Vol.2018-NL-238 No.6
2018/12/11

# Very Deep CNNによる文書分類における トピック分布を用いた事前学習

守屋 俊[1,a)]　岡本 千尋[2,b)]

**概要**：Conneau らによる Very DeepCNN(VDCNN) [3] は，文字レベルの埋め込みを行った後 ResNets [7] に類似した構造を持つ層の深い畳み込みニューラルネット (CNN) であり，高精度な文書分類が可能な手法の一つである．我々は，外部データセットを一切用いず，同一のデータセットから教師なし学習であるトピックモデルにより推定された複数のトピック分布を，ソフトラベルとして事前学習 (Mix-Soft と呼ぶ) を行ったのち，本来の目標ラベルを用いて CNN を再学習させる手法を提案する．本提案手法は，再学習を行う前までは，教師なし学習であるので，教師ラベル付きの文書が一部しかない場合も，再学習の際に，教師なしデータを用いることで，学習することが可能である．本研究では，実際に，AGNews, DBPedia, Yelp などのデータセットに対して Latent Dirchlet Allocation (LDA) 用いて，トピック分布を複数推定し，VDCNN の事前学習を行う．その後，VDCNN を再学習することで，事前学習を行わない場合に比べ，分類精度が改善することを示す．特に，事前学習として Mix-Soft を用いたとき，最も精度が向上することを示す．また，半教師あり学習としては，大幅な精度改善が可能であり，特に，ラベル付きデータの数が極めて少数に制限されている場合でも，一定程度の精度を達成できること示す．

Shun Moriya[1,a)]　Chihiro Okamoto[2,b)]

## 1. はじめに

文書分類のタスクにおいても，畳み込みニューラルネット (CNN) やリカレントニューラルネット (RNN) を用いたモデルの学習が有効であることが近年示されている [3], [8], [9], [10], [17]．しかし，実際に，ウェブスクレイピングなどで集めた文書のセットに対して何らかの自動分類を行うようなタスクに直面したとき，それと類似したカテゴリに分かれたラベルを持つデータセットが存在しないことが多い．この問題に対応する一つの方法としては，より大規模で汎用的なデータを用意して事前学習させ，それを一部に組み込むなどして，追加的に学習させることが考えられる．本論文では，外部データを用いずに，事前にある程度の学習 (オートエンコーダなど) を行う事前学習と区別するため，外部データを用いた事前学習を行うことを総じて転移学習と呼ぶこととする．画像認識や生成では，転移学習は応用上よく用いらており [4], [5], [16]，ImageNet のデータセットで事前学習させることが多い．自然言語処理の分野では，転移学習の有効性が示されたのは比較的新しく，例えば，ULMFiT [8] と呼ばれる，LSTM 言語モデルを大規模データで学習し，分類対象となるドメインのデータで LSTM 言語モデルを再学習して文書分類を行う手法が存在する．

一方で，深層学習を用いた文書分類における転移学習は，転移元と転移先データセットの類似度が高い場合に有効であることが知られている [14]，したがって，有効に転移学習行うためには，類似した規模の大きい転移元データセットを用意する必要がある．一般に，ニュース記事のカテゴリやレビュー文の評価の分類など，分類のための大規模なデータセットは存在するが，実際の問題に適用する際に対象とするデータに対して，それらの大規模なデータセットが，必ずしも類似しているとは限らず，実際に有効な転移元データを用意できないことが多い．

転移学習に変わる方針としては，外部データ (転移元データセット) を用意しない場合でも，教師なし学習によって実際に解きたいタスクとは別のラベルを獲得し，同一のデータで異なるラベルによる事前学習が有効となる可能性

[1] 東京工科大学 コンピュータサイエンス学部
[2] 東京工科大学大学院 バイオ・情報メディア研究科
[a)] c0115334ef@edu.teu.ac.jp
[b)] shibatachh@stf.teu.ac.jp

も考えられる．本研究では，実際に，外部データを使わなくても，トピックモデルによる教師なしの事前学習により精度が向上することを示す．これは，一般的な転移学習や distant supervision [13] などとは異なり，同一のデータによる事前学習によっても，精度が向上することを意味する．また，提案手法は，再学習の際に一部の文書のみを対象にすることで，半教師あり学習とも捉えることができる．実際に，データセット中のラベルありの文書の数を制限することにより，半教師あり学習としての精度の向上の効果がどの程度あるのかを実験を通して示す．

