文字レベル深層学習によるテキスト分類と転移学習

Text Classification and Transfer Learning based on Character-Level Deep Convolutional Neural Networks

佐藤 拳斗[1*]　折原 良平[1]　清 雄一[1]　田原 康之[1]　大須賀 昭彦[1]
Minato Sato[1]　Ryohei Orihara[1]　Yuichi Sei[1]　Yasuyuki Tahara[1]　Akihiko Ohsuga[1]

[1] 電気通信大学大学院情報システム学研究科
[1] Graduate School of Information Systems, University of Electro-Communications

Abstract: Temporal Convolutional Neural Network (Temporal ConvNet) is an emergent technology for text understanding. An acceptable input for the ConvNets is a sequence of words or a sequence of characters. The latter require none of the knowledge of words which depends on language-specific processing such as morphological analysis. Past studies showed that the character-level ConvNets worked well for text classification tasks in English and romanized Chinese corpus. In this article we apply the character-level ConvNets with an embedding layer to unromanized Japanese text understanding. We also attempt to reuse meaningful representations that are learned in the ConvNets with the embedding layer in the form of transfer learning. As for the application to two types of task in Japanese corpus, the ConvNets outperformed N-gram-based classifiers. In addition, our transfer learning frameworks worked well for our case studies.

1　はじめに

近年，メディア情報処理の分野において，深層学習 (Deep Learning) の研究が盛んになってきている．中でも，画像認識の分野においては，畳み込みニューラルネットワーク (Convolutional Neural Networks; CNNs, ConvNets) を深層化することにより飛躍的に精度が向上した．また，関連したドメインの知識やデータを転移して，目標ドメインの問題をより高精度で解くことを狙いとする転移学習 (Transfer Learning) に関する研究も盛んである．特にここでは，元ドメイン，目標ドメインの両者について教師ラベルが付いている帰納転移 (Inductive Transfer) のことを指す．深層学習の画像認識分野における転移学習の例として，ImageNet と呼ばれる大規模教師付き画像データセットで事前学習した (Pre-trained) ConvNet を用いて，学習対象となるデータセットについて，特徴を抽出し様々な機械学習手法と組み合わせたり，ネットワーク全体の重みの再学習 (Fine-tuning) を行うといったものがある．自然言語処理の分野における成功例としては，`word2vec`[8] と呼ばれる単語の概念ベクトルを分散表現 (Distributed Representations) として獲得する手法が有名である．自然言語処理の中でも，深層学習をテキスト分類や感情分析に用いる研究が盛んに行われている．

中でも，1 次元畳み込みニューラルネットワークをテキスト分類に適用する事例が報告されている [6, 9, 3, 10]．1 次元 ConvNet を用いた手法の中には，単語レベルで ConvNet を適用する手法と，文字レベル，特に表音文字レベルで ConvNet を適用する手法，両者を組み合わせた手法とがあり，文字レベルで ConvNet を適用する手法は形態素解析等の言語に依存したシステムが不要となる．先行研究において，文字レベル ConvNet は英語のみならず中国語のデータセットにおいても有効である事が確認されたが，その他の言語においても有効であるか不明である．また，画像認識の分野においては ConvNet がどういった特徴を抽出しているかや抽出した特徴の活用について議論されているが，文字レベル ConvNet は先行研究においてそういった議論がなされていない．

そこで本研究では，未だ検証されていない日本語のデータセットにおける文字レベル ConvNet の有効性を検証した．また，文字レベル ConvNet が抽出した特徴を活用すべく，どういった特徴を抽出しているかや転移学習に関する分析を行った．

本研究の主な貢献は以下の 3 つである．1 つ目は，日

*連絡先： 電気通信大学大学院情報システム学研究科
〒 182-0022 東京都調布市調布ケ丘 1-5-1
E-mail: sato.minato@ohsuga.is.uec.ac.jp

