BERT を用いた文書分類タスクへの Mix-Up 手法の適用

菊田尚樹
茨城大学工学部
情報工学科
17t4028n@vc.ibaraki.ac.jp

新納浩幸
茨城大学大学院理工学研究科
情報科学領域
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

1 はじめに

自然言語処理のタスクを機械学習のアプローチで解決しようとした場合，訓練データの構築コストが高いことが問題になる．このため従来より様々な試みがなされている．その中で近年のトピックの一つとしてデータ拡張 (Data Augmentation)[1] がある．データ拡張手法は大きく加工と生成の 2 種類に分けられる．例えば画像の識別であれば，訓練データ内の画像を反転させたり，切り取ったりした画像であってもその画像のラベルに変化はないので，そのようにして加工した画像を訓練データに追加することで訓練データを増やすことができる．あるいは GAN などを利用して人工的なデータを生成する手法もデータ拡張の一種といえる．データ拡張の生成手法の一つである Mix-Up 手法 [2] は，簡単に実装でき効果も高いことが知られている．Mix-Up 手法は画像識別を対象にした手法であるが，自然言語処理にも応用が可能である [3].

本論文では BERT [4] を用いた文書分類タスクに Mix-Up 手法を試みる．具体的には BERT は 2 文の入力が可能であるため，ラベルの異なる 2 文書を組み合わせて入力し，マルチクラスの出力を教師データとする．livedoor ニュースコーパスを用いた実験で提案手法の有効性を示した．

2 関連項目

2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers の略) は 2018 年に Google が作成して以降利用が注目されている高性能な事前学習済みモデル [4] であり，分類，単語予測，文脈判断などに活用ができる．本研究では Mix-Up 手法を利用することで，BERT を用いた文書分類タスクの精度向上を試みる．

2.2 Mix-Up 手法

Mix-Up とは，2017 年に Hongyi Zhang により発表された画像分野でのデータ拡張手法である [2]．画像データに関しては式 1，ラベルに関しては式 2 によってデータ拡張を行う．

$$x = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$y = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

x は画像データのベクトル，y はラベルの one-hot ベクトルである． λ は混合の割合を表わしている．

2.3 nlp に Mix-Up 手法を用いた先行研究

2019 年，Hongyu Guo により，Mix-Up 手法を nlp に用いた研究が行われた [3]．Mix の方法は前節の画像分野での方法と同じく，式 1 と式 2 によって行っている．nlp の場合は，x が単語の埋め込み表現または文の埋め込み表現である．

(例) 6:4 の割合で Mix-Up する場合

文書 1 のベクトル

[0.2, -0.3, 0.5, ⋯]

文書 2 のベクトル

[0.4, 0.1, -0.5, ⋯]

新たな文書のベクトル

[0.2, -0.3, 0.5, ⋯] × 0.6 + [0.4, 0.1, -0.5, ⋯] × 0.4
= [0.28, 0.22, 0.1, ⋯]

3 提案手法

先行研究での Mix-Up 手法を BERT での文書分類にも採用するとなると問題が生じてしまう．先行研究の手法では，NN の学習以前に文書の特徴ベクトルを作っておく必要があるが，BERT を利用した文書分類では NN の学習プロセスで文書の特徴ベクトルを得るため，先行研究とは順序が異なる．仮に先行研究の手法を採用するとなると，BERT による特

徴ベクトルの計算部分は学習の外でやることになり，分類精度が期待できない．(図 1 参照)

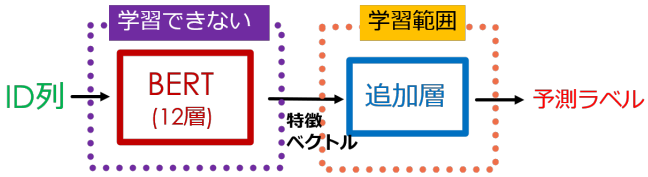

図 1　先行研究手法を BERT で採用する場合の学習範囲

そこで，本論文では以下の手法を提案する．

3.1　データの Mix 方法

2 つの文書において，BERT に入力する際の ID 列同士を連結するという手法である． id"2"は文の先頭を表し，後続の ID 列には不要なので除いた．この手法であれば文書の特徴ベクトルを得る BERT 部分の学習が可能である．(図 2 参照) また，BERT の入力最大列長は 512 のため，それを超えないように前半と後半の ID 列長はそれぞれ 252 までの制限をかけた．

(例)

1 つ目の ID 列

[2, 6259, 9, 12396, 14, 3596, 3]

