視覚文字表現と深層学習による文書分類法

島田, 大輔 / Shimada, Daisuke

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
58
(開始ページ / Start Page)
1
(終了ページ / End Page)
8
(発行年 / Year)
2017-03-31
(URL)
https://doi.org/10.15002/00014224

視覚文字表現と深層学習による文書分類法

DEEP LEARNING METHOD FOR DOCUMENT CLASSIFICATION
THROUGH IMAGE-BASED CHARACTER EMBEDDING

島田大樹
Daiki SHIMADA
指導教員　彌冨仁

法政大学大学院理工学研究科応用情報工学専攻修士課程

Languages such as Chinese and Japanese have a significantly large number of characters as compared to other languages, and each of their sentences consists of several concatenated words with wide varieties of inflected forms; thus appropriate word segmentation is quite difficult. In this study, we propose a new and efficient document classification technique for such languages. The proposed method is characterized into a new “image-based character embedding” method and character-level convolutional neural networks method with “wildcard training.” The first method encodes each character based on its visual structures and preserves them. Further, the second method treats some of the input characters as wildcards to prevent over-fitting of the classifier. We confirmed that our method showed superior performance to conventional ones for Japanese document classification tasks without data pre-processing.

***Key Words** : document classification, deep learning, convolutional neural network, Japanese character*

１．背景

テキストや音声の意味解析，知識獲得を主な目的とする自然言語処理 (natural language processing; NLP)は，web の普及と SNS (social network)などのアプリケーションの発達に伴って情報が急速に増大する昨今において，より重要性を持つ技術である．そのような背景から，文章からその意味情報を抽出する試みが数多く行われている．特に，近年の計算機性能の向上や big data と呼ばれる大規模データの整備から，統計や機械学習を応用した研究が目覚ましい成果を上げており，文書分類も NLP の主要な応用の一つである．

文書分類では，一般に，特徴抽出と分類という二段階を経て処理される．特徴抽出では，文中で意味を成す最小の単位（多くの場合で単語もしくは形態素）ごと,ないしはそれらの組み合わせを用いて文書を特徴づける数値表現を取得する．文書の特徴抽出手法としては，tf-idf (term frequency-inverse document frequency)[1] や，word2vec[2] ならびに GloVe (global vectors for word representation)[3]に代表される単語の分散表現，LSI (latent semantic indexing)[4]や LDA (latent Dirichlet allocation)[5]のようなトピックモデルが用いられ，良好な性能が報告されている．分類段階においては，ロジスティック回帰やニューラルネットワーク，サポートベクターマシン (support vector machines; SVM)など様々な分類器を適用することができ，これらの手法が文書分類に対して有効であることが示されている[6]．また，近年では，深層学習(deep learning)と呼ばれる超多層ニューラルネットワークを用いて特徴抽出と分類を同時に学習する方法が提案されている[7]．NLP における深層学習の枠組みでは，再帰構造を持つニューラルネットワーク (recurrent neural networks; RNN)が英文書分類問題で高い性能を示している[8]が，長い系列長の学習に問題が存在することや，計算の並列化が難しいといった欠点から必ずしも文書分類に適切な手法とは言えない．そこで，画像認識の文脈で提案された畳み込みニューラルネットワーク (convolutional neural networks; CNN)[9]を文書分類に適用させる試みがなされている[10]．CNN は，RNN よりも時間的に離れた系列間の関係性が学習可能であり，計算の並列化が比較的容易であるため，文書分類以外の言語処理タスクにも広く適用されている[11].

先に述べた文書分類の方法論では，日本語や中国語のような単語境界を明確に示さない記法をとる言語において，単語分割処理を行う必要が生じる．しかし，現在提案されている単語分割手法では辞書データに依る部分が大きく，新語や造語，文法的に正しくない文章が頻出する web 上のテキストや，文章の記法や習慣が異なる現代以前の文を扱う場合には期待する単語分割結果が得られない可能性が高い[12]. そうした単語単位の言語処理上の

問題を回避するべく，文字単位の文書特徴を抽出し，分類を行う研究がなされている[13,14]．文字単位の手法は，接頭語や接尾語といった単語より細かい粒度の情報を表現できるだけでなく，新語や造語に対しても頑健に動作するという利点を持つ．

