要約文章に基づいた画像生成による挿絵自動生成システム

Saito, Yuya / 齋藤, 優也

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
17
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2022-03-24
(URL)
https://doi.org/10.15002/00025262

要約文章に基づいた画像生成による挿絵自動生成システム
Automatic Illustration Generation System by Image Generation Based on Summarized Text

齋藤　優也
Yuya Saito
法政大学情報科学研究科情報科学専攻
E-mail: yuya.saito.5g@stu.hosei.ac.jp

Abstract

In this study, I propose a system for generating illustrations from the book to assist in image formation when reading. In the proposed system, each paragraph of the book is summarized by PEGASUS, and an illustration is generated using VQGAN-CLIP of Text-to-Image model based on the summarized text on the top of the original content. In the process of generating an illustration, a ranking method is proposed for determining parameters (the seed value and learning times) based on the CLIP image score and the loss value of image generation. In the experiment, VQGAN trained with 120,000 images and CLIP trained with 400 million sets of images and texts are used as pre-learning models. For the evaluation, Taro Urashima, a Japanese fairy tale with 19 paragraphs, is used to test the model, and the evaluation on the generated illustrations are conducted. There are 12 evaluators for questionnaire. The evaluation is based on the following three aspects: "Image Clarity", "Image Quality", and "Text to Relevance". From the viewpoint of "Image Clarity" and "Text to Relevance", it is highly evaluated, so the content of the text can be reflected in the illustration. From the results, it is concluded the automatic generated illustrations can reasonably support the understanding of its corresponding text.

1. まえがき

近年，様々なデータ化が進み，読書をする媒体も紙などの書籍から電子化された電子書籍へと変化している．スマートフォンの普及が進んだ背景もあり，場所や時間を問わずに簡単に読書を行うことが可能となっている．電子書籍を提供する Kindle など様々なサービスでは，読書をサポートするための様々な機能が搭載されている．分からない単語の意味の検索を行う単語検索機能や，メモをしたい箇所を保存しておくハイライト機能，文字を読みやすくするための拡大機能などがサポートされている．これらの機能は紙媒体の時代から行われていることを簡単に行えるようにしただけではあるが，電子書籍ならではの機能として，文章を自動で要約して読書の時間を短縮するような機能も注目を浴びている．そこで，本研究では文章の内容を解析して，読書のサポートをする機能として，自動で本文から挿絵を生成することによって，読者のイメージ形成をサポートする機能を提案する．小説における挿絵の影響は，読者に共通のイメージを与えるためのものとして有効である[1].

本研究によって提案されるシステムでは，各パラグラフに対して挿絵をそれぞれ生成する．各パラグラフの文章の内容を読み取るために，PEGASUS[2]による文章要約を利用し，パラグラフの内容を 1~2 文にまとめ上げる．要約された文章を基に Text-to-Image と呼ばれる文章から，その内容に見合った画像を生成する技術で挿絵を生成する．今回は VQGAN-CLIP[3, 4]と呼ばれるモデルを利用する．挿絵を生成する際は複数枚生成し，ランキングを行うことで最適な挿絵を生成できるようにする．本研究では，特に物語の文章から画像の生成を行う．物語の多くは元々文章と画像がペアとなっているため，挿絵として画像を生成することに適している文章であると考えられるからである．最終的に物語の各パラグラフに対して 1 枚ずつ挿絵を生成し，提案システムによる物語全体の挿絵を自動で生成することが目的である．

2. 関連研究

提案システムに関連する既存研究として，文章要約と文章から画像生成を行う Text-to-Image タスクに関する研究を紹介する．

2.1. 文章要約

自然言語処理分野において，文章要約は入力とする文章から簡潔でかつ正確な要約を出力することを目的に研究が行われてきた．一般的に，文章要約には，入力の文章内から重要な部分を断片的に得て要約する抽出型要約と，入力の内容に沿った文章を生成することによる抽象的要約の 2 通りに大きく分けることができる．

抽出型要約には，文章全体のトピックを算出し，そのトピックにあった文章を抽出するトピックベースの LSA(Latent Semantic Analysis)[5]を利用した手法など様々な手法が取り入れられてきた．