本語のデータセットに表音文字レベル ConvNet を適用することにより，訓練データのサンプル数が大きい場合において精度を向上させたことである．2 つ目は，文字レベル ConvNet が中間層において抽出した特徴量について N-gram 特徴量と比較，分析することにより，どういった特徴量が抽出されているか明らかにしたことである．深層学習の研究において，入力が画像以外の場合では抽出されている特徴を理解し難いといった背景があった．今回の試みにより，1 枚の畳み込みフィルタで複数の N-gram を表現できており，効果的な特徴を抽出できていることがわかった．3 つ目は，大規模データセットで学習した文字レベル ConvNet を用いることにより，小規模データセットでの精度を向上させたことである．テキスト分類向け ConvNet における転移学習に関する研究は未だなされていなかったが，今回の試みにより，画像認識分野における転移学習と同様に過学習を防ぎ，汎化性能を高める効果を確認した．

以降，本論文は 2 章では関連研究を，3 章では本研究で用いる文字レベル ConvNet のモデルと転移学習モデルを，4 章では実験とその結果を説明し，5 章ではディスカッションを行い，最後に 6 章ではまとめと今後の課題について述べる．

2 関連研究

2.1 分散表現学習

分散表現とは，単語やフレーズを固定長のベクトルで表現する情報表現の 1 つである．一例として，単語の意味的な情報をベクトルで表現する単語埋め込み (Word Embedding) 等がある．これらのベクトルは，ある単語の系列が与えられたとき次に出現する単語の条件付き確率をニューラルネットワークによってモデル化したニューラル言語モデル等によって学習することができる．単語埋め込みを獲得する高精度なモデルとして，Skip-gram モデルや CBoW モデル等 [8] がある．

2.2 テキスト向け ConvNet に関する研究

2.2.1 文字レベル ConvNet に関する研究

表音文字レベルの ConvNet をテキスト分類，感情分析に応用した研究として，Santos ら [3] の研究や，Zhang ら [10] の研究がある．Santos らの研究では，入力をランダムに初期化された文字の分散表現の系列として扱い，1 次元方向に畳み込み，プーリングを行い，そこで得られた特徴量と単語レベルの畳み込みを行った時に得られる特徴量を組み合わせる事によって精度向上を図った．Zhang らの研究では，文字を離散表現 (Discrete Representation) の 1 つである one-hot 表現と呼ばれる，対応する特定の文字に対応する次元だけ 1，他の要素が全て 0 というベクトルで表した．テキストをそれらの one-hot 表現の系列として扱い，1 次元畳み込み，1 次元最大プーリングの処理を行うのだが，その際ネットワークを深層化することにより精度向上を図った．その結果，英語のデータセットだけでなく，中国語のデータセットにおいても有効性を示した．その際，中国語は文字の種類が多く，one-hot 表現の次元が莫大になってしまうので，ローマ字化を行うことで one-hot 表現の次元を英語と同等にした．

2.3 画像認識分野における転移学習

深層学習の画像認識分野において，ImageNet と呼ばれる大規模教師付き画像データセットで事前学習した (Pre-trained) ConvNet を用いて，対象となるデータセットの特徴を抽出し，SVM で分類を学習するといった研究や，同データセットで事前学習した ConvNet に対して出力層だけを対象のタスクのものに付け替え，学習を進める Fine-tuning アプローチによる研究 [4, 1] がある．一般に，ConvNet の学習はネットワークの重みの初期値依存性が強いとされており，先の事前学習により得られたネットワークの重みを初期値として用いることで，フルスクラッチから学習するよりも良い結果が得られやすい．特に，訓練データのサンプル数が少ない場合，上記の様な初期値を得ることが過学習を防ぎ，汎化性能を高める鍵となる．

3 文字レベル ConvNet

3.1 概要

3.1.1 入力表現

本研究では，2 種類の入力表現を用いる．1 つは，先行研究で用いられていた簡便な one-hot 表現である．もう 1 つは，分散表現 (*i.e.*, 文字埋め込み (Character Embedding)) であり，これを起用することにより，日本語や中国語のデータセットで必要であった煩雑で不確実なローマ字化処理を省くことができる．

N 文字からなる文 ($\{c_1, c_2, \cdots, c_N\}$) が与えられた場合を考える．それぞれの文字 c_n は文字種類数 d 次元の one-hot 表現 $r_n \in \mathbb{R}^d$ に変換され，N 文字からなる文は $\{r_1, r_2, \cdots, r_N\} \in \mathbb{R}^{d \times N}$ となる．one-hot 表現とは，ある文字 c_n に対応するインデックスの値のみ 1，それ以外はすべて 0 のベクトルのことをいう．