一方で，文字単位での自然言語処理には，「類似した意味を持つ文字（意味的置換可能文字）」や「全く異なる意味を持つが形状的に類似する文字（視覚的置換可能文字）」について適切に扱うための処理が必要である．意味的置換可能文字とは，例えば括弧やカナならびに記号の全角/半角文字のように文字が意味するものは同一でありながら異なる文字コードが当てられている文字のことである．このような場合は，ある一種類の括弧に集約するなどの工夫が必要であるが，顔文字などへの影響を考慮するべきである．視覚的置換可能文字は，ハイフン（‐）やダッシュ（―），長音（ー）といったそれぞれが文中で果たす役割が異なるものの，文脈によって人間が柔軟に意味を解釈できる文字を指す．この場合は文脈から，どの文字を意図したものかを判別して訂正する必要がある．これらの文字に対しては辞書的に列挙した上で処理されることが多いが，「ギャル文字」や「ネットスラング」と呼ばれる新しい言語文化が次々と生まれておりそのような変化に追従することは困難である．

本論文ではまず，こうした自然言語処理上で注意するべき現象が文字の視覚的性質に起因することに着目し，視覚的文字表現を文書特徴に導入する image-based character embedding (ICE)を提案する． ICE は先に述べた問題を解決しうるだけでなく，日本語や中国語のように文字に構成性を有する場合に，従来の離散的な文字表現[15]よりも効率的に文字を表現できる可能性がある．

続いて，ICE を用いた汎用的な文書分類手法を提案する．提案する分類手法は，文字単位かつ文字の視覚的性質を抽出して文書分類を行うために，言語に関する事前知識が不要であり，記号や未知の文字，新語や造語などに頑健な手法である．本稿では，複数の日本語文書分類問題を設定し，従来的な文書分類手法と比較して提案手法が文書分類手法として様々な面で優れていることを示す．

２． 方法

（１）提案手法の全体

提案手法は，テキスト中の各文字を画像表現に変換してその視覚的特徴を抽出する image-based character embedding (ICE) と，文字単位で言語的特徴を抽出しながら文書分類を行う character-level document classification (CDC) の二つの処理から構成される．提案手法の概要を Fig.1 に示す．本研究では，両処理ともに表現能力が高く構造に柔軟性を持つ深層学習をベースにした手法を適用することを想定するが，ICE なら文字の視覚的性質を保存する手法，CDC なら文字単位に文書分類を行うことができる手法であれば良い．次節以降で，各処理の詳細について説明する．

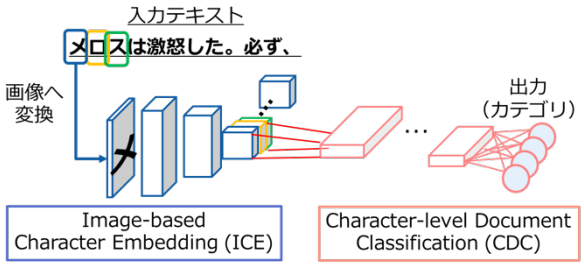

Fig.1 提案手法の全体像

（２）Image-based Character Embedding (ICE)

ICE は，入力されるテキストを画像化し，それぞれの文字から視覚的特徴を抽出する．画像認識分野では様々な特徴抽出手法が提案されているが，ICE に用いる特徴抽出手法としては，非常に視覚的性質が類似している文字については同じ特徴を抽出し，漢字のような文字についてはその構成性を表現しうる手法が望ましい．そこで，本研究では，画像に対して高い情報表現能力が報告されている stacked convolutional autoencoder (CAE)[16]を ICE の特徴抽出器として用いる．

a）Stacked Convolutional Autoencoder (CAE)の構成

Fig.2 に本研究で用いる CAE の構成を示す．CAE は，画像から特徴抽出を行う encoder $f(\cdot,\theta_f)$と，encoder が出力する特徴から元画像を復元する decoder $g(\cdot,\theta_g)$から構成される．それぞれ，学習可能なパラメータを持つ層を階層的に重ねたニューラルネットワークであり，各層では畳み込み (convolution) と最大値プーリング (max pooling)，もしくは逆プーリング (unpooling)を行い，空間的関係性を保持した二次元の特徴マップを得る．

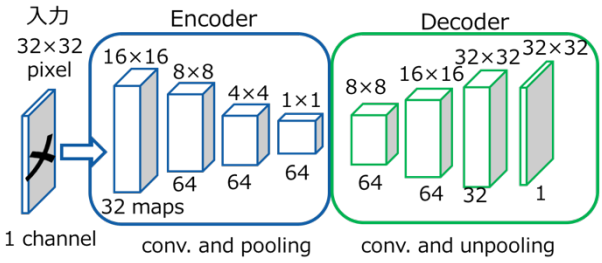

Fig.2 CAE の構成図

ある層における畳み込みは，前層のK個の入力特徴マップを$\boldsymbol{I}^1,\boldsymbol{I}^2,\ldots,\boldsymbol{I}^k,\ldots\boldsymbol{I}^K$ $(\boldsymbol{I}^k\in\mathbb{R}^{(n_y\times n_x)})$とし，現層の$L$個の畳み込みカーネルを$\boldsymbol{w}^1,\boldsymbol{w}^2,\ldots,\boldsymbol{w}^l,\ldots\boldsymbol{w}^L$ $(\boldsymbol{w}^l\in\mathbb{R}^{(n_w\times n_w\times K)})$，ならびに各畳み込みカーネルに対応するバイアス項を$\boldsymbol{b}^l$と定義すれば， Eq.1 に示すような新しい特徴マップ$\boldsymbol{h}^1,\boldsymbol{h}^2,\ldots\boldsymbol{h}^l,\ldots\boldsymbol{h}^L$を得る処理である．