## 2. 関連研究

Kim [10] の手法は，CNN を文書分類に適用した初期の手法であり，単語レベルの埋め込みを行い，比較的浅い層の CNN で構成されているものの，一定程度の精度を得ている．Zhang ら [17] 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，特に規模の大きいデータセットにおいて優れた分類結果を得ている．Conneau ら [3] は，ResNet に類似した構造を持つ層の深い文字レベル 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記録している．また，画像分類の転移学習については，Yosiniski ら [16] によって，CNN を用いた画像分類における転移学習の有効性が検証されている．また，文書分類の転移学習についての先行研究としては，Mou ら [14] は，ニューラルネットを用いた文書分類において，転移元と転移先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ことを示している．

## 3. 提案手法

提案する学習手法はおおよそ次のような 3 つの手順に分かれる (図 1)．

(1) まず，データセット中の文書集合 ($D$) を用いて，複数回トピックモデルの学習を行い事後確率から近似的にサンプルすることで，各文書 $d$ に対するトピック分布が複数得られる ($\theta_d^{(1)}, \cdots, \theta_d^{(N)}$)．

(2) 次に，同一のデータ $D$ を入力，得られたトピック分布を正解ラベルとして，深層ニューラルネット (DNN) の学習を行う．

(3) $D$ のうちカテゴリの教師ラベル ($t$) が付いている文書集合 $D'$ とする．(2) で得られた DNN の重み・バイアスを初期値として，$D'$ を入力，$t$ を正解ラベルとして，再度 DNN を学習させる．

提案手法は次の 2 つの学習の枠組みとして捉えることができる．まず，$D'$ と $D$ を等しく取ると，精度改善のための事前学習の手法であると言える．また，$D'$ を $D$ の真の部分集合となるようにとると，半教師あり学習の手法になるといえる．

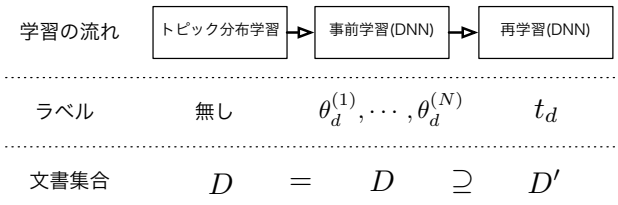


**図 1** 提案手法の概略図

### 3.1 トピックモデル

Latent Dirichlet Allocation(LDA) [1] は，データセット中の各文書に対するトピック分布を教師無しでベイズ推定するための生成モデルである．各文書は次のように生成されると仮定される．まず，各文書ごとに，トピックの確率分布が生成される．次に，文書中の単語の数だけ，次のようにして単語が生成される．文書の持つトピック分布からトピックが一つ選ばれ，そのトピックから単語が一つづつ生成される．また，トピックから単語が生成される確率は，文脈や前後の単語とは独立に定まると仮定される (bag-of-words)．以降では，文書 $d$ に対するトピックの分布を $\theta_d$, トピック $k$ に対する確率を $\theta_{dk}$ で表す．また，トピック $k$ から生成される単語の分布を $\phi_k$ で表す．また，データセット (文書集合) を $D$ であらわし，$\alpha$，$\beta$ を $\theta_d$, $\phi_k$ の事前分布 (Dirichlet 分布) を決めるパラメータとする．LDA の目的は，$\phi$， $\theta$ の事後確率:

$$P(\phi, \theta | D, \alpha, \beta) = \frac{P(\phi, \theta, D | \alpha, \beta)}{P(D | \alpha, \beta)}$$

をベイズ推定することである．しかし，上式の右辺の分母を直接計算することは困難なので，様々な近似手法が存在するが，マルコフ連鎖モンテカルロ (MCMC) 法の一種である，ギブスサンプリング を用いて近似する手法を用いると，十分に時間をかければ他の近似手法よりも精度の面で優れていることが知られている [15]．$\phi, \theta$ を MCMC により直接サンプリングする代わりに，トピック $k$ から単語 $i$ が生成された回数 $n_{ki}$ を確率変数とし，$\phi, \theta$ を周辺化し，$P(D, n | \alpha, \beta)$ に基づき $n$ を Gibbs Sampling によりサンプリングする方法は，Collapsed Gibbs Sampling (CGS) と呼ばれる [11]．本研究では，推定に CGS を用い，最終的な $\phi, \theta$ の推定値としては，最終的にサンプルされた $n$ (次の式中の $n^{\text{sampled}}$) を与えたときの，$\phi, \theta$ の期待値:

$$\phi^*, \theta^* = \mathbb{E}\left[ \phi, \theta \mid n^{\text{sampled}}, \alpha, \beta \right]$$

