単一トークン適応による大規模言語モデルに基づく文埋め込み

趙 開顔　呉 奇宇　苗 中濤　呉 梓隆　鶴岡 慶雅
東京大学大学院 情報理工学系研究科
{zhaokaiyan1006, qiyuw, mzt, zw2599, yoshimasa-tsuruoka}@g.ecc.u-tokyo.ac.jp

概要

文埋め込み学習では、対照学習に基づいたエンコーダのみのモデルが広く使われている。一方、大規模言語モデル (LLM) は様々なタスクで有効性が示されているにもかかわらず、LLM を文埋め込みの生成に利用する方法はまだ確立していない。本研究では、LLM の学習済みの知識や生成能力を維持しつつ、文全体の情報を捉えることができ、唯一の更新可能な特殊トークンである < *t2e* > を導入する新しいアプローチを提案する。文類似度タスクの実験結果から、単一の特殊トークン < *t2e* > を更新することで、本手法は他の微調整されたモデルと近い結果を達成できることが示された。

1　はじめに

文の埋め込みは自然言語処理 (NLP) の基本的なタスクであり、テキストや文を実数値ベクトルに変換する処理である。ベクトル空間上で類似した意味を持つ文が近くなるように学習し、文全体の意味を表すことができる。これにより、文や文書の検索 [1]、自然言語推論 [2]、および意味マッチングなど [3]、様々な下流タスクが容易になる。

文埋め込み学習の現在の一般的な手法は、Sentence-BERT [4] や SimCSE [3] のような、対照学習（CL）に基づくエンコーダのみのモデルを活用することである。文埋め込みタスクにおける対照学習の目的は、意味的に類似した文を近づけつつ、他の意味的に類似していない文を遠ざけることである [3, 5]。双方向のアテンション (bidirectional attention) 計算を通じて、エンコーダのみのモデルは、`[cls]` トークンや最後の層の平均出力を出力するなどの方法を用いて、文全体の情報を表現することができる [4, 6]。

一方で、大量のパラメータと豊富なトレーニングデータによって、大規模言語モデル (LLM) の強みは様々な下流タスクで示されている [7, 8, 9, 10]。

Methods	Representation	Generation	Computation
Prompt-based ICL	*poor*	*strong*	*low*
CL-based SFT	*strong*	*poor*	*high*
Attention manipulation	*strong*	*poor*	*high*
< *t2e* > **(ours)**	*strong*	*strong*	*low*

表 1　大規模言語モデルで文埋め込みを生成する異なる方法の特徴。

しかしながら、LLM を文埋め込みを生成するために適応させる方法はまだ確立していない。これは主に以下の理由が挙げられる：(1) LLM は双方向アテンションではなく、コーザルアテンション (Causal Attention) を使用し、連続的なテキストの生成を目的としている [11]。(2) 非自己回帰モデルのような `[cls]` トークンが存在せず、全体の文を表現することが困難である。

既存の LLM を使用して文埋め込みを生成する手法は、主に以下の 3 種類に分けることができる：(1) プロンプトベースのインコンテキストラーニング (In-Contect-Learning, ICL [12]) [11]、(2) CL ベースの教師あり微調整（SFT) [11, 13]、および (3) アテンションメカニズム (Attention Mechanism) の調整 [14]。しかし、表 1 に示すように、これらの方法はいずれも、強い表現能力、生成能力への影響、および低い計算量を同時に達成することができない。表 1 中の表現能力 (Representation) は、埋め込みが文全体の豊かな意味を表現できたかどうかを示している。生成能力 (Generation) は、LLM を文埋め込みを適応させることが LLM の学習済みの生成能力に影響を与えるかどうかを示しいる。低い計算量 (Computation) は、モデルをトレーニングするために必要な計算資源を指している。

プロンプトベースの ICL [11] では、*one word limitation* を使用し、文を 1 つの単語に要約し、その単語の表現を文の埋め込みとして扱う。しかし、ただ一つの単語だけでは文を十分に表現するのは不十分だと、我々は考えている。[11] のデモンストレーション例 (demonstration examples) では、複数

の文が同じ単語に要約される事例が存在している。例えば、「A man is playing a guitar.」と「The woman is playing the flute.」では、意味の違うこれらの文が同じ単語「Music」に要約されている。生成能力への影響と計算資源に関しては、SFT とアテンションの操作の両方がほとんどのパラメータを更新し、十分に学習された知識や生成能力に影響を与えたと同時に、膨大な時間と計算資源を占有する [15, 16]。

