E-006

再帰構造を用いた事前学習済み言語モデルの長文への対応
Enhancing Pre-trained Language Models for Long Texts with Recurrent Architecture

斎藤 大貴† 全 炳東†‡
Daiki Saito† Heitoh Zen†‡

1. はじめに

対話型言語モデルの普及に伴い，小説や論文，ソースコードといった特定ドメインに限らず，一貫性のある長文の生成が必要となっており，言語モデルを長文に対応させる手法の必要性が高まっている．

4096 トークン以上の入力を長文と定義したとき，長文に対応した言語モデルは既に幾つか提案されているが，提案論文で実装されるものは英語が基本であり，日本語などその他の言語で学習されたモデルを利用するには大規模データでの事前学習が必要となりコストが高い．

そこで，本研究では既存の事前学習済みモデルを利用することでコストを抑えながら長文に対応したモデルを作成する手法を提案する．既存の事前学習済みモデルとして BERT を利用し，Block Recurrent Transformers [1]を参考に再帰構造に拡張することで長文に対応する．

なお，評価実験の都合から英語で事前学習された BERT-tiny [2][3]を用いている．

2. 提案手法

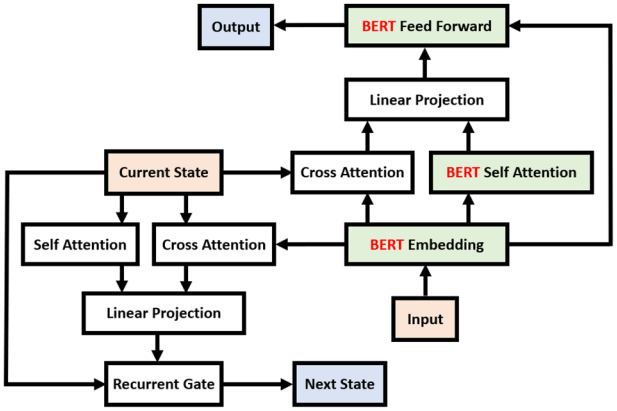

図 1　再帰構造に拡張した BERT

モデルの概略を図 1 に示す．バッチ正規化，ドロップアウトなど一部は省略した．

図の Input および Current State が入力，Output，Next State が出力を示している．BERT の名称を含む層が BERT の構造および事前学習済みパラメータを利用する部分であり，その他の層は Block Recurrent Transformers（BRT）と同様の構造である．

入力トークンを複数のブロックに分割し，このモデルを再帰的に用いることで，BERT に入力可能な最大入力系列長である 512 トークンを超える入力に対応できる（図 2）．

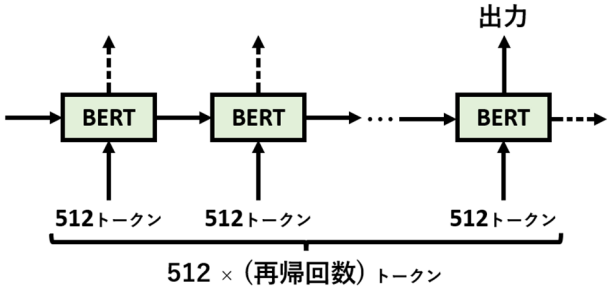

図 2　BERT を再帰的に用いる概略図

3. 実験

3.1 タスク

Arxiv Classification [4]および Hyperpartisan News Detection [5]によって評価する．前者は Arxiv 論文を 11 の分野に分類する多クラス分類であり，後者はニュース記事を入力とした True/False の 2 値分類である．

3.2 前処理

Arxiv Classification ではモデルが純粋に長文のテキストだけで分類できるかどうかを試すため，論文自体の要約である Abstract と論文の属性情報である題目と著者を削除した．Hyperpartisan News Detection では WEB の記事が含まれているため，HTML タグを削除した．

また，BERT のトークナイザを用いてトークン数の調査を行い，512 トークンに満たないデータは削除した．

3.3 設定

学習率を 1e-4，ミニバッチサイズを 4，出力層のドロップアウト率を 0.3，オプティマイザを AdamW，ミニバッチのサンプリングに用いる乱数のシードを 222 で固定して学習を行い，テストデータに対する正解率および F1 スコアで評価を行う．ベースラインとして再帰構造に拡張する前の BERT および事前学習なしの BERT を用いた場合を比較する．

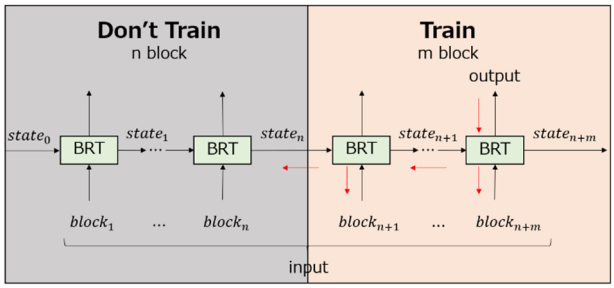

図 3　勾配計算する範囲を削減する例

† 千葉大学大学院 融合理工学府 数学情報科学専攻
Chiba University, Graduate School of Science and Engineering
‡ 千葉大学 情報戦略機構
Chiba University, Digital Transformation Enhancement Council

また，事前学習済みモデルを利用することで学習時間を減らす効果が期待されるが，更なる学習時間の削減のため，再帰構造に入力した 4096 トークンのうち，逆伝播によって勾配計算する範囲を減らすことが考えられる（図 3）．これについても比較を行う．