文字埋め込みの場合，更に変換を必要とする．文字埋め込みは下記の様に，埋め込み行列 $W^e \in \mathbb{R}^{d^e \times d}$ と

one-hot 表現 r_n の積により生成される．

$$r_n^e = W^e r_n \tag{1}$$

ここで，文字埋め込みの次元 d^e はユーザによって決定されるハイパーパラメータ，埋め込み行列 W^e は学習パラメータである．したがって，文字埋め込みの場合，N 文字からなる文は，最終的に $\{r_1^e, r_2^e, \cdots, r_N^e\} \in \mathbb{R}^{d^e \times N}$ に変換される．

以下の説明では，ベクトル s_n を用いてベクトル r_n あるいはベクトル r_n^e を表す．

3.1.2 畳み込み

Window Size が k である畳み込みフィルタの適用範囲 $z_n \in \mathbb{R}^{d \times k}$ を考えると，

$$z_n = \left(s_{n-(k-1)/2}, \cdots, s_{n+(k-1)/2}\right)^T \tag{2}$$

となる．これらから，i 番目の畳み込みフィルタによる z_n に対する畳み込みの出力は，

$$\left[u_n\right]_i = \left[W z_n + b\right]_i \tag{3}$$

ここで，$W \in \mathbb{R}^{f \times d \times k}$，$b \in \mathbb{R}^f$ は，それぞれ f 枚の畳込みフィルタから構成される重み，バイアスを表しており，誤差逆伝播法によって学習するパラメータである．畳み込み層において，パディングと呼ばれる仮想的な 0 ベクトルの系列を入力の系列の前後に設けなかった場合は，畳み込み層の出力の次元は $f \times (N - (k-1))$ となる．

3.1.3 プーリング

$M = (N-(k-1))$ としたとき，畳み込み層における i 番目の畳み込みフィルタの出力は，$(v_1, v_2, \cdots, v_M)_i \in \mathbb{R}^{N-(k-1)}$ と表すことができる．最大プーリングサイズ p を用いて，プーリングが適用される範囲 $\left[y_m\right]_i$ は，

$$\left[y_m\right]_i = \left(v_{m-(p-1)/2}, \cdots, v_{m+(p-1)/2}\right)_i \tag{4}$$

となり，プーリング層の出力の次元は，$f \times (N - (k-1))/p$ となる．

3.2 モデルの設計

本研究では Deep モデルと Shallow モデルを構築した．Shallow モデルは Window Size を超える長さの系列情報を考慮することができないが，Deep モデルは複数の畳込みとプーリングにより，より長い系列情報を考慮することができる．

本研究における英語あるいはローマ字化した日本語の入力の文字は以下の 68 文字である．

```
abcdefghijklmnopqrstuvwxyz0123456789
,;.!?:'/\|_@#$%^&*~`"+-=<>(){}
```

また，入力の文の長さは 1014 とした．出力層を除くいずれの層の活性化関数にも ReLU を用いた．誤差逆伝播法により学習を行う際の最適化アルゴリズムにはモーメント付き確率的勾配降下法 (Momentum SGD)[2] を用いた．その際，ミニバッチサイズ 50，モーメント項の係数は 0.9 で学習率は初期値 0.01，3 epoch 毎に半減させ合計 30 epoch 学習させた．ネットワークの重みは，Glorot ら [5] の一様乱数に基づく初期化を行った．

3.2.1 Deep モデル

Zhang ら [10] は，6 層の畳み込み層と 3 層の全結合層のモデルを構築した．そのモデルには，畳み込みフィルタの枚数の多い Large と少ない Small の 2 つがある．畳み込み層のパラメータについては表 1 に，全結合層のパラメータについては表 2 に示した．表 1 の列はそれぞれ，層の番号，Large の畳み込みフィルタの枚数，Small の畳み込みフィルタの枚数，畳み込みフィルタの Window サイズ，最大プーリングのサイズを表している．表 2 の列はそれぞれ，層の番号，Large の各層の素子数，Small の各層の素子数を表している．第 9 層は出力層となっており，扱う問題によって素子数が変わってくる (*e.g.*, ポジティブ/ネガティブの 2 値を判定する場合は素子数は 2 となる．)．このモデルを以下 Large-C6FC3，Small-C6FC3 とする．ここで，Cn_1FCn_2 は n_1 層の畳込み層と n_2 層の全結合層を持つネットワークを表す．また，正則化手法として 3 層ある全結合層の間に Dropout を設けている．いずれも Dropout の比率は 0.5 とした．