を推定値とする．

学習のためのデータとしては，次節で説明する事前学習で用いるものと全く同じものとする．具体的には，AGNews などに含まれる訓練用のデータを用いる．LDA は学習時にラベルを必要としないため，検証用のデータを LDA の

学習用のデータとして含めることも可能である．しかし，その場合，予測を行うたびに，新たにトピックモデルを学習し直すことを想定することとなる．したがって，本論文では検証用のデータは含めない．また，外部知識の効果を見るのが目的ではないため，トピックモデルの学習，事前学習，再学習全てにおいて，同じデータセットを用いる．

### 3.2 トピック分布を用いた事前学習

前節の学習で得られたトピック分布 $\theta^*$ を正解ラベルとして，VDCNN の事前学習を行う．その際に，トピック確率をそのまま正解ラベルとして使用する手法を**ソフトラベル** (Soft)，トピック確率が最大のインデックスのみを 1 とし，それ以外の要素を 0 とする手法を**ハードラベル** (Hard) と呼ぶこととする．また，トピックモデルによるトピック分布のラベルは，複数のトピック数でトピックモデルを作成することによって，1 つの文書に対して複数の事前学習用のラベルを作成する事ができる．こうして作成した複数のソフトラベルを用いて，マルチタスク [12] による事前学習を行う．この手法を**混合ソフトラベル** (Mix-Soft) と呼ぶこととする．VDCNN からの出力を $f(d) \in \mathbb{R}^K$ で表記する．ここで $K$ はカテゴリ数である．$f_k(d)$ は文書 $d$ に対するカテゴリ $k$ の確率を表す．事前学習の際の損失には，クロスエントロピー (以下 $H$) を使用する．

ソフトラベルの場合の目的関数 ($O_{\text{soft}}$) は次式となる：

$$O_{\text{Soft}}(D) = \mathop{\mathbb{E}}_{d\sim\mathcal{D}}[H(\theta_d^*, f(d))] = \mathop{\mathbb{E}}_{d\sim\mathcal{D}}[-\theta_d^* \cdot \log f(d)]\,.$$

ここで $\mathcal{D}$ はデータセット中の文書集合 $D$ 上の一様分布とする．

ハードラベルの場合法の目的関数は次式：

$$O_{\text{Hard}}(D) = \mathop{\mathbb{E}}_{d\sim\mathcal{D}}[-\text{onehot}_K(\text{argmax}_k\theta_{dk}^*) \cdot \log f(d)]\,,$$

で表される．ここで，$\text{onehot}_K(i)$ は，one-hot 表現，即ち $i$ 番目の要素のみが 1 で残りが 0 であるような $K$ 次元のベクトルを表す．

混合ソフトラベルの場合は，VDCNN の最終層が上記 2 つの場合と異なり，各トピック分布ごとに別々に重みが存在するものとする．VDCNN の学習目標となる $N$ 個のトピック分布を $\theta^{(1)}, \cdots, \theta^{(N)}$ とする．これらは，LDA の学習結果，つまりトピック分布の事後確率より近似的にサンプルされたものである．また，それぞれのトピックの次元数は異なって良いものとする．VDCNN の最終層における，各トピック分布ごとの重み (バイアスを含むものとする) を $W_F^{(1)}, \cdots, W_F^{(N)}$ とする．また，文書 $d$ を入力したときの VDCNN の各最終層の出力値を $f(W_F^{(1)}, d)$ で表す．このとき，混合ソフトラベルの場合の目的関数は次式：

$$\begin{aligned} O_{\text{MixSoft}}(D) &= \mathop{\mathbb{E}}_{d\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{i}\left[H(\theta_d^{(i)}, f(W^{(i)}, d))\right]\right] \\ &= \mathop{\mathbb{E}}_{d\sim\mathcal{D}}\left[-\frac{1}{N}\sum_{i=1}^{N}\theta_d^{(i)} \cdot \log f(W_F^{(i)}, d)\right] \end{aligned}$$