$$h^l_{x,y} = relu(\sum_{k=1}^{K}\sum_{i,j\in\phi} w^l_{i,j,k}\, I^k_{x+i,y+j} + b^l) \quad (1)$$

ここでϕは畳み込みの範囲を表し，畳み込みカーネルの中心座標を(0,0)としたとき，n_wの大きさに依存したインデクス集合である．例えば，$n_w = 3$としたとき，$\phi = \{(-1,-1),(0,-1),\dots,(1,1)\}$となる．活性化関数relu(·)は rectified linear unit (ReLU)を意味し，Eq.2 に示す関数である．

$$relu(x) = \max(0, x) \quad (2)$$

最大値プーリングは，特徴マップをs_pピクセルごとに大きさ$n_p \times n_p$のプーリング領域に分割し，各領域での最大値を出力する処理である．対して逆プーリングは，各ピクセル値を大きさ$n_u \times n_u$だけ空間的に拡大する処理である．

b）Stacked Convolutional Autoencoder (CAE)**の特徴抽出**

CAE で特徴抽出を行う際には，encoder の出力のみを用いる．また，活性化関数 ReLU では出力値に上限が存在しないため，後段の CLCNN の入力として利用するには不適切である．CAE が出力する特徴ベクトルの値域を制限するために，encoder の最終層の活性化関数は tanh を用いる．

c）Stacked Convolutional Autoencoder (CAE)**の学習**

CAE は，学習時には encoder と decoder の両方を用いる．学習は Eq.3 に示す誤差関数E_{CAE}を最小化するように，CAE のパラメータθ_f, θ_g（畳み込みカーネルおよびバイアスの値）を確率的勾配降下法によって最適化する．

$$E_{CAE} = \left\| \boldsymbol{x} - g(f(\boldsymbol{x}, \theta_f), \theta_g) \right\|_2^2 \quad (3)$$

CAE の各パラメータを最適化する確率的勾配降下法には，正規化された勾配の期待値を推定することで適切なパラメータ更新量を求める Adam 法[17]を用いる．時刻tにおいて，勾配の平均の推定量を$\hat{m}^{(t)}$，勾配の分散の推定量を$\hat{v}^{(t)}$，更新するパラメータを$\theta^{(t)}$としたとき，Adam 法でのパラメータ更新式は Eq.4 のように書ける．

$$\theta^{(t)} = \theta^{(t-1)} - \alpha \frac{\hat{m}^{(t)}}{\sqrt{\hat{v}^{(t)}} + \epsilon} \quad (4)$$

Eq.4 におけるαおよびϵはハイパーパラメータである．Adam 法は，従来の確率的勾配降下法に勾配の一次と二次のモーメントを用いた適応的な学習率決定法を導入した手法であると解釈される．

（３）Character-level Document Classification (CDC)

CDC では，前段の ICE で抽出された文字単位の視覚的特徴の列を入力にとり，その文書のカテゴリを出力する．本研究では，文字単位に言語的特徴をボトムアップに構成しながら分類を行う手法として，character-level convolutional neural networks (CLCNN)を用いる．

a）Character-level Convolutional Neural Networks (CLCNN) **の構成**

Fig.3 に本研究で用いる CLCNN の構成を示す．CLCNN は，CAE と同様に畳み込みと最大値プーリングを階層的に行い，最終的に入力テキストのカテゴリを推定する．