2 つ目の ID 列

[2, 11475, 9, 3741, 5, 12098, 75,]

Mix-up 後の新たな ID 列

[2, 6259, 9, 12396, 14, 3596, 3, 11475, 9, 3741, 5, 12098, 75, 3]

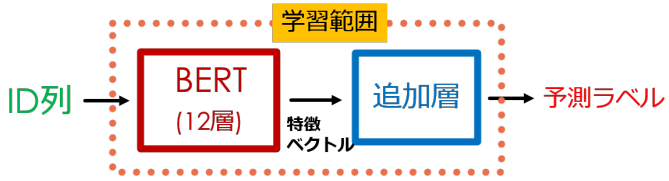

図 2　本研究手法での学習範囲

3.2　ラベルの Mix 方法

各ラベルをまず 0 と 1 からなる one-hot ベクトルで表現した．Mix させる場合は 0．5 が 2 箇所で残りが 0 のベクトルを作成した．Mix 後のラベルが 0.5 同士なのは 2 つの文書が等価値に混ざっていることを表している．

(例)

ラベル 3

[0, 0, 0, 1, 0, 0, 0, 0, 0]

ラベル 6

[0, 0, 0, 0, 0, 0, 1, 0, 0]

ラベル 3 と 6 の Mix

[0, 0, 0, 0.5, 0, 0, 0.5, 0, 0]

4　実験

4.1　概要

Mix-Up 手法を用いることで訓練データ数を増加させた状態での BERT による文書分類と，BERT を使った通常の文書分類の精度を比較し，本研究手法の効果を調べた．

4.2　条件

4.2.1　実行環境

実験は Google Colaboratory の GPU 環境を利用して行った．

4.2.2　使用した BERT モデル

BERT は，東北大学 乾・鈴木研究室が作成した日本語版 BERT の事前学習モデル[1] のうちの一つである，「bert-base-japanese-whole-word-masking」を使用した．

4.2.3　使用したコーパス

livedoor ニュースコーパス[2] を使用し，以下の 9 ジャンルに対しての文書分類を行った．

- ラベル 0 : 独女通信
- ラベル 1 : IT ライフハック
- ラベル 2 : 家電チャンネル
- ラベル 3 : livedoor HOMME
- ラベル 4 : MOVIE ENTER
- ラベル 5 : Peachy
- ラベル 6 : エスマックス
- ラベル 7 : Sports Watch
- ラベル 8 : トピックニュース

4.3　実験手順・手法

4.3.1　データの準備

本実験では livedoor ニュースコーパスより 6623 個の記事 (本文) を抽出し，表 1 のように振り分けた．

1)　https://github.com/cl-tohoku/bert-japanese
2)　https://www.rondhuit.com/download.html#ldcc

表 1　使用データの内訳

ラベル	train	val	test	sum
0	87	128	566	781
1	87	125	571	783
2	86	111	581	778
3	51	72	335	458
4	87	114	582	783
5	84	106	565	755
6	87	106	590	783
7	90	131	589	810
8	77	107	508	692
sum	736	1000	4887	6623

train は訓練時に用いたデータ，val はハイパーパラメータを調整するために分類精度をこまめに確認する際に使用した検証用データ，test は出来上がったモデルの最終的な分類精度を確認する際に使用したテストデータである．

また，以降の手順において各文書は BERT により形態素解析と ID 化を済ませた状態で扱っている．

4.3.2　訓練データの Mix-Up

訓練データについて，Mix-Up 手法を用いてデータ拡張を行った．Mix させる文書の選出方法は以下の 2 通りを試した．

Mix させる文書の選出方法 (1)

1 つ目の選定方法は全てのラベルの文書をランダムで Mix させるというものである．乱数を用いて 736 個の文書をランダムに並び替え，一番目から順に隣り合った 2 つの文書 (とそのラベル) を Mix させた．この方法の場合，文書間の隙間の数だけペアが作られるため，735(736-1) 個の拡張データができる．その結果訓練データは 736 個から 1471 個に拡張した．また，プログラムを起動するたびに選出される組み合わせは変わるようになっている．

Mix させる文書の選出方法 (2)

2 つ目の選定方法は，不足しているラベルの文書を対象に Mix-Up によってデータ数を増加させるというものである．本実験では表 1 に示す通り，訓練データではラベル 3 の文書が不足しているため，ラベル 3 の文書が必ず選定されるような動作にした．具体的には，51 個のラベル 3 文書からランダムに 1 つ選び，次にラベル 3 以外の 685 個のデータからもランダムに 1 つ選び，その 2 つの文書を Mix させるという手順を繰り返した．連結の順序はラベル 3 の文書が前半でラベル 3 以外の文書が後半である．この動作により 765 個の新たな訓練データを作成し，訓練データは 1501 個に拡張した．また，選出方法 (1) と同じようにプログラムを起動するたびに組み合わせは変わるようになっている．