で与えられるものとする．

### 3.3 再学習

前節の事前学習によって得られた VDCNN の重みを利用して，データセット中の教師ラベルを学習目標として再学習を行う．利用の仕方は，具体的には，最終出力層 (全結合+softmax 関数) の重み以外を，事前学習によって得られた VDCNN の重みによって初期化し，ネットワーク全体に対して再学習を行う．一方で，VDCNN の最終出力層の重み行列 $W_F$ は，適切なサイズを持ち，区間内の一様分布など一定の分布からとった初期値を持つ，新しいもの ($W_F^{\text{new}}$) に置き換える．これは，事前学習で用いたトピックと教師ラベルとの間は，特にそれらの id の順序において，全く関連がないためである．また，トピックの数自体も，教師ラベルのカテゴリ数とは異なるため，行列としての形自体，事前学習のものと再学習のものとでは異なる．直感的に言って，事前学習に用いるトピックにはなるべく多くの情報を含むようにしたいため，トピック数は，データセットによって決まっているカテゴリ数とは関係なく，十分多くとるように設計したほうが良いと考えられる．教師ラベルのカテゴリ数を $M$，文書 $d$ に対する教師ラベルを $t_d$ として，前節と同様の形で再学習の目的関数 ($O_{\text{Re}}$) を書くと，

$$O_{\text{Re}}(D) = \mathop{\mathbb{E}}_{d\sim\mathcal{D}}[-\text{onehot}_M(t_d) \cdot \log f(W_F^{\text{new}}, d)]$$

となる．

### 3.4 半教師あり学習

実際のテキスト分類の問題においては，文書の数自体は十分に多いものの，教師となる分類結果のラベルは少数しか用意できないことがしばしばある．我々の提案手法では，再学習の段階以外では教師ラベルを用いない．したがって，トピック分布の学習と事前学習では全データを用いて，再学習のときのみラベルつきデータを用いて学習を行うことができる．つまり，半教師あり学習 [2] を行うことができる．$D' \subset D$ をデータセット中のラベルのある文書集合とすると，この場合の再学習の目的関数は，$O_{\text{Re}}(D')$ とかける．本論文では，実際に，実験を通して，再学習のときに用いるデータの数を制限し，提案手法の半教師あり学習としての効果を検証する．

## 4. 実験

3 つの文書分類タスクに対する，予測結果のエラー率 (1-正解率) で比較を行う．

### 4.1 データセット

実験の対象としたデータセットを表 1 に示す．

AGNews データセットは，ニュース記事のカテゴリ分類タスクのデータセットである．DBPedia データセットは，DBPedia 記事のオントロジー分類タスクのデータセットである．Yelp Full データセットは，レビュー文の評価分類のデータセットである．上記のデータセットは，いずれも Zhang らによる Character-level CNN[17] で公開されているデータを使用した．

表 1　使用したデータセット

| データセット名 | 訓練用データ数 | 検証用データ数 | カテゴリ数 |
|---|---|---|---|
| AGNews | 120,000 | 7,600 | 4 |
| DBPedia | 560,000 | 70,000 | 14 |
| Yelp Full | 650,000 | 50,000 | 5 |

### 4.2 VDCNN の実装

Conneau らによる VDCNN [3] を，深層学習フレームワークである Chainer[*1] を使用して再現実装を行った．本研究では，ネットワーク全体での畳み込み層の数が 17 層のモデル (VDCNN17) を使用する．ただし，畳み込み層や埋め込み層などの重みの初期値の確率分布を Chainer のデフォルトのものにすると，[3] にある精度を再現できなかったため，次のように変更する．VDCNN の初期値は，畳み込み層は He の初期値 [6] で初期化し，それ以外の層に関しては，別の深層学習フレームワークである Torch[*2] のデフォルト初期値と同様に初期化する．

なお，[3] では，検証用データにおいて精度が下がったときにステップサイズを半減しており，その条件で再現実験を行い，我々の実装でも同程度の精度が達成できることを確認した．しかし，検証用データを見てステップサイズを変更すると，厳密には正しい精度が計測できていないと考えられる．そのため，本論文では，後述のように，3 epoch ごとにステップサイズを半減させたため，精度は [3] よりもわずかに低くなっている[*3]．

