L2-constrained Focal Loss を導入した BERT による文書の著者推定

落合晃汰　　青野雅樹
豊橋技術科学大学 知能・情報工学課程
ochiai@kde.cs.tut.ac.jp, aono@cs.tut.ac.jp

1 はじめに

近年,SNS やネットニュースなどのネットワークサービスの発達により, 情報をインターネットから簡単に入手することができる. しかし, このようなインターネット上の情報には著者の情報が含まれていないことが多い. 著者の情報は情報の信憑性を判断する際に重要であるため, 文章の著者推定問題は重要なタスクである. 匿名で記述された文章の著者推定, 盗作や転載の検出, 類似性の高い文章の推薦など, 幅広く応用できる. 本研究では,50 人の著者によって書かれた企業や業界に関するニュースからなる C50 データセットと 53 人の著者によって書かれたツイートからなる PAN 2016 Author Profiling Task の Twitter データセットを用いて,L2-constrained Focal Loss を導入した BERT を用いた文章の著者推定手法を提案する.

2 関連研究

2.1 BERT

BERT(Bidirectional Encoder Representationsfrom Transformers) [1] は 2018 年に Google が発表した言語処理モデルである. 複数の双方向 Transformer に基づくエンコーダーであり, 単語列を入力として各単語に対応する分散表現を出力する. また,BERT は大規模コーパスを用いて教師なしの事前学習を行うことで性能を向上させている. 事前学習には MaskedLM (Masked Language Model) と次文予測 (Next Sentence Prediction) の 2 つのタスクを用いてモデルの学習を行っている. これら 2 つのタスクによって事前学習されたモデルに転移学習やファインチューニングを施すことで高い精度を実現している.

2.2 L_2-constrained Softmax Loss

L_2-constrained Softmax Loss[2] は 2017 年に顔識別の DCNN(Deep Convolutional Neural Network) の精度向上のために提案された損失関数である.L_2-constrained Softmax Loss では Softmax Loss にネットワーク出力の L2 ノルムがある定数 α になるように制約を加える. これによって同一クラスのコサイン類似度が大きく, 違うクラスのコサイン類似度が小さくなるように学習されるという特徴がある. L2-constrained Softmax Loss の式は以下のようになる.

$$\begin{aligned} &\text{minimize} - \frac{1}{M}\sum_{i=1}^{M} log\frac{e^{W_{y_i}^T f(x_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(x_i)+b_j}} \\ &\text{subject to} \quad \|f(x_i)\|_2 = \alpha, \forall i = 1, 2, ...M \end{aligned} \tag{1}$$

2.3 Focal Loss

Focal Loss[3] は 2017 年に FAIR(Facebook AI Research) が発表した物体検出モデルである RetinaNet に使われている損失関数である. Focal Loss は物体検出において背景と背景以外のクラス間の不均衡の問題を解決するために提案された. このクラス間の不均衡によって, 学習の殆どが簡単な背景判定に支配されてしまう. この問題を解決するため Focal Loss では簡単に分類が成功している事例の損失を小さくしている. これにより, より難しく注目すべき事例が学習に強く寄与するようになる. Focal Loss の式は以下のようになる.

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

このとき γ はどのくらい簡単に分類が成功している事例の損失を減衰するかを決定するパラメータで, これによって簡単に分類が成功している事例の損失への寄与が小さくなる. これによって, より難しく注目すべき事例が学習に強く寄与できるようになる.

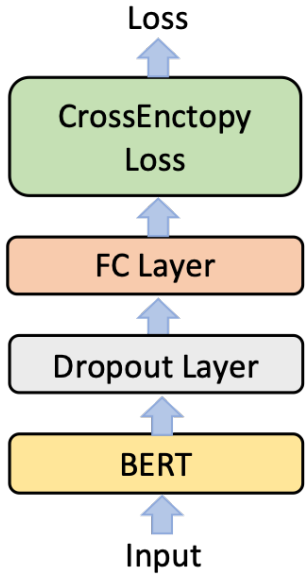

図 1 BERT ベースラインモデル

3 ベースラインモデルと提案手法

以下では提案する手法及びベースラインモデルについて説明する.