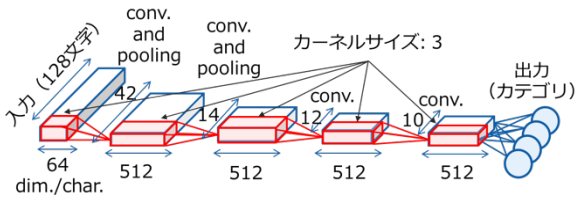

Fig.3 CLCNN の構成図

畳み込み，最大値プーリングともにテキストの時系列方向のみに処理を行う．CLCNN の各層における畳み込みは，K個の入力特徴ベクトルを$\boldsymbol{I}^1, \boldsymbol{I}^2, \dots, \boldsymbol{I}^k, \dots \boldsymbol{I}^K$ $(\boldsymbol{I}^k \in \mathbb{R}^{n_t})$，$L$ 個の畳み込みカーネルを $\boldsymbol{w}^1, \boldsymbol{w}^2, \dots, \boldsymbol{w}^l, \dots \boldsymbol{w}^L$ $(\boldsymbol{w}^l \in \mathbb{R}^{(n_t \times K)})$と定義すると，Eq.5 に示すような処理になる．

$$h^l_t = relu(\sum_{k=1}^{K}\sum_{i\in\phi} w^l_{i,k}\, I^k_{t+i} + b^l) \quad (5)$$

また，CLCNN の最終層で推定カテゴリを出力するために，softmax 関数を活性化関数として用いる．あるノードの出力u_cに対する softmax 関数の出力p_cは，総出力ノード数（カテゴリ数）をCとした時，Eq.6 のように書ける．

$$p_c = \frac{\exp(u_c)}{\sum_{j=1}^{C}\exp(u_j)} \quad (6)$$

softmax 関数の出力p_cは，入力テキストが与えられたときにカテゴリcに属する事後確率と解釈できる．

b）Character-level Convolutional Neural Networks (CLCNN) **の学習**

CLCNN の学習は，Eq.7 に示す誤差関数E_{CLCNN}を最小化するよう，CLCNN の各パラメータを確率的勾配降下法によって最適化する．Eq.7 において，理想的な出力（正解カテゴリ）はd_cとして定義される．

$$E_{CLCNN} = -\sum_{c=1}^{C} d_c \ln p_c \quad (7)$$

Eq.7 で示される誤差関数は，一般に交差エントロピー誤差と呼ばれるが，分類問題においては二乗和誤差関数よりも早く学習が進み，汎化性能が向上することが知られている[18].

本研究では，CLCNN に対して，CAE と同様に確率的勾配降下法に Adam 法を用いる．

（４）Wildcard Training (WT)

CLCNN を始めとする機械学習手法を用いた分類器の学習には，その汎化性能を向上させるために学習データの多様性が求められる．一般には，data augmentation と呼ばれる，データを加工して擬似的にデータの多様性を高める方法がとられる．自然言語処理の文脈では，シソーラスを利用して単語を類義語に置換する手法や，同じ意味をなす表現に言い換える手法が行われる．しかし，本研究のように単語分割が困難な言語では，このような手法は適合しない．そこで本研究では，単語分割不要な新しい data augmentation 手法を提案する．

提案する data augmentation 手法は，CDC への入力テキストの各文字を確率γにしたがってランダムに，ワイルドカード文字に置換するものである．本研究では，これを wildcard training (WT)と呼ぶ．日本語文での WT の適用については，予備実験の結果より，ワイルドカード文字をゼロベクトルと定義し，$\gamma = 0.2$と設定する．

３．評価実験

本研究では，提案手法の ICE および WT の有効性を確認するために，三種類のデータセットを用意して分類問題を設定した．提案手法として，CAE と CLCNN，WT をそれぞれ組み合わせた手法，そして従来の lookup-table による文字表現手法(LUT)[15]に CLCNN と WT を組み合わせた手法を試みた．比較手法には，(A) LUT に CLCNN を組み合わせた従来的な深層学習モデル，(B) 文字 3-gram 特徴から tf-idf 値を計算しロジスティック回帰によって分類するモデル(3-gram + tf-idf)，(C) 単語分割法によって分割される単語から tf-idf 値を計算しロジスティック回帰によって分類するモデル(word + tf-idf)，そして(D) トピックモデルとロジスティック回帰を組み合わせたモデルを用いた．提案手法の ICE に用いる CAE は，英数字，記号，ひらがな，カタカナ，漢字（JIS 第一水準，第二水準）を含む計 6,631 文字を学習させた．(B) 3-gram + tf-idf と(C) word + tf-idf については，最も良く出現する上位 n トークンを利用した．また，(D) トピックモデルは LSI と LDA の二種類を利用し，それぞれ単語の頻度ベクトル(bag of words; BoW)を入力にとり，一部品詞（助詞，助動詞，記号）の除去およびステミング（語形の変化の除去）を行った．