### 4.3 比較対象

事前学習無しでの VDCNN17(Baseline) をベースラインとし，トピック分布を用いた事前学習の効果を検証する．まず，トピックモデルのトピック数を 100 とし，トピック分布の推定を一度だけ行い，事前学習の手法として，ソフトラベルを用いた事前学習 (Soft) と，ハードラベルを用いた事前学習 (Hard) を比較対象として実験を行う．また，事前学習を行う際に，トピック数が 20，100，200 の，三つのトピック分布を用いた混合ソフトラベルを同時に学習した場合 (Mix-Soft) に関しても比較を行う．

また，半教師あり学習の枠組みでの効果を検証するために，再学習をする際の訓練データ数をそれぞれ 1/4，1/16，1/64，1/256，1/1024 とした際の比較をする．この実験では，ベースラインとソフトラベルを用いた事前学習の 2 つの手法に関して比較を行う．

[*1] https://chainer.org/
[*2] http://torch.ch
[*3] 例えば AGNews データセットに対しては，約 0.4% [3] に記載されている値よりもエラー率が高い．

### 4.4 学習パラメータ等の設定

VDCNN への入力となる文字の種類数は，VDCNN[3] の実験と同様に，小文字アルファベットや数字，その他記号から成る 69 種類とした．VDCNN の訓練時のバッチサイズは 128 とし，5000 ミニバッチで 1epoch として学習を行った．最適化手法は，MomentumSGD を使用し，慣性項を 0.9，学習率の初期値を 0.01 とした．学習率は，3epoch ごとに半減し，15epoch 学習を行った．

## 5. 実験結果

### 5.1 事前学習の効果と精度の比較

各枠組みと各データセットのテストデータに対するエラー率を表 2 に示す．ここでのエラー率は，各組み合わせごとに 5 回の実験を行い，各 epoch 終了時のテストデータに対するエラー率の平均をとり，エラー率が最小となった epoch のものを記載している．

いずれの事前学習ありの手法もベースラインよりも低いエラー率となっており，全てのデータセットにおいて，混合ソフトラベルを用いた事前学習 (Mix-Soft) が最小のエラー率を達成している．事前学習におけるトピック分布の扱いに関しては，ハードラベルよりもソフトラベルによる事前

表 2　テストデータに対するエラー率 (%)

| Framework / Data | AG | DBP | YelpF |
|---|---|---|---|
| Baseline | 9.20 | 1.38 | 36.26 |
| Soft | 7.95 | 1.24 | 35.36 |
| Hard | 8.53 | 1.34 | 35.94 |
| Mix-Soft | **7.64** | **1.21** | **35.31** |

