深層距離学習で最適化した文の埋め込み表現による未知クラスを含む日本語テキスト分類

大段 秀顕　　林 岳晴　　竹中 一秀　　湯浅 晃
株式会社 NTT データ
{ Hideaki.Odan, Takeharu.Hayashi, Kazuhide.Takenaka, Akira.Yuasa}@nttdata.com

概要

一般に「学習データに存在しない未知のクラス」を含む分類は難しいが，深層距離学習により特徴空間上でのクラスごとの分布を最適化すれば，分類精度の向上が期待される．本研究では未知クラスを含む日本語テキストの分類問題に対して，深層距離学習で最適化した BERT ベースの文の埋め込み表現を用い，平均 k 近傍距離による分類を行った．これにより BERT 文書分類の確信度ベースの手法や，他の文の埋め込み表現モデルを使用した場合と比較して，高い精度を達成した．また一般に難しいとされる「1 クラスあたりのデータ量が少ない場合」「不均衡な場合」「クラス数が多い場合」であっても他の手法を上回る精度が得られ，実用性の高さを示した．

1 はじめに

一般的な機械学習の分類モデルでは，評価データには「学習データに存在した既知クラス」しか存在しないことを暗黙の前提としている．そのため評価データに「学習データに存在しない未知クラス」が存在し，「対象データが既知クラスか未知クラスを判定したうえで正しく分類するタスク」は，通常の分類モデルとは異なるアプローチが必要になる[1]．実際，これに対するナイーブな手法として「分類モデルの確信度スコアが十分低い場合に未知クラスとしてみなす方法」が考えられるが，確信度スコアが過剰に高くなるケースが知られており[2]，高い精度での分類は難しいと想定される．

そこで本研究では画像領域でよく使用される深層距離学習に注目した．深層距離学習では特徴空間上で各クラスの分布を最適化し，互いにより識別可能になるような特徴量抽出が可能になる．ここで未知クラスの識別が既知クラスの識別と同じ観点の延長線上で可能と仮定するならば，深層距離学習により未知クラスを含む分類の精度向上が期待される．

本研究の新規性は，未知クラスを含む日本語テキスト分類タスクにおいて深層距離学習で最適化した文の埋め込み表現を用い，既存手法との比較検証を行ったことである．本提案手法で上記タスクを検証した文献は，著者の知る限り存在しない．具体的には，深層距離学習の損失関数 AdaCos[3]を用いて，文の埋め込み表現モデル Sentence-BERT (SBERT)[4]を訓練し，最適化したベクトルに対して平均 k 近傍法による分類を行い，性能を評価した．またデータ特性に応じた性能の変化も併せて調査した．

2 関連研究

2.1 文の埋め込み表現生成モデル

文の埋め込み表現（Sentence Embedding）生成モデルとは，文の意味を数値的なベクトルに変換するモデルのことである．最も単純なものとして TFIDF や，事前学習した単語の埋め込み表現 (word2vec, Glove[5], fastText[6] など) を加重平均する方法[7][8][9]がある．このほか転移学習可能なユニバーサルな表現を得ることを目的とした研究が多く存在する．例えば InferSent[10], Universal Sentence Encoder (USE)[11], SBERT などである．いずれも Stanford Natural Language Inference という自然言語推論用データセットで教師あり学習を行っているが，モデルの構造が異なる．InferSent は双方向 LSTM，USE では Transformer もしくは DAN，SBERT では BERT を用いている．論文[4]では Semantic Textual Similarity タスクにおいてこれら手法を比較しており，SBERT が最も高い精度となっている．

2.2 深層距離学習

深層距離学習（Deep Metric Learning）とは，特徴

空間上で同じクラスをより近くに，異なるクラスをより遠くに配置するような変換を学習させる手法のことである．このため互いにより識別可能な特徴量抽出が可能になる．主な応用タスクとして，画像領域での顔認証[3][12] [13]や異常検知[14]などがある．

深層距離学習の損失関数には複数手法が提案されている．基本的なものとして Contrastive Loss, Triplet Loss が挙げられるほか，最近では分類モデルのアーキテクチャをそのまま適用しながら Softmax 関数を改良した角度ベースの損失関数（CosFace[12] , ArcFace[13] , AdaCos など）も提案されている．