ここで，文字単位の文書分類手法は提案手法と(A)，(B)で，単語単位の文書分類手法は(C)と(D)である．

なお，CAE や LUT の特徴ベクトルの次元数や，トピックモデルのトピック数などは，複数のパラメータを試行したうち最も良かったものを採用した．

（１）小説文・論説文の著者推定

日本語小説文・論説文の著者 10 名（芥川龍之介，太宰治，梶井基次郎，菊池寛，国木田独歩，宮沢賢治，森鴎外，夏目漱石，新渡戸稲造，島崎藤村）の作品より各 10 編程度ずつ計 104 編（総文字数 2,632,225 文字）から，本文のみから著者を推定する実験を行った．全データのうち，81 編の小説文・論説文をモデルの学習に用い，残りの 23 編を性能評価に用いた．提案手法と従来手法との著者推定性能の比較を Table1 に示す．

Table1 小説文・論説文の著者推定性能の比較

methods	accuracy [%]
(proposed) CAE + CLCNN + WT	**82.61**
(proposed) CAE + CLCNN	52.17
(proposed) LUT + CLCNN + WT	78.26
(A) LUT + CLCNN	65.22
(B) 3-gram + tf-idf (n=50,000)	56.52
(C) word + tf-idf (n=50,000)	47.83
(D) LSI (# topics=60)	73.90
(D) LDA (# topics=30)	52.10

Table1 より，最もよく著者を推定できた手法は視覚的文字表現と WT を導入した提案手法であった．従来の文字表現手法を用いた場合でも，WT を導入することで，既存のトピックモデルなどの手法を大きく上回る効果が得られることが確認できる．

（２）記事の新聞社推定

朝日，毎日，産経，読売の各新聞社の政治，経済，国際カテゴリに属する記事から，200 文字以上ある記事を各社 5,610 件ずつ計 22,440 件（総文字数 69,124,142 文字）取得し，記事本文から新聞社を推定する実験を行なった．全データのうち，17,952 件をモデルの学習に用い，残りの 4,488 件を性能評価に用いた．提案手法と従来手法との新聞社推定性能の比較を Table2 に示す．

Table2 記事の新聞社推定性能の比較

methods	accuracy [%]
(proposed) CAE + CLCNN + WT	**87.81**
(proposed) CAE + CLCNN	80.95
(proposed) LUT + CLCNN + WT	76.42
(A) LUT + CLCNN	73.13
(B) 3-gram + tf-idf (n=300,000)	84.27
(C) word + tf-idf (n=49,684)	67.22
(D) LSI (# topics=2,000)	84.00
(D) LDA (# topics=70)	56.10

Table2 より，視覚的文字表現と WT を導入した提案手法が最も良い性能を示した．また，このタスクでは単語単位のアプローチよりも文字単位のアプローチが全体的に良好な性能を示していることがわかる．

（３）短文 SNS 投稿文のトピック推定

短文 SNS サイト「Twitter」に著者が指定した 14 種類のトピックいずれかに関して，2016 年 8 月 7 日から同年 10 月 6 日までの間に投稿されたものの一部（総文字数 2,167,984 文字）を抽出し，どのトピックについての投稿であるか推定する実験を行なった．データ数は各トピックにつき 5,000 件，計 70,000 件であり，投稿中にトピック名そのものが含まれる場合には削除した．全データのうち，56,000 件をモデルの学習に用い，残りの 14,000 件を性能評価に用いた．提案手法と従来手法とのトピック推定性能の比較を Table3 に示す．

Table3 SNS 投稿のトピック推定性能の比較

methods	accuracy [%]
(proposed) CAE + CLCNN + WT	**68.46**
(proposed) LUT + CLCNN + WT	59.14
(B) 3-gram + tf-idf (n=50,000)	64.46
(C) word + tf-idf (n=50,000)	49.28

Table3 より，視覚的文字表現と WT を導入した提案手法が最も良い性能を示した．このタスクでも，文字単位のアプローチがどれも良い性能を示した．特に，このタスクは各文書の文字数が最高 140 文字と制限されており，分類に利用できる情報が少ないにも関わらず, WT によるランダムな文字の置き換えを行なっても良好な結果を示した．

（４）ICE における表現次元数と性能の関係

前述した実験結果から，本研究で提案する CAE による ICE の有効性が，特に WT と組み合わせた場合において示された．しかしながら, ICE には主成分分析法 (principal component analysis; PCA)など他の視覚文字表現を適用することが考えられる．加えて，文字を効率的に低次元空間で表現するという観点では，unicode のような文字コードをそのまま利用することも検討できる．そこで本節では，CLCNN に対してどのような文字表現を用いることが最適化かどうかを明らかにする．

本実験では，記事の新聞社推定実験と同一のデータセットを利用する．実験で比較する手法は，文字表現に視覚的性質を利用するもの(ICE)として CAE と PCA，視覚的性質を利用しないものとして LUT と unicode の計四種類である．unicode による文字表現は，2 バイト表現を 16 次元のベクトルとして扱った．その他の文字表現は，8 から 256 次元までの複数の表現次元数を試行した．