このため、私たちは新しい効率的なアプローチとして特殊トークン $< t2e >$ を用いる手法を提案する。これは、入力文の後に追加されるプラグイン (plug-in) 特殊トークン $< t2e >$ で、LLM にテキストを埋め込み（text-to-embedding）に変換するように学習させる。トレーニング中、$< t2e >$ トークンのパラメータだけが更新され、他の LLM のパラメータはすべて凍結されている。

文類似度タスク STS [17] の実験結果からは、LLM 内のただ一つのトークン $< t2e >$ だけを更新することで、文全体の意味を捉えることができることを示している。LLM 内のすべてのパラメータを更新する他の方法と比較して、$< t2e >$ は計算資源のコストを極端に削減しつつ、すべて学習済みの生成能力を保持したまま、比較的に同等の結果を達成できることを示している。

2 関連研究

2.1 文埋め込み

文埋め込みは、文を特定の埋め込み空間内の固定サイズのベクトルに変換する技術である。単語埋め込み [18, 19] と同様に、ベクトル空間上で類似した意味を持つ文が近くなるように学習し、文全体の意味を表すことができる。ニューラルネットワークが広く用いられる前に、Doc2Vec [20] などの手法は文を含む段落の意味を捉えるため導入された。

近年では、対照学習 (CL) の能力を活用した SimCSE [3] の成功により、文埋め込みの処理において CL に基づく手法が注目を集めている [1, 21, 22]。意味的に類似した文を近づけつつ、他の意味的に類似していない文を遠ざけることにより、S-BERT（Sentence-BERT) [4] や Universal Sentence Encoder（USE) [23] などの従来なモデルと比較し、CL に基づくモデルは文から豊富な意味情報をより良く捉えることが報告されている。

2.2 大規模言語モデル

大規模言語モデル (LLMs) は、多数のパラメータを持つニューラルネットワークによって構築され、膨大な量のテキストデータで学習されたモデルである [7, 8, 9, 10]。LLMs は、テキスト補完 [12]、要約 [24]、質問応答 [25] など、さまざまな自然言語理解および生成タスクで優れた能力を示している。事前にトレーニングされた性質により、特定の下流タスクでの微調整が可能であり、様々なアプリケーションに適用できる汎用性が存在している。

しかしながら、LLM を文埋め込みを生成するために適応させる方法はまだ確立していないため、本研究では、唯一更新可能な特殊トークンである $< t2e >$ ををを導入する効率的なアプローチを提案する。

3 提案手法

このセクションでは、最初に CL を用いた文埋め込みの形式的な定義を述べ、その後に $< t2e >$ の導入と学習方法を紹介する。

3.1 背景

与えられた文 $x_i \sim \mathbb{X}$ に対して、文埋め込み学習はパラメータ化されたネットワーク f_θ を学習することを目指している。ネットワークは x_i に適用され、ベクトル、すなわち $\mathbf{h}_i = f_\theta(x_i) \in \mathbb{R}^d$ に変換し、文 x_i の意味を表現できる。対照学習のアイデアは、文 x_i に対して近い意味を持つ正例文 x_i^+ を構築し、これらを近づける一方で、x_i を他の意味的に関係ない負例文から遠ざけることである。一般的に使用される学習の目的関数は、以下の対照損失を最小化することである [3, 5]:

$$l_i^{(s)} = -\log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{h}_i, \mathbf{h}_j)/\tau}}, \tag{1}$$

ここで、$\mathbf{h}i^+ = f\theta(x_i^+)$、$sim(\cdot,\cdot)$ は類似性メトリック、N はミニバッチのサイズ、τ は温度パラメータ (temperature parameter) である。学習後、文脈情報が含まれている埋め込みの使用により、文の検索、文レベルの分類、およびテキストの類似度計算など、様々な下流タスクで応用できる。