表 1: Deep モデル (畳み込み層)

Layer	Large Frame	Small Frame	Window	Pool
1	1024	256	7	3
2	1024	256	7	3
3	1024	256	3	N/A
4	1024	256	3	N/A
5	1024	256	3	N/A
6	1024	256	3	3

3.2.2 Shallow モデル

本研究において，表 3，表 4 に示される，畳み込み層とプーリング層を各 1 層ずつ設けたシンプルなモデル (以下 Large-C1FC1，Small-C1FC1) を，Zhang らのモ

表 2: Deep モデル (全結合層)

Layer	Output Units Large	Output Units Small
7	2048	1024
8	2048	1024
9	Depends on the Problem	

デルとの比較・分析用に用意した．本研究で使用する日本語データセットの中に比較的規模が小さいものがあり，それに対して Zhang らのモデルの表現力が強すぎるため，過学習を起こすことが考えられる．それにより，Zhang らのモデルでは日本語データセットに対する ConvNet の有効性を検証することが難しいと考え，表現力の弱いモデルを分析のために用意した．

表 3: Shallow モデル (畳み込み層)

Layer	Large Frame	Small Frame	Kernel	Pool
1	1024	256	7	1008

表 4: Shallow モデル (全結合層)

Layer	Output Units Large	Output Units Small
2	Depends on the Problem	

3.3 文字レベル ConvNet に基づく転移学習

2.3 節において，深層学習の画像認識分野における転移学習について紹介したが，本研究ではそれらのアプローチに習い，Small-C6FC3 に基づく以下のモデルで実験を行い，評価する．

Scratch 大規模データセットで事前学習を行わない通常のモデル．

Pre-trained feature 大規模データセットで事前学習を行い，そのモデルの畳み込み層 6 層を特徴抽出器として用い，出力層にかけての全結合層の重みを乱数に基づいて初期化を行うモデル．誤差逆伝播法で学習を行うのは全結合層の重みのみとなる．

Fine-tuning 大規模データセットで事前学習したモデルの重みを初期値として誤差逆伝播法によりネットワーク全体の学習を進めるモデル．

4 実験

4.1 ベースライン手法

ベースライン手法として Bag-of-Words モデルと Bag-of-N-grams モデルを用意した．両モデルについて，下記の通り辞書を作成し，tf-idf 特徴量を作り，多項ロジスティック回帰を用いて学習，分類を行った．

Bag-of-Words 英語のデータセットにおいては，ストップワードを除く最頻出の 5000 語を辞書として用いた．日本語のデータセットにおいては，日本語形態素解析器である `MeCab` を用いて形態素解析を行い，最頻出の 5000 語を辞書として用いた．

Bag-of-N-grams 英語のデータセットにおいては，1-5gram のうち最頻出の 5000 語を辞書として用いた．日本語のデータセットにおいては，逆かな (ローマ字) 漢字変換ツールである `KAKASI`[1]を用いて，ローマ字に変換した後に英語のデータセットと同様の処理で辞書を作成した．

4.2 データセットと結果

本研究では，先行研究の再現を行い，且つ 2 種類のタスクで評価すべくカテゴリ分類と感情分析用の英語データセットを用意した．また，それに合わせて同 2 種類のタスクの日本語データセットを用意した．これらの概要を表 5 に示す．

AFPBB データセット (AFPBB) フランス通信社 AFP の日本語版[2]の 2006 年 5 月から 2016 年 5 月の記事を独自に収集した．“Lifestyle”，“Politics”，“Science”，“Sports” の 4 カテゴリがある．