Fig.4 に，文字表現とその表現次元数に対する分類性能の関係を示す．どの文字表現の性能についても，WT を導入した CLCNN で分類した結果である．

Fig.4 より，視覚的性質を用いた文字表現がどちらも高い性能を達成した．中でも CAE を文字表現に用いた場合のほうが，低い表現次元数でより高い性能を示し，CAE が文書分類に適切かつ効率的な文字表現であることが分かる．一方で，視覚的性質を利用しない文字表現であっても，従来の文書分類手法を上回る程度の性能を達成した．

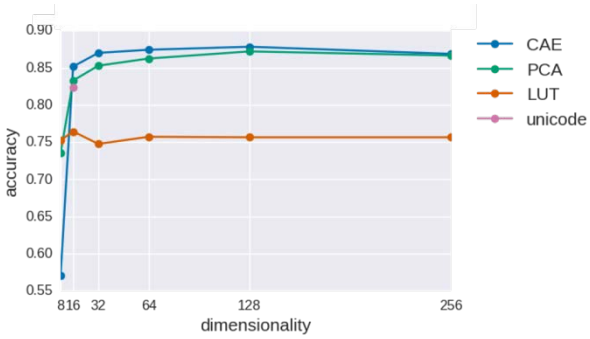

Fig.4 文字表現とその表現次元数に対する分類性能

４． 考察

本章では，前章で示した結果について，他手法による性能との差や CAE，CLCNN それぞれが学習した特徴量について解析することで考察する．

（１）各分類問題における従来手法との性能差について

前述したように，小説文・論説文の著者推定，記事の新聞社推定，SNS 投稿のトピック推定の各分類問題に対して，ICE と WT を導入した提案手法が最も良い性能を示した．

しかし，データセットによっては WT の効果に差が見られる．著者推定では CAE，LUT の両文字表現についても WT による性能向上が大きいが，新聞社推定ではどちらも性能向上が少ない．これは，データセットの量に起因すると考えられる．特に，著者推定用のデータセットは，新聞社推定用データセットと比較して規模が小さい．従って，WT はデータセットの多様性を高め，分類器の汎化性能を向上させることから，データ規模が小さいほど効果があると考えられる．このことは，WT の有無によって学習時の誤差の推移からも明らかである．WT の有無による学習誤差と評価誤差の推移を Fig.5 に示す．

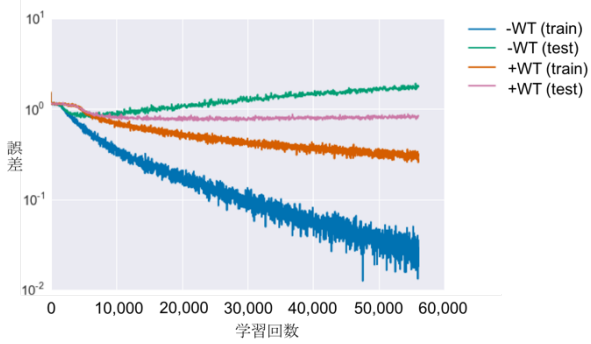

Fig.5 WT の有無による学習誤差と評価誤差の推移

Fig.5 より，WT を導入していない CLCNN (-WT)では学習誤差が減少するにつれて評価誤差が上昇しており，学

習データに過度に適合している一方で，WT を導入した CLCNN (+WT)では評価誤差の上昇が抑えられている．これらのことから, WT は機械学習における正則化に相当すると解釈できる．

（２）ICE における CAE の有効性

全ての実験タスクにおいて，ICE による有効性が示された. この結果は, 文字の視覚的性質を利用した ICE が，意味的置換可能文字や視覚的置換可能文字を吸収する効果と，文字の表意性や構成性を捉えることが出来たことによるものだと考えられる．そこで本節では，ICE に関する分析として特徴ベクトル次元を 64 次元として学習させた CAE について，文字表現空間上での学習済み文字間の類似度（マンハッタン距離）を計算した．まず，三種類のクエリ文字に対して，文字表現空間上でのマンハッタン距離が非常に近かった文字例を Table4 に示す. Table4 より，各クエリ文字についても視覚的に類似した文字が文字表現空間上においてもよく類似した表現となっている．特に，半全角が異なるだけの文字や視覚的置換可能文字が高い類似度を示している．従って，ICE はその視覚的文字表現の特性から，意味的置換可能文字や視覚的置換可能文字を吸収している．

Table4 文字表現空間上でマンハッタン距離が近い文字組の例

クエリ文字	類似度上位 3 件	マンハッタン距離
一（漢数字）	-（ハイフン）	6.07
	_（アンダーバー）	7.01
	~（チルダ）	7.06
ロ（カナ）	コ（カナ）	6.40
	口（漢字）	8.74
	ヨ（カナ）	9.94
(（半角カッコ）	（（全角カッコ）	0.00
	l（英字）	7.98
	「（カギカッコ）	8.50