3.1 LSTM ベースラインモデル

LSTM ベースラインモデルは fastText[4] を Embedding layer として用いた LSTM モデルである.fastText の分散表現は UMBC Webbase コーパス, および statmt.org ニュースデータセットでトレーニングされた 100 万語のベクトルとなっている. モデルは 1 層の単方向 LSTM であり, 隠れ層の次元は 768 次元で分散表現の次元数は 300 次元とした.

3.2 BERT ベースラインモデル

本研究で使用する BERT は Transformer が 12 層でトークンの次元数は 768 となっているモデルで,Wikipedia コーパスと BooksCorpus で事前学習されたものを使用している [5]. BERT ベースラインモデルはこの学習済み BERT をファインチューニングしたモデルである. 図 1 にモデルの概要を示す.

3.3 L2-constrained Softmax Loss を導入した BERT モデル

BERT ベースラインモデルに変更を加えたモデルとして L2-constrained Softmax Loss を導入したモデルを構築する. 出力の L2 ノルムが定数 α になるように制約を加えて,Softmax Loss を用いることで実現する.

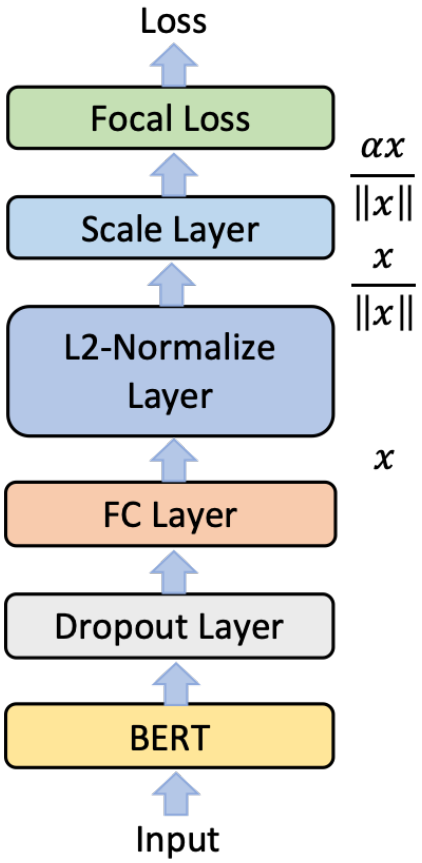

図 2 L2-constrained Focal Loss を用いた BERT モデル

3.4 Focal Loss 導入した BERT モデル

Focal Loss を導入した BERT モデルでは BERT ベースラインモデルで使用している Softmax Loss を Focal Loss に変更することで実現する.

3.5 L2-constrained Focal Loss を導入した BERT モデル

さらに変更を加えたモデルとして,L2-constrained レイヤーと Focal Loss を導入したモデルを構築する.3.3 節の L2-constrained Softmax Loss を用いた BERT モデルで使用している Softmax Loss を Focal Loss に変更することで実現する. 図 2 にモデルの概要を示す.

4 実験

提案した BERT モデルを使用して文書の著者を推定する.

4.1 データセット

本研究では,C50 dataset[6] と Twitter データセット [7] の 2 つのデータセットを使用する.