楽天レビューデータセット (Rakuten review) 楽天データセットを国立情報学研究所情報学研究データリポジトリ[3]から入手した．楽天データセットの内，2012 年 4 月から 2012 年 12 月の楽天市場のレビューデータからレビューの評価の★の数が 1 つと 5 つのデータを用いて極性分類を行った．

AG ニュースデータセット (AG's news) AG's corpus of news articles (以下 AG ニュースデータセット) を Gulli らのウェブサイト[4]から入手した．記事を多く含む “World”，“Sports”，“Business”，“Sci/Tech” の 4 カテゴリを用いた．

[1] http://kakasi.namazu.org
[2] http://www.afpbb.com
[3] http://www.nii.ac.jp/dsc/idr/rakuten/rakuten.html
[4] https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

表 5: データセット概要

Dataset	Task	Language	#Classes	#Train	#Validation	#Test
AFPBB	News Categorization	Japanese	4	48,000	1,500	1,500
Rakuten review	Sentiment Analysis	Japanese	2	600,000	40,000	40,000
AG's news	News Categorization	English	4	360,000	20,000	20,000
Amazon review	Sentiment Analysis	English	2	600,000	80,000	80,000

表 6: 各データセットに対する結果 (分類精度)

Model	AFPBB	Rakuten review	AG's news	Amazon review
Bag-of-Words	**0.9470**	0.9548	0.8689	0.8822
Bag-of-N-grams	0.9260	0.9510	0.8767	0.8813
Small-C1FC1 one-hot	0.9385	0.9585	0.8701	0.9113
Large-C1FC1 one-hot	0.9410	0.9637	**0.8849**	0.9254
Small-C1FC1 embedding	0.9130	0.9677	N/A	N/A
Large-C1FC1 embedding	0.9215	**0.9699**	N/A	N/A
Small-C6FC3 one-hot	0.9120	0.9637	0.8710	0.9216
Large-C6FC3 one-hot	0.9295	0.9668	0.8793	**0.9319**
Small-C6FC3 embedding	0.9400	0.9669	N/A	N/A
Large-C6FC3 embedding	0.9380	0.9689	N/A	N/A

Amazon レビューデータセット (Amazon review)
Amazon レビューデータセット [7] を，Stanford Network Analysis Project (SNAP) のウェブサイト[5]から入手した．"Books" 等をはじめとする合計 8 カテゴリからレビューの評価の★の数が 5 つと 1 つのデータを用いて極性分類を行った．

結果を表 6 に示す．AFPBB データセットに対する結果は，Bag-of-Words モデルが最も良い結果となったが，これはデータセットの規模が比較的小さいことに起因すると考えられる．また，C6FC3 one-hot モデルは小規模なデータセット且つローマ字化による削減された入力の情報量に対して表現力が強すぎることから過学習を起こしている．逆に C6FC3 embedding モデルはローマ字化されていない完全な入力の情報量に対してより適切な表現力であったため精度が向上した．楽天レビューデータセットに対する結果は，Deep モデルと Shallow モデルでは同じ入力形式同士を比較してもさほど差がなかった．AG ニュースデータセットに対する結果は，Large-C1FC1 モデルが最も良かったが，どのモデルも適切に特徴語を抽出することができ，さほど差がなかった．Amazon レビューデータセットに対する結果は，Large-C6FC3 モデルが最も良かった．Bag-of-Words モデルと Bag-of-N-grams モデルで用いられる最頻出語に基づく辞書は，感情分析タスクにおける特徴語を含んでいないと考えられるが，柔軟な ConvNet モデルは特徴語の抽出に成功していると考えられ，約 3-5%分類精度が向上した．

[5]https://snap.stanford.edu/data/web-Amazon.html

4.3 転移学習の実験設定と結果

本研究では大きく 2 種類の転移学習の実験を行った．1 つは，日本語データセットにおける入力となる埋め込みベクトルの意味表現の事前学習を大規模データセットで行う実験，もう 1 つは，ネットワーク全体を別の大規模データセットで行う実験である．

4.3.1 埋め込みベクトルの事前学習

文字レベルの系列に対し Skip-gram モデル [8] を適用することで，入力となる文字埋め込みを事前学習した．Skip-gram モデルの Window Size は 5 文字分，出力層にはネガティブサンプリングを，その際のサンプリング数は 5 とした．事前学習には楽天レビューデータセットの学習用データセットを用い，転移先データセットには AFPBB データセットを用いた．

