2T-09

文字レベル残差畳み込みニューラルネットワークを用いた日本語文書分類に対する転移学習

守屋 俊 † 柴田 千尋 ‡

東京工科大学コンピュータサイエンス学部 †‡

1 はじめに

畳み込みニューラルネットワーク (CNNs: Convolutional Neural Networks) は主に画像認識など，画像を適用の対象として発展してきたが，近年，文書分類にも適用され，高精度に分類ができることが知られてきている．とくに，Zhang ら [1] は，単語を文字レベルに分解したのち，比較的深い層をもつ CNN により，文書分類を行い，優れた分類結果を得ている．しかし，日本語の文書に対する文字レベル CNN を用いた文書分類の研究例はまだ少ない．その原因として，文書分類に利用できる日本語データセットの少なさが挙げられる．その問題に対処する手法として，他のタスクの学習結果を用いる，転移学習と呼ばれる手法が挙げられる．本論文では，従来の文字レベル CNN に対し，残差ネットワーク (Residual Network) を適用し，より層を深くしたネットワークに対して，転移学習を行うことで，より精度を高めることを狙う．また，様々なデータセットにこの手法を適用し，その効果を確認する．

2 関連研究

Conneau ら [2] は，Residual Network を用いたより深い構造の文字レベル CNN (VDCNN: Very Deep CNN) により文書分類を行い，Zhang らの文字レベル CNN よりも高い分類精度を出している．本研究では VDCNN のネットワーク構造を参考にし，日本語の文書分類に適した形に変更したネットワークを使用する．転移学習についての先行研究として，Mou らは，ニューラルネットワークを用いた文書分類において，転移元と転移先のタスクの類似度が高い場合に転移学習が上手く機能すると報告している [3]．また，佐藤は，文字レベル CNN を使用した日本語文書分類において，ネットワーク全体の転移学習では類似タスク間の場合に精度が向上したとの結果を報告している [4]．本研究では，VDCNN に対して類似タスク間での転移学習を適用し，VDCNN においても転移学習が有効かを検証する．

3 提案および実験の方法

ニュースカテゴリの分類予測について次の方法で転移学習を行う．まず, 使用したモデルは，節 3.1 で述べるような，VDCNN と同様の構造をもつ文字レベルの CNN とする．日本語では漢字が存在するため，埋め込みベクトルのサイズを，十分な表現能力が持てるように，50 次元とする．データセットとしては, 異なるニュース記事の集合を，転移元，転移先のデータセットとし，そのジャンルをカテゴリとして分類するものとする．対象となるニュース記事の集合が異なるため、当然カテゴリーも異なるが，本論文では，節 3.3 で述べるように，比較的近いと考えられるものを選んだ．

3.1 使用したネットワーク

17 層の畳み込み層を持つ VDCNN を参考に Chainer[*1] を用いて実装した[*2]．入力となる文字列の長さは，1024 文字とする．ネットワークの畳み込み層の初期化には，He らの初期値を使用する．以下では，実装した VDCNN の層を下層から順に，embed1，conv1, res2,···,res5，fc6,···,fc8 とする．各 res 層の中には 4 個の畳み込み層が含まれている．ネットワークの学習には確率的勾配降下法を使用し，その慣性項を 0.9，学習率の初期値は 0.01 に設定した．また，3 epoch ごとに学習率を半減させ，30 epoch 学習を行った．

Transfer Learning for Text Classification using Residual Convolutional Neural Networks

†Shun Moriya Tokyo University of Technology

‡Chihiro Shibata Tokyo University of Technology

[*1] https://chainer.org/

[*2] https://github.com/9shikixp/vdcnn-transfer

3.2 転移の枠組み

転移の枠組みとしては，AFPBB ニュースデータセットで学習した文書分類ネットワークのうち，embed1 から res5 層までの重み・バイアスを livedoor ニュースコーパスの文書分類ネットワークの初期値として用いる．fc6 から fc8 の重み・バイアスについては，ランダムに初期化する．転移した重みとバイアスについては，固定する場合 (Fixed) と固定しない場合 (Init) の 2 通りで実験を行う．

3.3 対象とするデータセット

対象とするデータセットを表 1 に示す．

表 1　使用したデータセット

	データセット名	訓練データ数	テストデータ数	カテゴリ数
転移元	AFPBB	46,800	5,200	4
転移先	livedoor	2,700	300	4

AFPBB データセットは，AFPBB ニュースの Web ページ[*3]から独自に収集したものを使用する．politics，environment-science-it，lifestyle，sports の 4 カテゴリのニュース記事を取得した．livedoor データセットは，株式会社ロンウィットによって公開されているニュースコーパス[*4]の内，AFPBB のデータとカテゴリの性質が比較的近いと考えられる 4 カテゴリ (topic-news，it-life-hack，livedoor-homme，sports-watch) を使用した．

4　実験結果および考察

転移元のネットワークの学習は，AFPBB データセットで行う．転移先のデータセットは，livedoor データセットを使用する．その際，転移無しでゼロから学習を開始した場合 (Scratch) と，3.2 節で述べた 2 通りの枠組みで転移を行った場合のテストデータに対する正解率について比較を行った．加えて，転移先のデータ数が少ない場合の転移学習の効果を明らかにするため，転移先の訓練データの数を，元の 1/2, 1/4, 1/8, 1/16 に減らした場合についても比較を行った．テストデータに対する正解率の比較結果を表 2 と図 1 に示す．

表 2　テストデータに対する正解率 (最終 5epoch の平均)

model	original	1/2	1/4	1/8	1/16
Scratch	0.87	0.75	0.73	0.61	0.50
Fixed	0.80	0.76	0.74	0.69	0.65
Init	**0.94**	**0.92**	**0.89**	**0.87**	**0.80**

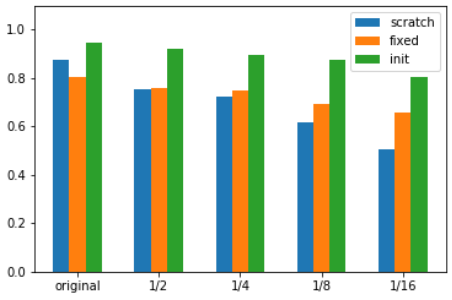

図 1　テストデータに対する正解率の比較

転移先のすべてのデータサイズに対して，ネットワークの初期値として転移元のものを使った場合 (Init) が最も精度が高いことがわかる．VDCNN においては，少なくとも今回行ったような類似データセット間の転移学習を用いることは極めて有効な手段である．一方，転移後の重みやバイアスを固定したとき (Fixed) には，訓練データの数が十分に多い (Original) の場合には，Scratch のものよりも精度が低く，逆効果となっているが，それ以外の場合では優劣は逆転しており，その差は訓練データの数を減らすにつれて大きくなっている．また，データサイズを減らさないで転移無し (Scratch) の精度と，データサイズが 1/8 でネットワークの初期値として転移元のものを使用した場合の精度がほぼ同じである．加えて，各データサイズを減らすにつれて，Scratch と Init の精度の差が広がっていく．したがって，訓練データの数が少なければ少ないほど，Init 方式での転移が有効であるということが考えられる．

参考文献

[1] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[2] A. Conneau, H. Schwenk, L. Barrault, and Y Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1107–1116, 2017.

[3] L. Mou, Z. Meng, R. Yan, G. Li, Y. Xu, L. Zhang, and Z. Jin. How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*, 2016.

[4] 佐藤拳斗．文字レベル深層学習によるテキスト分類と転移学習．2016.

[*3] http://www.afpbb.com/

[*4] https://www.rondhuit.com/download.html#ldcc