3 提案手法

まず文の埋め込み表現生成モデルとして，高い精度を持つ SBERT を用いる．ただし元論文[4]と異なり，自然言語推論データセットではなく、対象タスクのデータセットを用いた深層距離学習によるファインチューニングを行う．これはユニバーサルな表現よりも，特定のタスクに特化した表現の方が高い精度を実現できると考えたためである．

次に深層距離学習の損失関数についても，最適なものを選択した．まず Contrastive Loss, Triplet Loss は学習過程のサンプル選択依存性が高く，収束や精度が安定しないため採用しなかった．一方で角度ベースの損失関数の多くはハイパーパラメータ依存性が高い．今回は中でもハイパーパラメータを自動調整できる AdaCos を採用した．

最後に深層距離学習で最適化された埋め込み表現を平均 k 近傍距離で分類する．具体的には，まず学習・テストデータの埋め込み表現を出力し，特徴空間上にマッピングする．その後，各テストデータ点の学習データ点に対する平均 k 近傍距離を算出する．平均 k 近傍距離が閾値以上のものを未知クラスとし，それ以外については k 近傍法による多数決で所属クラスを推論した．なお近傍点数 k として k=min(N, 5) を採用した（N は最小クラスサイズ）．

4 実験設定

4.1 データセット

本研究ではデータ特性に応じた性能評価を行うため，表 1 に示す 4 種類のデータセットを用意した．

まず共通するデータソースとして Yahoo! 知恵袋データ（第 2 版）[15]を用いた．これは，ヤフー株式会社が国立情報学研究所に提供しているデータセットで，大/中/小分類カテゴリ・質問文・回答文等から構成される．本データセットは量が多いものの，整っていない日本語テキストも多く含まれるため，なるべく品質が高いと思われる「質問ステータス：質問者による知識」「役立つ質問」「BA 不適合投票率：0」「回答ステータス：質問者による BA」を満たすデータのみを使用した．また分類対象テキストとしては質問文と回答文を結合したものを用い，正解ラベルは中分類カテゴリを選択した．

この共通データソースに対して，さらに既知/未知クラスの量を調整することで，4 種類のデータセットを準備した．これらは実験#1 をベースラインに，以下を評価するために用意している：

- 実験#2：クラス当たりのデータ量が少ない場合
- 実験#3：クラスごとのデータ量が不均衡な場合
- 実験#4：クラスごとのデータ量が非常に不均衡で，ごく少量データのクラスも存在し，かつ全体のクラス数も多い場合（最も難しい場合）

4.2 前処理

ノイズ除去のため，「改行コードの除去 」「HTML 特殊文字コードの復元」「HTML タグの除去」「URL 文字列の除去」「ファイルパス文字列の除去」「正規化」「数字のゼロ置換」を実施した．

4.3 モデル

実験では提案手法のほかに，以下の目的のもと，比較用モデルを 5 つ用意した．

- 未学習 SBERT : SBERT のモデル構造ではなく，深層距離学習自体の効果を測るため．
- TFIDF, fastText, USE：ユニバーサルな文の埋め込み表現生成モデルと比較するため．
- BERT 文書分類：分類モデルによるナイーブな実現手法と比較するため．

これらのうち BERT 文書分類以外のモデルについては，提案手法と同様に埋め込み表現を平均 k 近傍法によって分類した．一方で BERT 文書分類は異なる手法で分類を行ったので後述する．

提案手法　東北大学・乾研究室の日本語 BERT（cl-tohoku/bert-base-japanese-v2）[i]をベースに，sentence-transformers[ii]ライブラリを用い，BERT 出力

[i] https://github.com/cl-tohoku/bert-japanese/tree/v2.0
[ii] https://github.com/UKPLab/sentence-transformers

を平均プーリングする SBERT モデルを用意した．深層距離学習は AdaCos の損失関数を実装し，バッチサイズ 4，最適化関数 AdamW，学習率 5e-5，スケジューラ WarmupLinear，重み減衰係数 0.01，patience10 回の早期終了の条件のもと行った．