結果を表 7 に示す．いずれのモデルも事前学習を行うことによって精度が向上したが，特に Shallow モデルでの精度が大幅に向上した．小規模データにおいて Shallow モデルでは埋め込み層の学習が正しく行われない問題が転移学習によって解消された．

4.3.2 ネットワーク全体の事前学習

用いたデータセットの概要を表 8 に示す．事前学習を行うための大規模データセットとして "Books"，"Video Games" をはじめとする計 16 カテゴリを，転移学習先のデータセットとして "Movies and TV"，"Electronics"，

表 7: AFPBB データセットへの転移学習の結果．数字は分類精度を表している．char2vec は Skip-gram モデルで埋め込み層を事前学習したモデルを示す．

Model	Random	char2vec
Small-C1FC1 embedding	0.9130	0.9355
Large-C1FC1 embedding	0.9215	**0.9465**
Small-C6FC3 embedding	0.9400	0.9455
Large-C6FC3 embedding	0.9380	0.9440

“Home and Kitchen” カテゴリを用いた．また，事前学習用のデータセットに転移学習先のカテゴリは含まれない．Small-C6FC3 で事前学習用のデータセットに対して学習，評価を行った結果，分類精度は 0.9168 となった．

表 8: 転移学習用データセット概要

Dataset	#Train	#Validation	#Test
Pre-train	1200,000	120,000	120,000
Movies	150,000	30,000	30,000
Electronics	150,000	30,000	30,000
Home	60,000	9,000	9,000

実験結果を表 9 に示す．いずれも Fine-tuning が最も良い結果となった．以前の実験と合わせて，データセットの規模が小さいほど，Scratch モデルは Bag-of-N-grams 等と同等あるいはそれ未満の分類精度となることがわかる．このような場合には，転移学習が非常に有効であることがわかる．

表 9: Amazon レビューデータセットへの転移学習の結果．数字は分類精度を表している．

Model	Movies	Electronics	Home
Bag-of-Words	0.85503	0.86033	0.8524
Bag-of-N-grams	0.85603	0.87550	0.8693
Scratch	0.85827	0.87854	0.8581
Pre-trained feature	0.88697	0.88183	0.8988
Fine-tuning	**0.8992**	**0.90523**	**0.9123**

5 ディスカッション

5.1 畳み込みフィルタに強く反応する N-gram について

先行研究では，畳み込みフィルタの重みの可視化を行うことによりどの文字についてよく学習されているか調査していたが，具体的にはどういった特徴を抽出しているか理解困難である．そこで，畳み込みフィルタに強く反応する (*i.e.*, 畳み込みフィルタとの畳み込み演算の出力値が大きい) N-gram 特徴量について調査した．本研究で用いた ConvNet のモデルの第 1 層目の畳み込みフィルタの Window Size は 7 であったので，評価用のデータセットから 7-gram を生成し，ある畳み込みフィルタとの畳み込み演算を行い，その出力値のランキングを作った．表 10 は AFPBB データセットに対して Large-C1FC1 one-hot で学習を行った結果である．これらの文字列にマッチングしている記事の部分文字列を調べた結果，“〜は数字 (1 桁)-数字 (1 桁) で逆転 (あるいは引き分け)” となっているものが多々見受けられ，これはスポーツの記事に多く含まれていた．実際，表 10 の N-gram の文字列を先頭から全て OR で連結した正規表現 `[,a][132][\-][213][d][e][ghbf]` に一致する文字列を含む記事数をカテゴリごとに集計した結果，“Sports” が 20 件，その他カテゴリは全て 0 件となり畳み込みフィルタの特徴抽出が働いていることがわかる．つまり，少なく見積もっても約 $2\times3\times1\times3\times1\times1\times4 = 72$ 通りの 7-gram をほぼ単一の特徴量として 1 枚の畳み込みフィルタで表現できていることになる．また，仮に “逆転” や “`gyakuten`” を特徴量とするとしても，選挙の逆転当選や支持率の逆転を報道する記事があった場合，この特徴量のみから “Sports” であると判断することは誤りあり，畳み込みフィルタで学習された特徴量の優位性が見て取れる．

