, "Alice++: Adversarial training for robust and effective temporal reasoning," in Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, 2021, pp. 377–386.

[173] L. K. Pereira, K. Duh, M. Cheng, Fei andAsahara, and K. Ichiro, "Attention-focused adversarial training for robust temporal reasoning," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7352–7359.

[174] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue, vol. 16, no. 3, pp. 31–57, 2018.

[175] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," arXiv preprint arXiv:2005.04118, 2020.

[176] L. Hu, Y. Liu, N. Liu, M. Huai, L. Sun, and D. Wang, "Seat: Stable and explainable attention," arXiv preprint arXiv:2211.13290, 2022.

[177] P. I. Khan, S. A. Siddiqui, I. Razzak, A. Dengel, and S. Ahmed, "Improving health mention classification of social media content using contrastive adversarial training," IEEE Access, vol. 10, pp. 87 900–87 910, 2022.

[178] P. I. Khan, I. Razzak, A. Dengel, and S. Ahmed, "A novel approach to train diverse types of language models for health mention classification of tweets," arXiv preprint arXiv:2204.06337, 2022.

[179] X. Deng, J. Zhang, R. Liu, and K. Liu, "Classifying asd based on time-series fmri using spatial–temporal transformer," Computers in Biology and Medicine, vol. 151, p. 106320, 2022.

[180] J. C. Ditz, B. Reuter, and N. Pfeifer, "Comic: Convolutional kernel networks for interpretable end-to-end learning on (multi-) omics data," arXiv preprint arXiv:2212.02504, 2022.

[181] S. Mishra, M. Verma, Y. Zhou, K. Thadani, and W. Wang, "Learning to create better ads: Generation and ranking approaches for ad creative refinement," in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2653–2660.

[182] S. Mishra, M. Kuznetsov, G. Srivastava, and M. Sviridenko, "Visualtextrank: Unsupervised graph-based content extraction for automating ad text to image search," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3404–3413.

[183] S. Mishra, C. Hu, M. Verma, K. Yen, Y. Hu, and M. Sviridenko, "Tsi: an ad text strength indicator using text-to-ctr and semantic-ad-similarity," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4036–4045.

[184] H. Kamigaito, P. Zhang, H. Takamura, and M. Okumura, "An empirical study of generating texts for search engine advertising," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021, pp. 255–262.

[185] Y. S. Kanungo, S. Negi, and A. Rajan, "Ad headline generation using self-critical masked language model," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021, pp. 263–271.

[186] S. Kitada, H. Iyatomi, and Y. Seki, "Ad creative discontinuation prediction with multi-modal multi-task neural survival networks," Applied Sciences, vol. 12, no. 7, p. 3594, 2022.

[187] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, ser. Association for Computational Linguistics (ACL), 2014, pp. 1532–1543. [Online]. Available: http://dx.doi.org/10.3115/v1/D14-1162

[188] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017. [Online]. Available: http://dx.doi.org/10.1162/tacl_a_00051

[189] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," CoRR preprint arXiv:1604.06737, 2016.