未学習 SBERT 提案手法と同じ日本語 BERT をベースとした SBERT を用いたが，比較のためファインチューニングを行わない状態とした．

TFIDF MeCab[16]（mecab-ipadic-neologd[17]辞書）を用いてテキストを形態素解析し，品詞フィルタ処理として「名詞_サ変接続」「名詞_ナイ形容詞語幹」「名詞_一般」「名詞_形容動詞語幹」「名詞_固有名詞」「動詞_自立」「形容詞_自立」のみを残し，原形化を行った単語群を用いて，埋め込み表現を求めた．

FastText TFIDF と同様の前処理を実施後，各単語を fastText[iii]ライブラリで単語ベクトル化し，TFIDF 値を重みとして加重平均して，埋め込み表現を求めた．fastText モデル自体は，2020/9/2 時点での日本語 Wikipedia ダンプに対して WikiExtractor[iv]で XML タグを除去したのち MeCab（mecab-ipadic-neologd 辞書）で分かち書きしたデータを，skip-gram 方式・300 次元で学習させたものを用いた．

USE TensorFlow Hub に存在する多言語モデル universal-sentence-encoder-multilingual-large(v3)[v]を使用して，埋め込み表現を求めた．

BERT 文書分類 提案手法と同じ日本語 BERT をベースに Huggingface Transformers[vi]ライブラリの文書分類クラスを用いて，バッチサイズ 32，最適化関数 Adam，学習率 5e-5，patience10 回の早期終了の条件のもと学習させたものを用いた．また分類手順としては，まずテストデータに対して各既知クラスに該当する確信度スコアを出力した．次に「確信度スコアが十分低いとみなす条件」を満たすものについては未知クラスを割り当て，それ以外は確信度スコアが最大のクラスへ割り当てるように推論した．なお，「確信度スコアが十分低いと見なす条件」として，「確信度スコアの最大値が閾値以下」「確信度スコアの最大値と次点の差分が閾値以下」の両方を実験したが，今回は精度の良かった前者を採用している．

4.4 評価方法

既知未知判定の単独精度 まず単純な既知/未知クラスの 2 値分類の識別能を評価した．その際，テストデータにおける既知/未知クラス比率は 1:1 に設定していること，また閾値選択によらない評価を行いたいことを鑑み，評価指標として AUROC（Area Under Receiver Operating Characteristic Curve）を採用した．

未知クラスを含む分類精度 次に未知クラスも含め全クラスを正しく推論できるかどうかを accuracy で評価した．なお本評価を行うためには，既知未知判定を確定させるため，閾値を選択する必要がある．本番適用する場合であれば，テストデータの正解ラベルは不明のため，閾値は平均 k 近傍距離もしくは確信度スコアの分布だけから決定すべきである．しかし今回はあくまで各モデルの潜在能力を評価することが目的のため，正解ラベルが所与として精度最大となる閾値を選択したうえで評価した[vii]．

5 実験結果

実験結果を表 2 に示す．ここから以下が分かる：

- 全ての実験条件・評価指標にて，本提案手法が最大精度を達成できている．
- 特に実験#2~4 は，実験#1 と比べて「1 クラス当たりのデータ量が少ない」「クラスごとのデータ量の不均衡性がある」「クラス数が多い」ケースとなっている．ビジネス利用を想定すると実際のデータでは不均衡であったり，データ量が十分になかったり，クラス数が多い場合があることを考慮すると，提案手法はいずれの場合も比較的高い精度を実現でき，実用性が高いといえる．
- 本提案手法と未学習 SBERT を比較すると，全ての実験条件・評価指標にて 20～30%近くの差分が出ている．ここから SBERT のモデル構造ではなく，深層距離学習自体による性能向上効果が非常に高いことが分かる．

[iii] https://github.com/facebookresearch/fastText
[iv] https://github.com/attardi/wikiextractor
[v] https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
[vi] https://github.com/huggingface/transformers

[vii] 提案手法の深層距離学習が十分高い精度で進めば，平均 k 近傍距離は既知/未知クラスで値が乖離するため，M 字型の分布をすることを別実験で確認している．このような場合であれば M 字の真ん中の極小点を閾値として選択できる．