また，本研究では CAE が漢字などに見られる構成性を獲得しているか確認するため，前述した文字間の類似度から類似した文字を列挙し，各文字間で最も変化の大きかった文字表現空間上の軸を調査した．その結果，ある特定の二軸が漢字の部首に強く関係した形状特徴を表現することが示唆された. Fig.6 に，「侮」という文字を CAE の文字表現空間上で特定の二軸に対応する値のみを徐々に変化させて画像を復元した結果を示す．Fig.6 の各列，各行がそれぞれ 64 次元の文字表現空間上における 36 次元目と 61 次元目の軸に対応する．Fig.6 より，最も左下の文字が「侮」と読めるのに対して右上の文字は「海」に，左上の文字は「悔」に読める文字が現れている．このことから，CAE が獲得した一部表現に文字の表意性や構成性に相当数する表現が含まれていると言える．

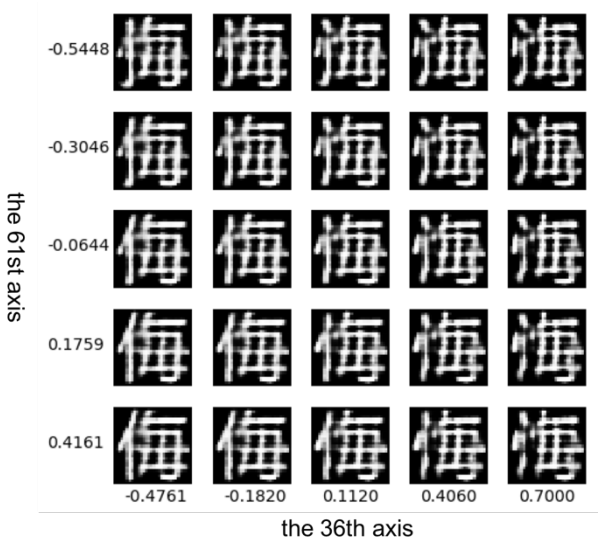

Fig.6 「侮」の特定の二軸の値を変化させた結果

CAE の高い視覚情報の表現能力は，CAE と PCA を比較した実験で低次元表現を用いる場合に CAE の方が僅かに高い性能を示したことにも関連すると考えられる．特に，CAE が文字の構成性に関連した表現を獲得していることは，効率的な文字表現を実現できていることを意味している．Fig.7 に文字表現次元を 64 次元としたときの，CAE，PCA の各手法での文字表現から復元した文字画像を示す．Fig.7 より，CAE の方がより鮮明かつ文字の特徴をよく捉えた画像が復元されていることが示されている．このことから，畳み込みによる形状特徴の位置不変性や非線形性を持つ CAE の方が，ICE の文字表現手法として PCA より適していると考えられる．

Fig.7 CAE，PCA の各表現（64 次元）からの復元画像

（３）CLCNN で構築される特徴について

前節まででは，提案手法の CAE と WT の有効性について議論したが，本節では CLCNN が学習で獲得した特徴量について考察する．本研究では，入力文字列を処理する第 1 層目の畳み込みカーネルに着目し，どのような文字列に反応する言語的特徴を獲得したか調査する．調査

のために，学習済み CAE+CLCNN+WT モデルの畳み込みカーネル内の各重みベクトルについて，最もコサイン類似度が高くなる文字表現ベクトルの文字を探索した．これによって CLCNN がよく反応する三文字の組み合わせが，ひとつのモデルあたり 512 種類（本研究の CLCNN 第 1 層目の畳み込みカーネル数）生成される．例として，新聞社推定問題を学習した CLCNN の畳み込みカーネルの一部を可視化した図を Fig.8 に示す．

Fig.8 新聞社推定モデルの第 1 層目の
畳み込みカーネル（一部）を可視化した例

しかしながら，Fig.8 の各カーネルの可視化結果の通り，最も類似度の高い文字を直接割り当てているため，それぞれが意味を持つ文字列を成していない．そこで，本研究では，より定量的な CLCNN の特徴量評価を行うため，可視化された文字列と各データセットで出現する文字種の内訳を比較する．Table5 に，各モデルの畳み込みカーネルおよびデータセットに出現する文字種（ひらがな，カタカナ，記号）の内訳を示す．