一方で抽象的要約においては，人が要約を作るように，文章の意味を理解したうえで適切な要約を生成する手法である．抽象型要約を実現するためには，長文の内容理解能力，文章の情報整理能力，新たな文章を生成する能力など，自然言語処理分野でも難しい処理が求められる．これらの複雑な処理を実現するために，機械学習が多く

Supervisor: Prof.　Runhe Huang

取り入れられている．文章データと要約文章を含む高品質な教師付きデータセットが多く公開されていることも機械学習が盛んになっている理由である．特に，RNN や Transformer を応用した Encoder-Decoder モデルが主流となっている．さらに，近年自然言語処理分野で大きな注目を浴びている BERT[6]を応用した PEGASUS によって，高品質な要約を生成することが可能となった．

2.2. Text-to-Image

Text-to-Image タスクとは，任意のテキストから，そのテキストの内容に沿った画像を生成するタスクのことである．従来，このタスクではあらかじめテキストと画像のペアを用意し，入力のテキストに対してそれらの画像を組み合わせる手法がとられてきた．しかし，機械学習分野の発展や豊富なデータセットの取得が可能となったことから，汎用的な画像が生成可能となった．このタスクは，アート生成やコンピュータによるデザインの補助などに応用できると考えられている．

様々な手法が研究されているタスクであるが，近年では特に，深層学習モデルを取り入れた研究が盛んになりつつある．深層学習モデルの一つである敵対性ネットワーク(GAN)を応用し，テキストを入力として与えることで直接画像を生成する手法がある．その他にも機械学習モデルでは，OpenAI から発表された DALL-E[7]と呼ばれるモデルが注目を浴びている．自然言語処理分野で革新的な性能を示した GPT-3[8]を応用し，膨大なデータセットを学習させることで汎用的な画像を生成することが可能となった．これにより，モデルが未知のデータに対しても画像生成が可能となるゼロショット学習が可能となり，優れた性能を示すことができた．

3. 提案システム

本研究では，読者のイメージ形成補助のための挿絵自動生成を目的としたシステムを提案する．本研究による貢献として，小説や物語などから挿絵を生成することで，読者に内容のイメージを持たせることができ，スムーズな読書が出来ると考えられる．また，提案システムを用いた挿絵の自動生成によって，AI によるアート生成分野として，機械によるデザインの補助に貢献があると考えられる．

また，システム内で行う挿絵生成において，生成画像の質を向上させるために，パラメータ決定のための画像ランキング方式を新たに提案する．画像生成に使用するパラメータの探索方法として，画像をランキング化し，それらを基にパラメータ決定を行う．

これらを実現するために，対象となる物語や小説などの本文を入力テキストとして，そのテキスト内容に見合った画像を生成するための挿絵自動生成システムを提案する．提案システムの全体図は図 1 となる．

入力されたテキストを前処理によって整形したのち，機械学習モデルによって文章要約を行い，要約文章を獲得する．文章を要約することによって，物語の本文の内容を簡潔にまとめ上げることが出来ると考えた．その後，Text-to-Image のモデルによって，要約文章から画像を生成する．生成される画像には，要約文章の内容が反映さ

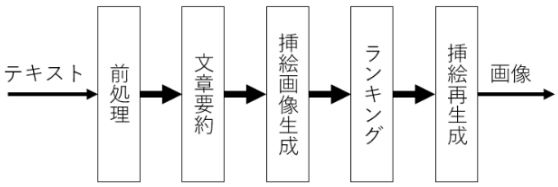

図 1：提案システムのアーキテクチャ

れるため，挿絵としての役割を果たすことが期待できる．

また，挿絵の質を向上させるために，複数枚画像を生成したのち，新たに提案するランキング方式によって，画像生成に必要なパラメータを決定する．最後に，決定したパラメータを用いて出力とする挿絵を再生成する．

3.1. 前処理

対象とする絵本の本文に対して前処理を行う．前処理は自然言語処理分野で一般的な手法で，文章中のノイズを除去し，パフォーマンスを安定させることが目的である．今回行う前処理は，以下の表 1 のとおりである．