4.3.3　作成した分類器

分類器として使ったニューラルネットワークのモデルは BERT の直後に全結合層を一層追加したものを用いた．長さ 512 以下の ID 列を BERT に入力し，768 次元の文書の特徴ベクトルを出力として得る．それを全結合層に入力し，9 次元の各ラベルに対する予測値を出力として得るという流れである．細かな設定を以下に示す．

損失関数

交差エントロピー：今回ラベルは one-hot 表現のため nn.CrossEntropyLoss() に直接入力することはできないため，nn.LogSoftmax() を利用し，交差エントロピーの定義通りに計算して損失を求めた．

最適化関数

確率的勾配降下法 (SGD) : 学習率は検証データを用いて分類精度と学習効率の両方を考慮した結果，0.01 を採用した．

訓練データのバッチサイズ

バッチサイズは実行環境である Google Colaboratory にて可能な最大値であった 10 を採用した．

エポック数

検証データでの分類精度の上がり幅を考慮した結果，10 エポック学習すれば精度は収束値に達していると判断した．

4.4　実験結果

通常版，選出方法 (1) の Mix-Up 版，選出方法 (2) の Mix-Up 版の 3 つにおいて，訓練データで 10 エ

ポック学習したモデルをそれぞれ 10 個用意し，テストデータに対する正解率を求めた．各モデルを比較した箱ひげ図を図 3 に示す．また，平均値の比較を表 2 に示す．※正解率は小数点第 4 位を四捨五入している．

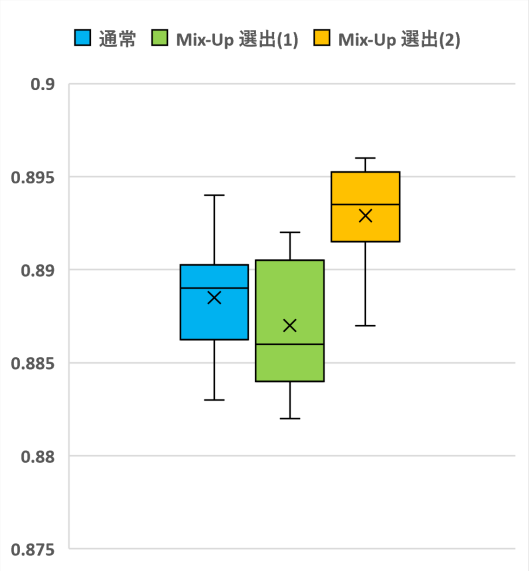

図 3　実験結果

表 2　平均値の比較

	正解率 (平均)
通常	0.889
Mix-Up 選出 (1)	0.887
Mix-Up 選出 (2)	0.893

精度の高い順に，選出方法 (2) の Mix-Up，通常の BERT，選出方法 (1) の Mix-Up という結果になった．

5　考察

選出方法 (1) では通常よりも精度が落ち，選出方法 (2) では精度が上がっていることから Mix させる文書の選定方法が精度に大きく影響を与えることがわかった．本実験からは，不足しているラベルを中心に Mix させてデータ拡張していくのが有効であると言える．様々な選出方法を試して，最終的な結論を出したい．

また，今回は Mix させる二つの文書を等価値 (割合を 0.5 : 0.5) にして実験を行ったが，連結させる ID 列長に差を設けてその割合をラベルの one-hot 表現に採用するといった手法も試してみる価値はあると考えている．

6　おわりに

本論文では BERT を用いた文書分類タスクに Mix-Up 手法を試みた．具体的には BERT は 2 文の入力が可能であるため，ラベルの異なる 2 文書を組み合わせて入力し，ラベルは one-hot ベクトルにて 0.5 の要素を二つ作ることで混合した．livedoor ニュースコーパスを用いた実験を行い，Mix-Up させる文書の選出方法によって精度に差が出ることを示した．

謝辞

本研究は JSPS 科研費 JP19K12093 および 2020 年度国立情報学研究所公募型共同研究（2020-FC03）の助成を受けています．

参考文献

[1] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, Vol. 6, No. 1, p. 60, 2019.

[2] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017.

[3] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers forlanguage understanding, 2018.