E-043

まぎらわしい分類タスクにおける言語モデルのファインチューニングとその解釈

中田 駿平 †　　正田 備也 ‡
立教大学大学院 †　　立教大学大学院 ‡

1 はじめに

近年，BERT や ChatGPT をはじめとする言語モデルに大きな注目が集まっている．言語モデルの優れた能力の 1 つが高い文脈把握能力である．文書を単なる単語の集まりではなく文脈をふまえたうえで扱う．これにより高度な文書埋め込みや文書生成を実現しているとされる．本研究では文書埋め込みにおける言語モデルの文脈把握能力について考察する．

扱うタスクは学術論文のメタデータを使用した文書分類問題である．論文のアブストラクトをもとに，その論文にラベルされた分野を特定する．論文は広義に微分幾何学に類別される近接分野のものを使用する．これらの論文は互いに出現単語・出現頻度が類似している．そのため Bag of Words（以下，BoW）による分類が困難な，まぎらわしいタスクであると予想される．本研究では BoW モデルと言語モデルの 2 種類のアプローチを採用する．両者の分類性能を比較することにより言語モデルの文脈把握能力の優位性を明らかにする．

2 実験概要

2.1 データセット

論文のメタデータは arXiv API Access[*1]から取得する．対象分野は math.AG, math.CV, math.DG, math.MG, math.SG[*2]の 5 分野とする．いずれも広義に微分幾何学に類別される近接分野である．このうち 2014 年 1 月 1 日から 2023 年 12 月 31 日までの 10 年間に出版されたもの対象とする．該当する論文件数はそれぞれ 19250 件, 6179 件， 15678 件， 3808 件， 2435 件で合計 47350 件である．実験ではこれらのデータを訓練用， 検証用， 試験用 = 8 : 1 : 1 の比率で層別に分割して使用する．

2.2 構築するモデル

構築するモデルは埋め込みを行う言語モデルと分類を行う全結合層からなる．言語モデルは HuggingFace の MTEB リーダーボード[*3]の中から成績上位のものから選択する．本研究では bert-large-cased と同じパラメータサイズが 335M 規模のモデルを 7 つ選択した．なお，ベースラインとなる BoW モデルは埋め込みには TF-IDF を，分類には線形サポートベクトルマシンを使用する．

Params	Value
max token length	256
train batch size	32
gradient accumulation	16
optimizer	AdamW

表 1　全モデル共通のハイパーパラメータ．Params 列はハイパーパラメータの種類，Value 列はその設定値を表す．

2.3 ファインチューニング

学習には NVIDIA GeForce RTX 4090（VRAM 24GB）を使用する．学習は全パラメータを対象とするフルファインチューニングで行う．学習時間は検証ロスを監視して patience=3 で早期終了させる．結果としてどのモデルも 40 分程度で学習が完了した．

全モデル共通のハイパーパラメータは表 1 の通りである．max token length と train batch size は使用する GPU のメモリサイズに合わせて可能な限り大きく設定した．max token length については全文書中約 8 割が 256 トークン以下の長さである．学習率は検証ロスが最小になるようモデルごとに最適な値を設定した．

2.4 モデルの評価

ファインチューニングの結果，各モデルの分類性能は表 2 の通りとなった．どの言語モデルも BoW モデルより高い性能を発揮している．最も性能が高かったのは bge-large-en-v1.5 であり正解率と F1 スコアはそれぞれ 91.59%, 88.77% であった．しかし，いずれの言語モデルも性能の差はわずかであった．このことからパラメータサイズが 335M 規模の言語モデルであれば，どのモデルを選択しても概ね同等の性能が得られることがわかる．一方，BoW モデルも比較的高い性能を発揮している．正解率と F1 スコアはそれぞれ 90.54%, 87.20% であり，言語モデルとの差は 1.05%, 1.57% であった．本実験は言語モデルの文脈把握能力の優位性を明らかにする狙いがあったが，大きな差とまではいえない結果となった．

3 積分勾配法による解釈

3.1 積分勾配法とは

最も高性能であった bge-large-en-v1.5 を対象に積分勾配法による解釈を試みる．積分勾配法はニューラルネッ

Fine-Tuning Language Models and Their Interpretation in Ambiguous Classification Tasks
† Shumpei Nakada, Rikkyo University
‡ Tomonari Masada, Rikkyo University
*1 https://info.arxiv.org/help/api/index.html
*2 https://www.arxiv.org/archive/math
*3 https://huggingface.co/spaces/mteb/leaderboard

Model	Acc.	F1
bert-large-cased	90.90	87.70
e5-large-v2	91.15	88.60
bge-large-en-v1.5	**91.59**	**88.77**
gte-large	91.00	87.86
mxbai-embed-large-v1	90.60	87.37
GIST-large-Embedding-v0	91.53	88.76
UAE-Large-V1	90.83	87.75
TF-IDF*LinearSVM (BoW)	90.54	87.20