表 1：前処理一覧

ストップワードの除去
“a”や”the”，”is”などの情報量の少ない単語をテキストから取り除くことで，重要な単語に焦点を当てることが可能となる．
不要な記号の除去
文章中に含まれる，”!”や”?”などの記号を除去し，”．”に変換させる．文章内の曖昧性を減らし，パフォーマンスを安定させる．

日本語の物語などを対象とするときは，一度英語に翻訳を行ってから，表 1 の前処理を行う．後述するが，提案システムに使用するモデルは英語で学習させているため，入力を統一するために，英語に翻訳する必要がある．英語翻訳には Google が提供する，Google 翻訳を使用する．Google 翻訳は，大量の文章を一度に翻訳できる点，また精度がよく無料で使用できる点で優れている．

3.2. 文章要約 による重要文の生成

文章要約では，Text-to-Image の画像生成モデルに与える入力の文章を生成することを目的とする．一般的な小説や物語などにおいて，挿絵が含まれるときは，文章やパラグラフなどの内容を表す重要な場面であると考えられる．つまり挿絵を生成するためには,文章やパラグラフの内容を理解する必要がある.本研究と同様な挿絵を推薦する方法として,物語の場面を分割して,その中から重要な文章を取り出す手法がある.しかし,この手法では,文章の中で重要な文章を取り出して,その文章から挿絵を生成する研究[9]がある.しかし,この手法は重要文の抽出精度があまりよくなかった.そのため,文章の挿絵を生成する文章としては適していないと考えた.そこで，提案手法において,各パラグラフの内容を,要約文章を生成することで簡潔に内容をまとめる手法をとる.文章要約を用いるこ

とで，パラグラフ内の文章の内容を，重要な文章として 1~2 文程度にまとめあげることができる．そして，そこで得た要約された文章はパラグラフの重要な内容を示すため，挿絵に適した文章が生成できると考える．

文章要約の手法としては， 2.1 で示したように様々な手法がある．その中でも，機械学習を用いた抽象型の文章要約モデルを，提案システムでは使用する．抽象型の文章要約モデルは，その学習方法は人間が文章要約する方法に似ているため，学習が上手くいけば，要約の精度も高い．今回の抽象型の文章要約モデルは，Google が 2019 年に提案した PEGASUS という事前学習モデルを使用する．PEGASUS は，文章要約のタスクにおける 12 個のベンチマークで最高性能を達成することが可能となった．よって，本研究の文章要約によって，文章内の内容を反映させた文章を生成するという目的を達成できると考えた．

以上の方法で，文章の内容をまとめ上げた要約文章を生成する．また今回の提案システムでは，挿絵を生成する箇所は各パラグラフに 1 枚を目安とする．そのため，要約文章もパラグラフごとに生成する．次のステップで，生成した要約文章を基に挿絵を生成する．

3.3. 挿絵画像生成

挿絵画像生成では，実際に本の挿絵を生成することが目的である．生成する挿絵には，3.2 によって得た各パラグラフの重要な内容を示す要約文章を利用する．要約文章の内容を，生成画像に反映させることで挿絵を生成することができる．文章から画像を生成する Text-to-Image タスクにおいて，実現のために考慮すべき点が 2 つある．

一つ目は，質の高い画像の生成である．実画像とは，ピクセルで表現した高次元の空間の中でも，それらの情報が綺麗に配置されたごく一部である．つまり，一から画像を生成するためには，画像のピクセル数に伴う高次元なデータ分布の学習が必要となる．敵対的ネットワークである GAN は，このような高次元のデータの分布をモデル化することが可能となった．そのため，GAN を利用することで，実画像に近い多種多様な画像を生成することができる．