Table5 各モデルの畳み込みカーネルおよび
データセットに出現する文字種の内訳

		出現率 [%]		
		かな	カナ	記号
小説文・論説文の著者推定	データセット	60.93	0.90	8.10
	畳み込みカーネル	2.99	1.89	1.63
記事の新聞社推定	データセット	34.05	7.01	13.94
	畳み込みカーネル	6.58	3.78	2.73
短文 SNS のトピック推定	データセット	37.07	12.74	12.21
	畳み込みカーネル	14.06	9.05	1.56

Table5 より，他のデータセットよりも記号の出現率が多い新聞社推定問題したモデルでは他よりも記号に反応するカーネルが多い傾向にあり，同様にカタカナの出現率が高い短文 SNS のトピック推定問題を学習したモデルではカタカナに反応するカーネルが多い．このことから，CLCNN は学習するデータセットに応じて抽出する言語的特徴を柔軟に獲得しており，結果高い分類性能を示したと考えられる．

一方で，ひらがなの出現率が非常に高い著者推定問題のモデルでは，むしろ他のモデルよりひらがなに反応するカーネルが少ないという結果になった．これは，著者推定問題のデータセットが他のデータセットよりも漢字の多様性が高く，著者推定に対してはひらがなやカタカナで書かれる文字列よりも漢字で書かれる文字列の重要度が高いことを示唆していると言える．

５．結論

本研究では，深層学習による文字単位の文書分類手法に視覚的文字表現と単語分割不要な data augmentation 手法を導入した新しい汎用的な文書分類手法の枠組みを提案した．複数の日本語文書分類タスクにおいて，従来の単語単位および文字単位の文書分類手法を上回る性能を達成し，提案手法が有効であることを示した．

提案手法の解析では，文字をその視覚的情報を基に表現することから，従来の自然言語処理では対処の難しい形状が類似した文字の取り扱いや，括弧や句読点とピリオド，カンマのような置換されても意味が変動しない文字の取り扱いが適切になされていることが示唆された．加えて，CAE による視覚的文字表現の学習の結果，漢字の表意性や構成性を一部考慮した表現が得られることが確認された．これにより，従来的な文字表現手法と比較して，提案手法が効率的かつより文字の意味を保存した文字表現であることが明らかになった．

CLCNN が学習した特徴量を解析した試みでは，CLCNN がデータセットに含まれる文字種の出現率の異なりや，分類に寄与する文字を学習によって獲得することが明らかになった．このような日本語の自然言語処理における CLCNN の特徴量解析を行なった研究は少なく，CLCNN の理解という側面からも本研究は一定の成果を示した．

謝辞：本研究にあたり，全般にわたるご指導をしてくださった彌冨仁准教授，および彌冨研究室の皆様に深く御礼申し上げます．

参考文献

1) K. S. Jones and P.W. Daly, “A Statistical Interpretation of Term Specificity and its Retrieval,” The Journal of Documentation, Vol. 28, pp. 11–21, 1972.

2) T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Procs. of Neural Information Processing Systems (NIPS), pp. 3111–3119, 2013.

3) J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," Procs. of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.

4) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,”

The Journal of The American Society for Information Science, Vol. 41, No. 6, pp. 391–401, 1990.

5) M. D. Hoffman, D. M. Blei, and F. Bach, “Online Learning for Latent Dirichlet Allocation,” Procs. of Neural Information Processing Systems (NIPS), pp. 856–864, 2010.

6) T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” Procs. of European Conference on Machine Learning (ECML), pp. 137–142, 1998.

7) Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, Vol. 521, No. 75532, pp. 436-444, 2015.

8) D. Tang, B. Qin, and T. Liu “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification,” Procs. of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1422–1432, 2015.

9) Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, 1998.

10) X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” Procs. of Neural Information Processing Systems, pp. 649–657, 2015.

11) J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-Recurrent Neural Networks," The Computing Research Repository (CoRR), arXiv: 1611.01576, 2016.

12) 山田勉, Twitter 分析のための形態素解析の最適化, 言語処理学会年次大会発表論文集, pp. 578-581, 2014.

13) 松浦司, 金田康正, n-gram 分布を用いた近代日本語小説文の著者推定, 情報処理学会研究報告自然言語処理（NL）, Vol. 1999, No. 95, pp. 31-38, 1999.

14) 佐藤挙斗, 折原良平, 清雄一, 田原康之, 大須賀昭彦, 文字レベル深層学習によるテキスト分類と転移学習, 人工知能基本問題研究会, Vol. 102, pp. 13-19, 2016.

15) R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," The Journal of Machine Learning Research, Vol. 12, pp. 2493-2537, 2011.

16) J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” Lectures Notes in Computer Science, Vol. 6791, pp. 52–59, 2011.

17) D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” In International Conference on Learning Representations (ICLR), 2015.

18) P.Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” Procs. of International Conference on Document Analysis and Recognition, pp. 958-963, 2003.