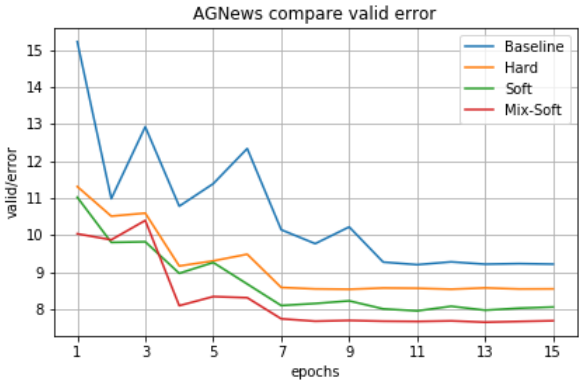


図 2　AGNews のエラー率 (%) の推移

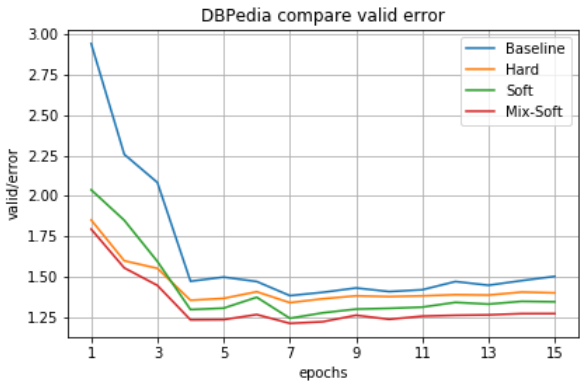


図 3　DBPedia のエラー率 (%) の推移

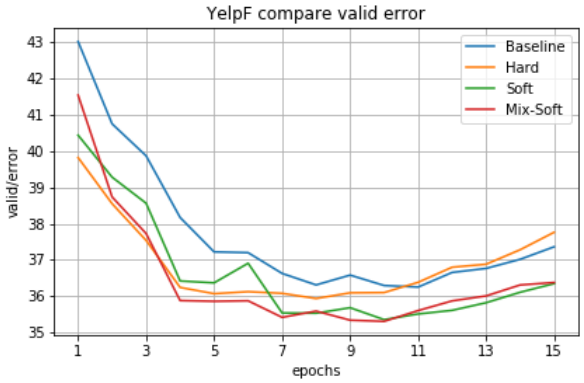


図 4　Yelp Full のエラー率 (%) の推移

学習が良い結果となっている．また，Baseline と Mix-Soft のエラー率を比較すると，AGNews では 1.56%，DBpedia では 0.17%，Yelp Full では 0.95%改善している．

各データセットごとのエラー率の推移を図 2-4 に示す．

事前学習ありの枠組みでは，学習の初期段階から低いエラー率となっている．また，Yelp Full データセットにおいては，Soft と Mix-Soft の差が小さく，Mix-Soft による事前学習の効果が小さいことが読み取れる．

### 5.2　半教師あり学習に対する提案手法の効果

次に，再学習時の訓練データ数を減らした際のエラー率を表 3-5 と図 5-7 に示す．ここでのエラー率は，データサイズ 1/1 に関しては，表 2 のものと同じ値である．それ以外のデータサイズについては，各組み合わせごとに 1 回ずつ実験を行い，各 epoch 終了時のテストデータに対する最小のエラー率のうち，エラー率が最小となった epoch のものを記載している．

全てのデータセット，全ての訓練データサイズにおいて，ソフトラベルによる事前学習のエラー率が下回っている．

表 3　AGNews のエラー率 (%)

| | 1/1 | 1/4 | 1/16 | 1/64 | 1/256 | 1/1024 |
|---|---|---|---|---|---|---|
| Baseline | 9.20 | 11.86 | 17.63 | 27.02 | 47.68 | 67.58 |
| Soft | **7.95** | **10.11** | **14.00** | **17.63** | **17.18** | **21.90** |

表 4　DBPedia のエラー率 (%)

| | 1/1 | 1/4 | 1/16 | 1/64 | 1/256 | 1/1024 |
|---|---|---|---|---|---|---|
| Baseline | 1.38 | 1.86 | 2.80 | 5.60 | 14.17 | 33.91 |
| Soft | **1.24** | **1.61** | **2.17** | **3.11** | **4.70** | **5.82** |

表 5　Yelp Full のエラー率 (%)

| | 1/1 | 1/4 | 1/16 | 1/64 | 1/256 | 1/1024 |
|---|---|---|---|---|---|---|
| Baseline | 36.26 | 39.83 | 45.31 | 51.42 | 58.66 | 69.08 |
| Soft | **35.36** | **39.62** | **43.67** | **48.21** | **53.27** | **58.48** |