二つ目は，テキストと画像の扱いについてである．ある画像のキャプションを人が生成するときに，そのキャプションの表現は人それぞれとなる．このように，テキストと画像の関係は，人間においても常に一対一の関係ではない．そのため，これらの関係性をモデル化する場合においても，テキストと画像の特徴を上手く扱う必要がある．このようなテキストと画像の関係については，画像分類や画像キャプションといった分野での研究がある．これらは，画像からテキストの方向性で様々な出力を得ることが出来るが，テキストと画像の関係性がよく学習出来たモデルであると考えられる．以上の 2 点に考慮して，提案システムにおける挿絵画像の生成には，画像生成モデルには VQGAN を，テキストと画像の関係性を扱うモデルとして CLIP を応用し，Text-to-Image モデルとして VQGAN-CLIP を利用する．VQGAN は，多種多様な画像を生成することが出来る一般的な GAN よりも高解像度の画像を生成することに長けており，先ほど述べた質の高い画像生成という面で優れたモデルである．CLIP は，画像分類モデルであり，大規模なデータセットを学習させていることで高い分類精度を誇っている．つまり，先ほど述べたテキストと画像の関係性について，非常にうまく学習させたモデルであるといえる．

挿絵の生成を行う VQGAN-CLIP の入力には，画像を生成するための文章と，生成のためのパラメータが必要となる．入力として必要なパラメータには，生成される画像の内容を決定する seed 値，画像生成の際に最大何回まで学習を繰り返すかを決める iter 値，学習率を示す step size などがある．これらのパラメータは自分で決定する必要があり，最適なパラメータを探すことは困難である．そこで，Seed 値に関しては，次セクションで述べる方法によって，パラメータを決定するランキング方法を提案する．

3.4.ランキングによるパラメータ決定

3.3 で画像生成に必要なパラメータ述べたが，今回は生成される画像の内容を決定するためのパラメータである seed 値を探す方法として，本研究では画像のランキング方式を提案する．実際に入力される seed 値は int 型の実数となっており，その組み合わせは非常に多くなっている．そこで，今回提案するランキング方式では，あらかじめ入力パラメータの候補としてランダムに seed 値を用意し，それらの候補をランキング化することで，最適な seed 値を探す．

まず，候補となる seed 値をランダムに決定する．今回は 25 個の seed 値を候補として用意する．これらの seed 値を利用して，先ほど説明した VQGAN-CLIP を用いて画像を生成する．25 個の seed 値を用いて，それぞれ画像を生成し終えた後，それらの画像を基にランキングのスコアを計算する．各画像のランキングのスコアである $Score_{rank}$は，以下の式(1)に基づいて算出する．

$$Score_{rank} = \frac{Score_{CLIP}}{Loss_{image}} \quad (1)$$

$score_{CLIP}$は， CLIP による画像とテキストの類似度のスコアを表す．CLIP は，テキストと画像の関係性をうまく学習させたモデルであり，入力となるテキストと生成した候補の画像との類似度を計算した値になる．これにより，候補の画像の中から入力文章の内容により近い画像を見つけるための指標として扱う．図 2 の画像を用いたときに，CLIP を用いた時の画像と文章の類似度スコアを表 2 に例として示す．表 2 が示すように，4 枚の画像に対して、入力文章である「3 人の女性が海辺にいる」との類似度が計算でき，入力文章の内容を反映させている画像を探すことが可能となる．

ただし，VQGAN-CLIP で生成を繰り返した時に，学習の良し悪しで画像と入力テキストとの類似度に影響を及ぼす可能性がある．そこで，画像の生成を終了した時の学習の状態を示す loss 値を$loss_{image}$とすることで，学習の良し悪しを考慮してランキングスコアを計算する．

図 2：系統の異なる 4 枚の画像


図 3：提案システムによって生成される挿絵の例

表 2：入力文章と画像による類似度の結果

入力文と各画像の CLIP による類似度				
画像	1.jpg	2.jpg	**3.jpg**	4.jpg
類似度	15.55	12.411	**25.77**	13.95
入力文：3 人の女性が海辺にいる				

$score_{CLIP}$は，テキストと画像の類似度を示すスコアであり，高い数値であればあるほど入力テキストと画像は類似しているといえる．また，$loss_{image}$は，画像生成終了時の学習の状態を示す値であるため，低い値であればあるほど良い画像が生成できているといえる．このことから，$Score_{rank}$が高い画像は，テキストの内容をよく反映することが出来る seed 値として有効であり，この seed 値を最適なパラメータとして決定する．