表 1　各モデルの実験結果

System	C50		Twitter	
	Parameter	Accuracy(%)	Parameter	Accuracy(%)
Combined Features (Bag-of-Words + Character 3-grams + Noun-cluster based Features) (LIBSVM)	-	64.49	-	-
Modified Tree + Combined Features (Bag-of-Words + Character 4-grams + Noun-cluster based Features) (SVM-Light)	-	-	-	55.26
LSTM	-	43.96	-	27.40
$BERT_{base}$ (Softmax Loss)	-	67.20	-	75.04
$BERT_{base}$ (L2-constrained Softmax Loss)	α=20	69.60	α=12	76.09
$BERT_{base}$ (Focal Loss)	γ=3	69.60	γ=2	75.91
$BERT_{base}$ (L2-constrained Focal Loss)	α=20,γ=2	**70.32**	α=20,γ=2	**76.32**

4.1.1　C50 データセット

C50 データセットは RCV1(Reuters Corpus Volume 1) の一部からなるロイターニュースの記事のデータセットである. 企業及び産業について書かれた英語のニュース記事のデータセットであり,50 人の著者によって書かれた記事がまとめられている. 訓練データは著者あたり 50 本の合計 2500 本の記事で構成され, テストデータには訓練データと重複しない著者あたり 50 本の 2500 本が使用されている. C50 データセットの最大トークン数は BERT の tokenizer で分割した際 1874 語である.BERT で読み込める最大トークン数は 512 語であるため, そのまま読み込むことができない. そのため本実験では先頭の 512 語のみを使用する.

4.1.2　Twitter データセット

Twitter データセットは PAN 2016 Author Profiling Task データセットのツイートからなる. PAN 2016 Author Profiling Task データセットから,53 人のユーザーを収集し, ユーザーあたり 900 ツイートを使用する. 訓練データは著者あたり 800 のツイートの合計 42400 のツイートで構成され, テストデータには訓練データと重複しないユーザーあたり 100 のツイートの 5300 のツイートが使用されている.

4.2　評価指標

データ数は各クラス間で同じであるため, 評価指標には Accuracy のみを用いる.

4.3　モデルパラメータ

最適化には AdamW を用い, ミニバッチサイズは 8, ドロップアウト率は 0.1 とした. 20 エポック学習を行った際のモデルの性能を評価する.

4.3.1　パラメータ調整

L2-constrained Softmax Loss を導入した BERT モデル,Focal Loss を導入した BERT モデル, L2-constrained Focal Loss を導入した BERT モデルについてはそれぞれパラメータ調整を行った. L2-constrained レイヤーのパラメータ α は 8 から 32 までを 4 刻みで変更して実験を行い,Focal Loss のパラメータ γ は 1 から 5 までを 1 刻みで変更して実験を行った.

4.4　実験結果

節 3 で示した 5 つの手法について実験を行った. なお L2-constrained Softmax Loss と Focal Loss のパラメータの値は節 4.3.1 で示した方法で実験し, 各手法で最高性能を示した値を使用する. SVM(support vector machine) を使用した著者推定の先行研究 [8] の結果と本実験の結果をまとめたものを表 1 に示す. LSTM ベースラインモデルでは SVM を用いた先行研究の精度を上回ることはできなかった. しかし,BERT ベースラインモデルは両方のデータセットで先行研究を上回った. また, 両方のデータセットにて L2-constrained Softmax Loss を導入した BERT モデルと Focal Loss を導入した BERT モデルは BERT ベースラインモデルよりも分類精度は向上した. さらに,L2-constrained Focal Loss を導入した BERT モデ