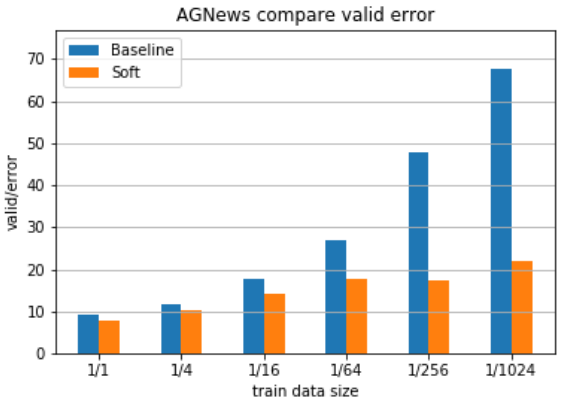


図 5　AGNews のエラー率 (%) の比較

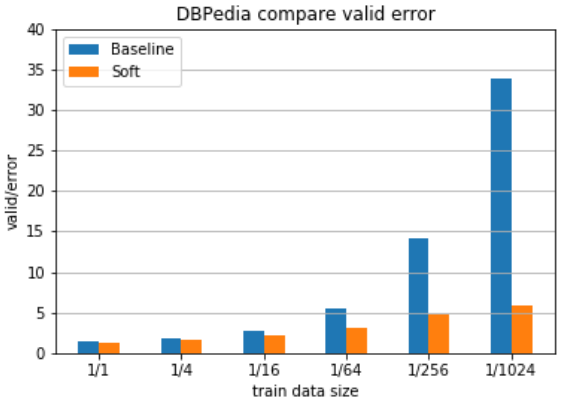


図 6　DBPedia のエラー率 (%) の比較

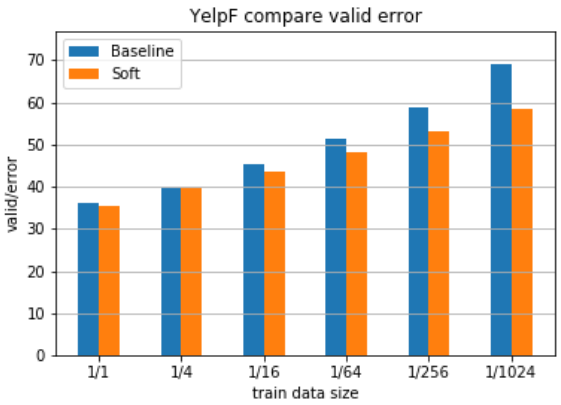


図 7　Yelp Full のエラー率 (%) の比較

訓練データサイズを減らすほどに，ベースラインとソフトラベルによる事前学習のエラー率の差が大きくなっている．特に，AGNews においては，訓練データサイズを 1/1024 とするということは，わずかに 117 個のラベル付きデータしか再学習に用いないということを意味している．それにもかかわらず，提案手法を用いると，8 割近くの正解率 (78%) を達成でき，一方で，ベースラインの正解率 (32%) はチャンスレベルである 25%に近い値となっている．したがって，特に教師ありのデータ数が少ない場合においては，極めて大きく精度が改善していることがわかる．

## 6. おわりに

本研究では，深層ニューラルネットの事前学習の枠組みとして，トピック分布を用いた事前学習を提案した．実験の結果，混合ソフトラベルを用いた事前学習が最も有効であるということがわかった．また，半教師あり学習の枠組みでは，訓練データサイズが少なくなるほど，ソフトラベルによる事前学習が有効であることを示した．本提案手法を用いると，例えば，新たに集められた文書セットに対して独自のカテゴリに分けるような自動分類を行うようなタスクに対しても，少数のラベルをつけるだけで，高い精度を達成できることが期待できる．

### 参考文献

[1] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022 (2003).
[2] Chapelle, O., Scholkopf, B. and Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], *IEEE Transactions on Neural Networks*, Vol. 20, No. 3, pp. 542–542 (2009).
[3] Conneau, A., Schwenk, H., Barrault, L. and Lecun, Y.: Very deep convolutional networks for text classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 1107–1116 (2017).
[4] Ehteshami B. B., e. a.: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer, *The Journal of the American Medical Association*, Vol. 318, No. 22, pp. 2199–2210 (2017).
[5] Gatys, L. A., e. a.: Image style transfer using convolutional neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423 (2016).
[6] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034 (2015).
[7] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).
[8] Howard, J. and Ruder, S.: Universal language model fine-tuning for text classification, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 328–339 (2018).
[9] Johnson, R. and Zhang, T.: Deep pyramid convolutional neural networks for text categorization, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 562–570 (2017).
[10] Kim, Y.: Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* (2014).
[11] Liu, J. S.: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem, *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 958–966 (1994).
[12] Liu, X., Gao, J., He, X., Deng, L., Duh, K. and Wang, Y.: Representation learning using multi-task deep neural networks for semantic classification and information retrieval (2015).
[13] Mintz, M., Bills, S., Snow, R. and Jurafsky, D.: Distant supervision for relation extraction without labeled data, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, Association for Computational Linguistics, pp. 1003–1011 (2009).
[14] Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L. and Jin, Z.: How Transferable are Neural Networks in NLP Applications?, *arXiv preprint arXiv:1603.06111* (2016).
[15] Teh, Y. W., Newman, D. and Welling, M.: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems 19*, MIT Press, pp. 1353–1360 (2007).
[16] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H.: How transferable are features in deep neural networks?, *Advances in neural information processing systems*, pp. 3320–3328 (2014).
[17] Zhang, X., Zhao, J. and LeCun, Y.: Character-level convolutional networks for text classification, *Advances in neural information processing systems*, pp. 649–657 (2015).