3.5. 画像再生成

3.4 で決定したパラメータであるランキングが最も高い Seed 値を，VQGAN-CLIP のパラメータとし，同じテキストで画像を再度生成する．3.4 で生成した際は，あくまで seed 値を決めるための画像生成であるため，生成画像の質をあまり考慮していない．そこで，画像の再生成の目的としては，最終的に提案システムの出力となる挿絵の画像の質を向上させるために行う．VQGAN-CLIP は，入力テキストと生成画像の差をなくすように繰り返し学習を行う．今回の再生成では，3.4 で生成したときよりも，入力と出力の差である loss 値が小さくなるように再生成を行う．先ほど生成したときは，学習終了時の最終的な loss 値だけを参照していたが，再生成では学習回数 50 回ごとに画像と loss 値を記録し，iter 値の回数分の学習が終了したときに，loss 値が最も低いときの画像をシステムの出力とする．

図 3 は，提案システムによって生成される，最終的な出力となる挿絵の一部例である．元となっている文章は，「煙をあびた浦島太郎は，髪の毛がまっ白でシワだらけの，背中のまるまったおじいさんになってしまいました．」である．

4. 実験

提案システムにおける，自動生成された挿絵の妥当性を評価するために，本研究ではアンケートによって評価実験を行う．

4.1. 実験内容

本研究が提案するシステムの一部である VQGAN-CLIP は，教師なし学習であるため，モデルの評価が難しい．GAN などの教師なし学習では，教師あり学習と異なって，正解データがないためにモデルをどのように評価するかという問題がある．一般的に使われる性能評価の指標として，Inception Score や FID などの指標があるが，ある程度の実際の画像が必要である．本研究のシステムでは，入力される文章に制限を設けていないため，生成される画像も多種多様となることもあり，特定の画像を用意することが困難であると考えた．そのため，生成された画像の評価には客観的評価として，アンケート方法を用いる方法が最も適切であると考えた．

アンケートの内容は，Zhang ら[10]が行う Text-to-Image のアンケート評価を基に，以下の 3 項目について 5 段階評価を行う．値が小さければ，悪い評価であり，値が大きいほど，よい評価を得ることとなる．

① Image Clarity：生成された画像内にあるオブジェクト(人やモノ，背景など)を理解できるか．
② Image Quality：生成された画像の質はどうか．
③ Text to Relevance：生成画像に基となったテキスの内容が反映されているかどうか．

また，アンケートを行うための挿絵は，童話の「浦島太郎」の文章から生成する．今回アンケートを行う対象者は日本人となっており，日本人にとって親しみ深い童話のひとつであるため，生成画像に関する適切な評価を行うことが出来ると考えたためである．評価者は合計 12 名である．なお，本研究の提案システムは，英語で学習されているため，一度「浦島太郎」を英語に翻訳し，それらをシステムの入力としている．今回は，合計 19 段落の本文に対して，それぞれ 1 枚ずつの計 19 枚の画像が挿絵として生成される．

4.2. 実験環境

文章要約に用いた PEGASUS は，cnn_dailymail を使用する．CNN と Daily Mail によって書かれた 30 万件強のニュース記事を含む英語のデータセットである．文章要約の機械学習に使われる一般的なデータセットである．

VQGAN には，ImageNet のデータセットを学習に用いたモデルを使用する．ImageNet は画像認識分野における

標準的なデータセットであり，訓練用の画像データが 120 万枚用意されており，学習に十分なデータ量である．

CLIP は，提案元である OpenAI が公開されている事前学習モデルを使用する．ウェブ上から収集した 4 億にも及ぶ画像とテキストのペアを学習に利用している．