表 2　各モデルの分類性能．Model 列は埋め込み用の言語モデル，Acc. 列はマルチラベル正解率，F1 はマクロ F1 スコアを表す．いずれの値も単位はパーセントである．太字は最も性能が高かった値である．

トワークモデルと相性のよい解釈性手法である．寄与度の加法性など解釈性手法として望ましい性質を備えており，近年よく用いられる．積分勾配法により計算される寄与度は，モデルが問題を解く際にどのトークンが重要な役割を果たしているかの目安であると解釈できる [1]．なお値を計算する際は積分ステップ数をサンプルごとにチューニングすることが重要である [2]．本研究では積分ステップ数を 1 から 200 まで変化させ数値誤差指標が最小となる結果を採用した．

3.2　サンプル文書の解釈

サンプル文書は実際の論文のアブストラクトをベースに ChatGPT にパラフレーズさせて生成する．これらの文書に対する積分勾配法の結果が図 1 である．

上段の文書では「metric measure spaces」が正の寄与を示している．測度距離空間は math.MG における主要な研究テーマなので納得感がある．特に「spaces」は他分野でも出現しうるありふれた単語である．「metric measure」と共起して正の寄与を示していることは，モデルが一連のトークンの意味を文脈に基づき判別している証拠であると考えることもできる．中段の文書では「for sequences of metric measure spaces with unbounded dimensions」が全体的に正の寄与を示している．（次元が非有界な）測度距離空間列は math.MG における主要な研究テーマなので納得感がある．一方で「convergence theory」は負の寄与を示している．「convergence」と「sequences」は密接に関係するトークンであり，モデルの文脈理解という観点では直感に反する．下段の文書では「einstein metric」が正の寄与を示している．「metric」は math.MG においても頻出する単語だが，「einstein」と結合することで math.DG 固有の特徴となっていると考えられる．一方で「ka ##hler」や「fan ##o」といった単語（人名）も分野を特定する上で重要なキーワードとなるが，あまり高い寄与を示していない．サブワード分割された各トークンにあてられた寄与度も一貫性がなく，そもそも一つの単語として認識できていない印象である．

そのほかの点として「this」「of」「examine」といったありふれた単語が正負に大きな寄与をしている場合がある．これらはどの分野にも出現しうる単語であり直感に反する．また複数トークンにわたり大きな寄与をしていても，単に寄与度の高いトークンが連続して出現しているだけの可能性もある．したがってこの事実をもってモデルの文脈把握能力の証拠とするのは不十分であり追加の分析が必要である．

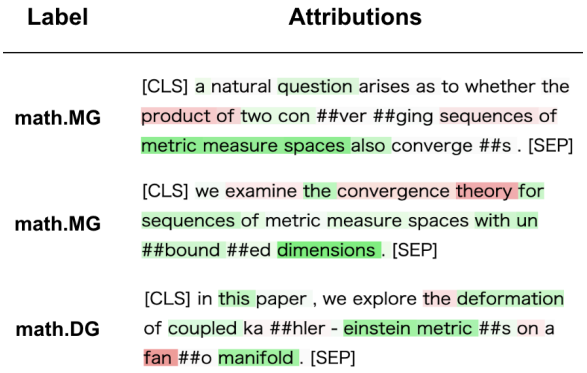

図 1　サンプル文書における各のトークンの寄与度．Label 列は正解ラベル，Attributions 列は各トークンの寄与度を可視化したものである．トークンにあてられた色は緑なら正の寄与を，赤なら負の寄与を表し，色が濃いほど大きな寄与であることを示す．

4　おわりに

本研究ではパラメータ数が 335M 規模の言語モデルを対象にその文脈把握能力について考察した．近接分野の学術論文の分類問題を扱ったため難易度の高いタスクであることが予想されたが，言語モデルをファインチューニングすることで高性能な分類モデルを構築できた．一方で BoW モデルでも言語モデルに匹敵する性能を発揮した．そのため言語モデルの文脈把握能力について大きな優位性は示すには至らなかった．また，積分勾配法による言語モデルの解釈も試みた．分析結果に一定の解釈性は認められたが，単に重要単語が連続して出現した可能性も排除できず，言語モデルの文脈把握能力を明らかにする証拠とまでは至らなかった．

今後は浅いニューラルネットモデルを積分勾配法によって解釈することを試みる．本研究で構築したモデルと比較することにより，Transformer ベースの言語モデルが持つ高度な文脈把握能力についてより明確な知見を得ることを目指す．

参考文献

[1] M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks., ICML'17 - Vol. 70, pp. 3319-3328, 2017.
[2] 牧野雅紘，浅妻祐弥，佐々木翔大，鈴木潤，Integrated Gradients における理想の積分ステップ数はインスタンス毎に異なる，言語処理学会 第 30 回年次大会 発表論文集, pp. 1826-1830, 2024.