表 10: ある畳み込みフィルタによる畳み込み演算の出力値ランキング (バイアス項を除く)．AFPBB データセットに対して Large-C1FC1 one-hot で学習した結果を用いて評価用データセットで計算した．

7-gram	Value	7-gram	Value
`,1-2deg`	0.50774	`a3-2def`	0.47848
`a3-2deg`	0.50613	`a1-3deh`	0.47644
`a1-1deh`	0.49199	`a3-1deh`	0.47461
`a2-2deg`	0.48423	`a2-1deg`	0.47368
`a3-2deb`	0.48332	`,3-2deh`	0.46938

5.2 転移学習の結果について

5.2.1 埋め込み層の事前学習の結果について

埋め込み層の事前学習に用いたデータセットは感情分析タスク用で，転移先のデータセットはニュースのカテゴリ分類タスク用と異なるものだったが，精度が向上した．これは，文字埋め込みの概念ベクトルはデータセットのドメインが全く異なるものでもほとんど同様なものが獲得できるからと考えられる．

5.2.2 ネットワーク全体の事前学習の結果について

本研究で用いた Zhang らのモデルは，非常に Deep であり，学習する必要のあるパラメータが膨大となっている．そのため，データセットのサンプル数が少なければ適切なパラメータを学習することは難しく，過学習を起こしてしまう．4.3.2 節の実験においては全てのデータセットにおいて訓練サンプル数が比較的少なく，Bag-of-Words や Bag-of-N-grams モデルと同等やそれ以下の精度となっていたが，転移学習を導入することにより，精度が大幅に改善した．今回の実験ケースからは，同一言語，同一タスクであれば，学習した畳み込みフィルタを転移学習で活用することができることがわかった．

6 まとめ

本研究では，日本語のデータセットにおいて表音文字レベル ConvNet を適用することにより，データのサンプル数が大きい場合に精度を向上させた．また，文字レベル ConvNet が中間層において抽出した特徴量について N-gram 特徴量と比較，分析することにより，どういった特徴量が抽出できているか明らかにした．それにより，1 枚の畳み込みフィルタで効果的な特徴を抽出できていることがわかった．加えて，転移学習により，小規模データセットにおける精度を向上させた．文字の埋め込み層の転移と畳み込み層の転移の 2 種類の実験を行った結果，前者は事前学習と転移先のデータセットのドメインが異なっていても精度向上が期待でき，後者は少なくとも同一ドメインの場合に精度向上が期待できることがわかった．テキスト分類向け ConvNet における転移学習に関する研究は未だなされていなかったが，今回の試みにより，画像認識分野における転移学習と同様に過学習を防ぎ，汎化性能を高めることができた．

今後の課題としては，中国語等の日本語以外の表意文字を含む言語における埋め込み層の利用とその事前学習の検証や，あるタスクの大規模データセットで事前学習したネットワーク全体を，全く別のタスクのデータセットへ転移学習する手法の検討などが考えられる．

謝辞

本研究は JSPS 科研費 26330081, 26870201, 16K12411 の助成を受けたものです．

本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所／東京大学 本位田 真一 教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．

本研究では，楽天株式会社が提供し国立情報学研究所が配布している「楽天公開データ」を利用した．関係者の皆様に大変感謝致します．

参考文献

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the Performance of Multilayer Neural Networks for Object Recognition. In *Proc. ECCV*, pp. 329–344, 2014.

[2] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in Optimizing Recurrent Networks. In *Proc. ICASSP*, 2013.

[3] Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. Classifying Relations by Ranking with Convolutional Neural Networks. In *Proc. ACL*, pp. 626–634, 2015.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. CVPR*, 2014.

[5] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proc. AISTATS*, 2010.

[6] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proc. EMNLP*, pp. 1746–1751, 2014.

[7] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring Networks of Substitutable and Complementary Products. In *Proc. KDD*, pp. 785–794, 2015.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. NIPS*, pp. 3111–3119, 2013.

[9] Aliaksei Severyn and Alessandro Moschitti. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *Proc. SIGIR*, pp. 959–962, 2015.

[10] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Proc. NIPS*, pp. 649–657, 2015.