使用する GPU は，NVIDIA GeForce GTX 1080，8GB を用いて，画像の生成を行う．

5. 実験結果と考察

4 で行ったアンケートの結果として、各項目の結果と考察について述べる。

5.1 Image Clarity に関する結果

画像内のオブジェクトに関する項目である Image Clarity についての結果は，表 3 の通りである．

表 3：画像内オブジェクトに関する結果

Image Clarity					
評価	1	2	3	4	5
人数	0 人	1 人	4 人	**6 人**	1 人

画像内に含まれるオブジェクトを理解できるかどうかについて，4 という評価が最も多くなった．生成された画像を見たときに，画像内の人物やモノ，背景などをある程度理解することができたと考えられる．ただし，最も高い 5 段階の評価を得られなかった原因としては，一部の画像では学習がうまく行かなかったことも原因と考えられる．

5.2 Image Quality に関する結果

生成された画像の質に関する項目 Image Quality についての結果は，表 4 に示す．

表 4：画像の質に関する結果

Image Quality					
評価	1	2	3	4	5
人数	0 人	**6 人**	3 人	1 人	2 人

生成画像の質に関しては，2 が最も多く，あまりいい結果を得られなかったと言える．これはターゲットとなる文章と生成モデルが影響を及ぼしていると考えられる．本研究では，小説などの本文を要約した文章をターゲットとしている．要約した文章は，複数の状況を 1 つの文章にまとめているため，長い文章が生成されやすくなっている．一方，画像生成をするために学習されたモデルは，比較的に短い文章とその内容の画像によって学習が行われている．そのため，要約された文章が長くなる事や，複数の状況が 1 つの文章内に含まれるような時は，特に画像生成がうまく行かなかったと考えられる．

5.3 Text to Relevance に関する結果

生成画像と入力文章との関係に関する項目 Text to Relevance については，表 5 の結果となった．

表 5：生成画像とテキストに関する結果

Text to Relevance					
評価	1	2	3	4	5
人数	0 人	0 人	4 人	**7 人**	1 人

表 5 の結果は 4 が最も多くなり，比較的に評価が高いと思える．先程述べたように，要約された文章が長くなりがちであるが，文章内の重要な単語が，生成された画像に上手く反映されていると考えられる．

5.4 全体の評価結果

各項目についての平均スコアを表 6 に示す。

表 6：各評価項目の平均スコア

各項目の平均スコア		
Image Clarity	Image Quality	Text to Relevance
3.58	2.92	3.75

表 6 から読み取れるように，画像内のオブジェクトや文章の内容を画像に反映することができていると考えられ，本研究の目的である，挿絵の生成という部分で有用性を示すことが出来たと考える．しかし，生成する画像の質に関しては，課題が残る結果となった．学習させるモデルのデータセットの工夫や，要約文章に関する入力を工夫することで改善ができると考えられる．

6. 今後の課題

本研究で提案するシステムの現状として，テキストの意味を理解し，その内容に沿った画像の生成が可能であると考えられる．しかし，5 の結果からわかるように，生成画像の質の向上が課題に挙げられる．そのための解決策が 4 点考えられる．

まず 1 点目は，5 で述べたように，要約文章の影響についてである．要約文章は，段落の内容によって異なり，簡潔に要約が出来る段落もあれば，複雑な構造になってしまう段落もある．先ほどの行った評価と共に，19 段落の中から良いと思った画像と，悪いと思った画像を評価者に選択してもらった．そうしたところ，図 3 の画像が最も良い評価を受け，以下の図 4 が最も悪い評価であった．図 3 を生成する要約文章は，「煙をあびた浦島太郎は，髪の毛がまっ白でシワだらけの，背中のまるまったおじいさんになってしまいました．」である．図 4 の作成に利用した要約文章は，「浦島は，その箱を大切に持って，亀の背中に乗って帰りました．」である．

図 3 は，文章全体を通して，煙を浴びた後の浦島太郎がどのような状態に変化しているかを示している文章である．一方で，図 4 の文章に関しては，浦島太郎が箱を持つ行動と，亀の背中に乗って帰る行動の 2 つの情報が含まれている．このような例から，画像生成に用いる文章としては，複数の行動などが含まれる場合に，画像の質が下がってしまう可能性があることが分かる．そのため，要約文章を生成した際に，複数の状況を示す文章になったときは，異なる処理が必要であると考えられる．