3.4 結果と考察

表 1 に結果を示す．BRT-BERT-tiny は提案手法であり，train-n/m-block は m ブロック入力し n ブロック学習した場合を意味している．

データ量の多い Hyperpartisan News Detection タスクでは提案手法がベースラインを上回ったが，データ量の少ない Arxiv Classification タスクでは提案手法がベースラインを下回った．ここから，十分なデータで学習した場合には提案手法によって既存モデルを長文に対応させることができると考えられる．十分なデータで学習する例として，wikipedia などの汎用的なデータを用いて追加の事前学習を行うことが考えられる．

また，学習時間削減のために勾配計算する範囲を減らした場合には，Hyperpartisan News Detection タスクにおける train 4/8 block において事前学習していない BERT を用いた場合と同等の精度であり，削減できる学習時間に対して精度の減少率が大きかった．一方，訓練時に入力する最大系列長を削減した train 4/4 block ではベースラインを上回り，train 8/8 block に近い精度を示した．train 2/2 block では大きく精度が減少していることから，学習時間を削減する場合には最大入力系列長を 4 ブロックに抑えると良いと考えられる．

分類結果を目視で確認するため，エンコーダ最終層の隠れ状態ベクトルを umap [6]を用いて 128 次元の空間を 2 次元に写像した散布図を図 5，図 6 に示す．比較的良好なクラスタリングができているが， Arxiv Classification タスクではクラスタによっては十分に分離できていないことが読み取れる．

表 1　実験結果

Model（Max Token）	Arxiv 29106 data		Hyperpartisan 300110 data	
	Accuracy	Macro F1	Accuracy	F1
BERT-tiny（512）	**79.17**	**78.67**	95.06	95.26
BRT-BERT-tiny-not-pretrained（4096）	76.60	76.07	95.69	95.81
BRT-BERT-tiny-train-8/8-block（4096）	78.24	77.97	**96.22**	**96.37**
BRT-BERT-tiny-train-4/8-block（4096）	78.30	77.38	95.68	95.82
BRT-BERT-tiny-train-2/8-block（4096）	76.29	75.85	95.10	95.21
BRT-BERT-tiny-train-4/4-block（4096）	78.52	77.54	96.15	96.29
BRT-BERT-tiny-train-2/2-block（4096）	70.20	69.22	94.28	94.53

4. おわりに

提案手法によって既存の事前学習済みモデルである BERT-tiny を再帰構造に拡張することで，Hyperpartisan News Detection タスクにおいて精度が向上することを示した．

今後の課題として，下流タスクにおいて十分なデータが得られない場合の精度を改善すること，デコーダ型モデルを用いた生成タスクでの精度を確認すること，日本語で事前学習されたモデルを用いた実験をすることがある．

1 つ目の課題については，MLM タスクによって追加の事前学習を行うことが考えられる．ただし，大規模データで学習することは学習時間を削減する目的から外れるため，テキストの長さで制限をかけ，データ数を削減することが考えられる．

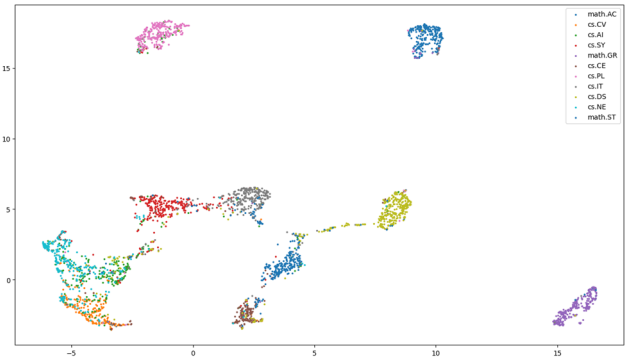

図 4　埋め込みの可視化（Arxiv）

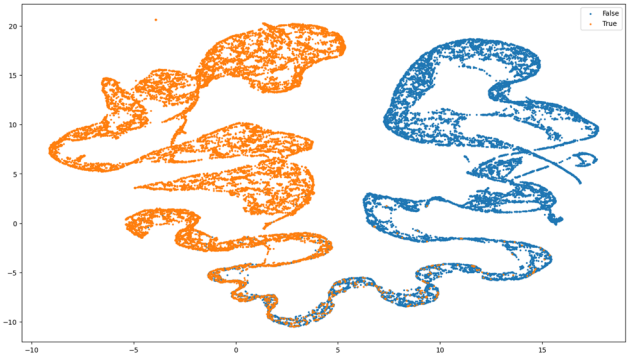

図 5　埋め込みの可視化（Hyperpartisan）

参考文献

[1] DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, Behnam Neyshabur. “Block-Recurrent Transformers”, arXiv:2203.07852 [cs], (2022).
[2] Prajjwal Bhargava, Aleksandr Drozd, Anna Rogers, “Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics”, arXiv:2110.01518 [cs], (2021).
[3] Iulia Turc, Ming-Wei Chang,Kenton Lee, Kristina Toutanova, “Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation”, CoRR, abs/1908.08962, (2019).
[4] Jun He, Liqun Wang, Liu Liu, Jiao Feng, Hao Wu, “Long Document Classification From Local Word Glimpses via Recurrent Attention Learning”, IEEE Access, vol 7, p 40707 – 40718, (2019).
[5] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, Martin Potthast, “Semeval-2019 task 4: Hyperpartisan news detection”, In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 829–839, (2019).
[6] Leland McInnes, John Healy, Nathaniel Saul, Lukas Großberger, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, Journal of Open Source Software, vol 3, 861-, (2018).