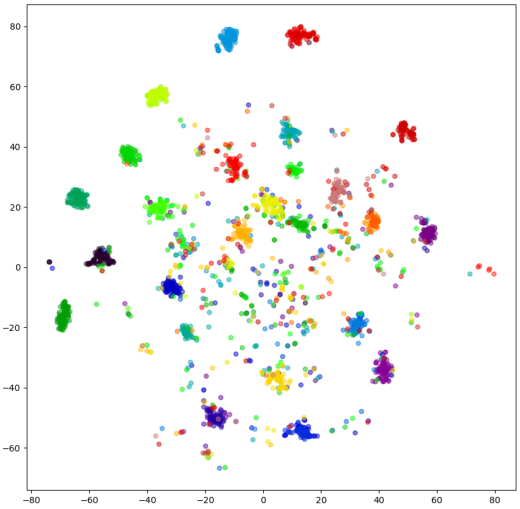

図 3 Twitter データセットでの BERT モデルの分散表現

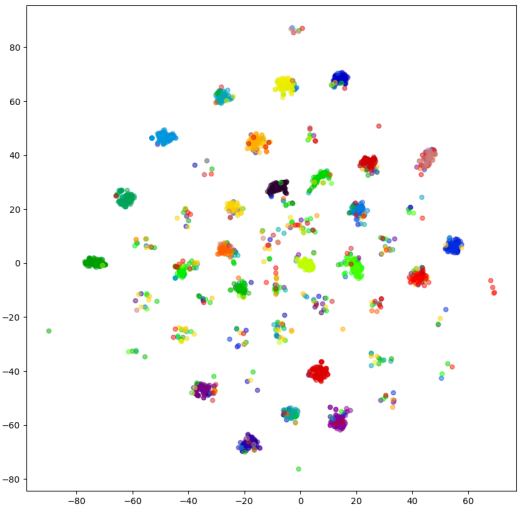

図 4 L2-constrained Focal Loss を導入した BERT モデルの分散表現

ルはより高い精度となった. 図 3, 図 4 に Twitter データセットでの BERT ベースモデルの出力 ([CLS] トークンの 768 次元の分散表現) と L2-constrained Focal Loss を導入した BERT モデルの出力をそれぞれ t-SNE により次元圧縮したものを示す.

4.5 考察

表 1 より,L2-constrained Softmax Loss を導入した BERT モデルと Focal Loss を導入した BERT モデルそれぞれについて BERT ベースラインモデルと比較して精度が向上することが確認できた. また,L2-constrained Softmax Loss を導入した BERT モデルと Focal Loss を導入した BERT モデルそれぞれの精度は C50 データセットについては 69.60%で同じであり,Twitter データセットについては 0.18%の差のみであった. そのため, 最高精度を記録した L2-constrained Focal Loss を BERT に導入したモデルでもバランス良く作用して精度の向上につながっていると考えられる. また, 図 3, 図 4 より L2-constrained Focal Loss を BERT に導入することで分散表現についても BERT ベースラインモデルと比較して各ラベルがよりまとまったクラスタを形成していることが確認でき, これにより分類の精度が向上していると考えられる.

5 おわりに

本研究では,L2-constrained Softmax と Focal Loss を導入した BERT を用いた文書の著者推定モデルを提案した. 実験では, ベースラインと比較し精度が向上した. さらに,BERT の分散表現についても各ラベルがよりまとまったクラスタを形成するようになっていることが確認できた. 今後の課題としては BERT の複数の層の出力の平均や和を利用したモデルや BERT から発展した Transformer ベースのモデルについても L2-constrained Focal Loss を利用することができないかについての検討も行っていきたいと考えている.

謝辞

本研究の一部は, 科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した.

参考文献

[1]Jacob Devlin, Ming-Wei Chang, Keton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[2]Rajeev Ranjan, Carlos D. Castillo, and Rama Chellappa. 2017. L2-constrained Softmax Loss for Discriminative Face Verification. arXiv preprint arXiv:1703.09507.

[3]Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. arXiv preprint arXiv:1708.02002.

[4]English word vectors · fastText. https://fasttext.cc/docs/en/english-vectors.html. Accessed: 2021-01-05.

[5]bert-base-uncased · Hugging Face. https://huggingface.co/bert-base-uncased. Accessed: 2021-01-05.

[6]C50 dataset. https://archive.ics.uci.edu/ml/datasets/Reuter_50_50. Accessed: 2021-01-05.

[7]Pan 2016 author profiling task. http://pan.webis.de/clef16/pan16-web/author-profiling.html. Accessed: 2021-01-05.

[8]Shofi Nur Fathiya. Author Identication Focusing on Semantic and Syntactic Features Extracted from Long and Short Texts. Toyohashi University of Technology, 2017.