図 4：最も悪い評価を受けた生成画像

続いて解決策の 2 点目は，オブジェクトの配置を正しく行うことである．5 の結果から，人物やモノなどの各オブジェクトはうまく画像化することが出来ていると考えられるが，それらを正しく配置できていないと考えらえる．例えば図 4 では，箱の中にカメがいるようにも見て取れる．オブジェクトだけ生成し，その配置を人が変更することなどで，画像の質を向上することが可能であると考えられる．

画像の質向上の解決策の 3 点目として，挿絵としての一貫性が挙げられる．本研究のシステムでは，絵本や小説など各段落に対して，文章を要約し，画像を生成している．このため，生成される画像に統一感がない．これは，挿絵として画像を生成するには良いとは言えない．基本的に，小説や絵本などの本は，1 つのストーリーとして，時系列的な流れに基づいて話が展開されていく．しかし，提案システムには，この時系列的な流れを無視して，段落ごとに画像の生成を行うため，挿絵生成に一貫性を持たせることは，画像の質向上につながると考えられる．具体的な解決策としては，登場人物など重要なオブジェクトをあらかじめ用意しておき，画像生成する際にこの情報を与えることによって，段落間での挿絵の揺らぎが少なくなると考えられる．

最後に 4 点目は，日本語への対応である．元々，文章要約や画像生成に使用するモデルは，すべて英語のデータセットを用いている．そのため，今回実験で使用した「浦島太郎」は，すべて一度英訳し，その翻訳後の文章をシステムの入力とした．日本語独特な文章構造や表現などの部分に関して，学習済みのモデルに対して，影響を与えた部分もあると考えられる．今回使用したデータセットや学習済みのモデルは，英語をベースとした大規模なデータセットであるため，それと同等の日本語データセットを用意することで解決できると考えられる．

7. むすび

今回，読者のイメージ形成のサポートとして絵本の文章から画像を生成するシステムを提案した．提案システムでは，生成に必要である入力となる文章を決めるために，文章要約の方法を用いて，事前学習モデルでのある PEGASUS を利用して，要約文章の生成を行った．その後，要約文章を入力テキストとして Text-to-Image のモデルである VQGAN-CLIP から，テキストの内容にあった画像を生成することが可能となった．また，画像生成の際に必要となるパラメータの決定を，複数枚の画像をランク付けし，最も適したパラメータを決定するランキング方式を提案した．その後，再度最適なパラメータを用いて，画像を生成し，各段落に対して挿絵を作成する．

提案システムの有用性を確かめるために，評価実験を行った．VQGAN などの教師なし学習では，定量的な評価が難しいことから，定性的な評価として，アンケートを実施した．結果として，人物やモノ，背景などのオブジェクトや，入力となるテキストの内容を画像にうまく反映することが出来たといえる．一方で，画像の質の部分では，高い評価が得られず，課題が残る結果となった．解決策として，要約文章の生成や挿絵としての一貫性に工夫を凝らす必要があると考えられる．

謝　辞

本研究を進めるにあたり，ご指導を頂いた卒業論文指導教員の黄潤和教授，助言して頂いた副指導教員の細部博史教授, 若原徹教授に感謝致します．また，日常の議論を通じて多くの知識や示唆を頂いた黄研究室の皆様に感謝致します．

文　献

[1] 田中麻巳. "小説の読みに与える挿絵の影響." *読書科学* 58.2 (2016): 74-86

[2] Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR, 2020.

[3] Esser, Patrick, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[4] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020 (2021).

[5] Ozsoy, Makbule Gulcin, Ferda Nur Alpaslan, and Ilyas Cicekli, "Text summarization using latent semantic analysis.", Journal of Information Science 37.4 (2011): 405-417.

[6] Devlin, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[7] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." arXiv preprint arXiv:2102.12092 (2021).

[8] Brown, Tom B, et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[9] 下窪聖人, "BERT で得られる分散表現に対してコサイン類似度を用いた小説の挿絵推薦", 法政大学院情報科学研究科修士論文, 2020.

[10] Zhang, Han, et al. "ERNIE-ViLG: Unified Generative Pre-training for Bidirectional Vision-Language Generation." arXiv preprint arXiv:2112.15283 (2021).