- 他のユニバーサルな表現生成モデルと比較しても，本提案手法が優れており，特定のタスクに特化した表現の方が高い精度を実現できるという仮説通りといえる．ただ USE については比較的高い精度となっており，ある程度の性能で十分な場合には，ファインチューニング不要な USE も選択肢の一つと言える．
- 本提案手法と BERT 文書分類モデルを比較すると，1 クラス当たりの学習データ量が少なくなる場合（実験#2, #4）に BERT 文書分類モデルが大きく性能劣化するのに対して，本提案手法は性能劣化が少ないことがわかる．ここから同じ BERT ベースでも，深層距離学習により少量データに対する頑健性が高くなっていることがわかる．

6 おわりに

本研究では未知クラスを含む日本語テキストの分類問題に対して，深層距離学習で最適化した文の埋め込み表現を用いることで，一般に難しいとされるケースでも既存手法と比べて高い精度を達成でき，実用性が高いことを示した．

今後の改善としては，BERT 以外のベースモデルや他の損失関数への拡張が考えられる．また本提案手法は正解ラベルを必要とするが，自然言語処理では正解ラベルのアノテーションコストが高いという課題がある．そこで日本語テキストに対する data augmentation を適用することで，教師なしの対照学習へと拡張することも考えられる．

表 1 実験条件

			実験#1		実験#2		実験#3		実験#4	
定性	クラス数		中		中		中		大	
	学習データ量		中		小		中～大		小～大	
	学習データの不均衡性		なし		なし		中		大	
定量	クラス分類		既知	未知	既知	未知	既知	未知	既知	未知
	クラス数		12	2	12	2	12	2	80	7
	データ数		600	100	120	20	1138	107	2082	248
		訓練	400	0	80	0	825	0	1468	0
		検証	100	0	20	0	206	0	366	0
		テスト	100	100	20	20	107	107	248	248
	1 クラス当たりデータ数	最大	50	50	10	10	226	56	226	56
		最小	50	50	10	10	53	51	1	9

表 2 実験結果

タスク	評価指標	モデル	実験#1	実験#2	実験#3	実験#4
既知未知判定	AUROC	本提案手法	**0.89**	**0.81**	**0.85**	**0.77**
		未学習 SBERT	0.59	0.56	0.60	0.53
		TFIDF	0.79	0.71	0.77	0.67
		fastText	0.78	0.75	0.74	0.65
		USE	0.80	0.77	0.82	0.73
		BERT 文書分類	0.70	0.38	0.75	0.65
未知クラスを含む分類	accuracy	本提案手法	**0.73**	**0.70**	**0.71**	**0.63**
		未学習 SBERT	0.50	0.50	0.56	0.51
		TFIDF	0.62	0.55	0.58	0.50
		fastText	0.64	0.63	0.69	0.52
		USE	0.64	0.63	0.65	0.54
		BERT 文書分類	0.63	0.50	0.67	0.52

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo！知恵袋データ（第 2 版）」を利用した．

参考文献

[1] GENG, Chuanxing; HUANG, Sheng-jun; CHEN, Songcan. Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence, 2020.

[2] GUO, Chuan, et al. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR, 2017. p. 1321-1330.

[3] ZHANG, Xiao, et al. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 10823-10832.

[4] REIMERS, Nils; GUREVYCH, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. “Enriching Word Vectors with Subword Information” Transactions of the Association for Computational Linguistics (2017) 5: 135–146.

[7] S. Arora, Y. Liang, and T. Ma, “A simple but tough-to-beat baseline for sentence embeddings,” International Conference on Learning Representations (ICLR), 2017.

[8] K. Ethayarajh, “Unsupervised random walk sentence embeddings: A strong but simple baseline,” in Proceedings of The Third Workshop on Representation Learning for NLP, 2018, pp. 91–100

[9] Z. Yang, C. Zhu, and W. Chen, “Parameter-free sentence embedding via orthogonal basis,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 638–648.

[10] CONNEAU, Alexis, et al. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364, 2017.

[11] CER, Daniel, et al. Universal sentence encoder. arXiv preprint arXiv:1803.11175, 2018.

[12] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. arXiv preprint arXiv:1801.09414, 2018.

[13] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698, 2018.

[14] MASANA, Marc, et al. Metric learning for novelty and anomaly detection. arXiv preprint arXiv:1808.05492, 2018.

[15] ヤフー株式会社 (2011): Yahoo! 知恵袋データ（第 2 版）. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.1.2

[16] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)

[17] 佐藤敏紀; 橋本泰一; 奥村学. 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会発表論文集, 2017, 875-878.