3.2 $< t2e >$

$< t2e >$ は text-to-embedding（テキストから埋め込みへ）を示している。これは、LLM に追加で挿入し

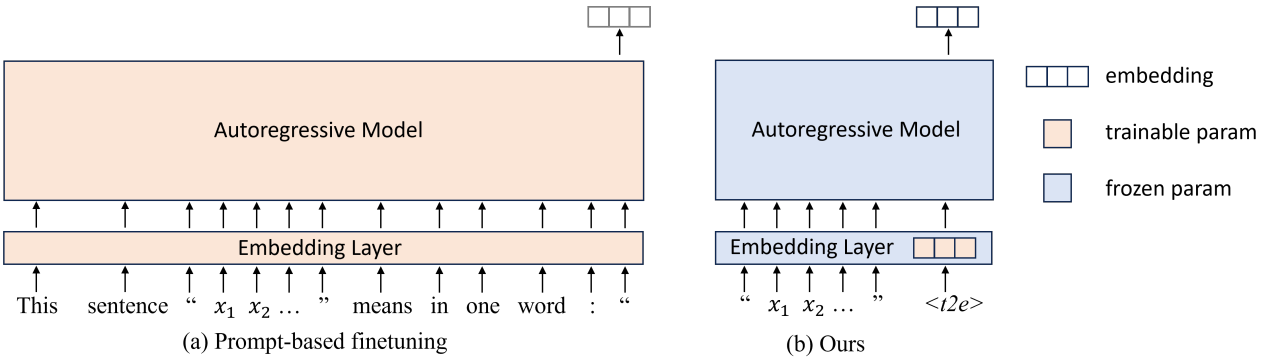

図 1 提案手法と既存のモデルの比較。既存の手法はモデル内のほとんどのパラメータを更新する一方で、提案手法は単一特殊なトークンのみを更新して文埋め込みを生成する。

た特別なトークンであり、事前学習された生成能力の忘却を同時に防ぎつつ、低コストで文レベルの情報を獲得することができる。

文 x_i をモデルに入力する前に、$< t2e >$ がまず x_i の後ろに連結され、$[x_i, < t2e >]$ の入力を構築する（図 1 (b) を参照）。この特別なトークンを使用することにより、LLM にこの文の表現を生成させる。$< t2e >$ の最後の層の出力が文埋め込みとして利用される。モデルを訓練する際に使用した目的関数は式 1 に示した対照学習損失である。

図 1 (a) に示すように、モデル内のほとんどのパラメータを更新する方法とは異なり、学習中には $< t2e >$ の重みのみが更新され、他のすべてのパラメータは凍結されている。したがって、トレーニングプロセスは他の既存の方法と比較して非常に低コストで済みことができる。

4 実験

$< t2e >$ の汎用性を証明するため、OPT-125m [1] [16] を含む小規模な自己回帰モデルと LLaMA2-7b [2] [8] を含む LLM の両方で実験を行った。

使用したデータセットは、アンカー文、正例文、ハード負例文 (hard negative) を含む教師付きの SimCSE の NLI データセット [3] [3] である。

ベースサイズのモデルは、バッチサイズ 128 で 1 つの A100 80g GPU 上で学習し、学習率と重み減衰 (weight decay) は 0.01 に設定されている。一方、7b サイズの大規模モデルは、各 GPU 上でバッチサイズが 32 である 2 つの A100 80g GPU でトレーニングし、学習率と重み減衰は同じく 0.01 である。

モデルの性能を、SentEval Benchmark [4] [26] からの 7 つの STS タスクで評価し、いくつかのベースラインと比較した。まず、SBERT [4] と SimCSE [3] を最も一般的に使用される 2 つのベースラインとして選び、それに加えて、ICL と完全に教師付きで微調整された OPT および LLaMA を適用した PromptEOL [11] を選択する。最後に、CL 損失のみで学習された OPT および LLaMA の比較を行った。

実験結果は表 2 に示している。

† が付いてる平均スコア結果から見ると、学習なしの小規模および大規模モデルは文の完全な意味を捉えることができないことが分かる。報告されたベースサイズモデルと 7B サイズモデルの両方において、スコアの最高値は完全な微調整によって達成されている。また、ベースサイズモデルに $< t2e >$ の導入により、モデルは SimCSE よりも性能が少々下回っているが、SBERT および PromptEOL に基づく ICL よりも優れている。 7B サイズのモデルに対しては、性能が大量のパラメータを更新するモデルにかなわいだが、学習なしの LLaMA2 より、大幅に効果が向上している。

我々の結果が SOTA（State of the Art）モデルを超えていないにもかかわらず、トレーニングなしのモデルと比較して大幅な改善をもたらすことができる。また、我々モデルが必要とする計算資源が限られていることを考慮すると、$< t2e >$ は効率的かつ効果的な手法であると言える。

我々の $< t2e >$ と既存方法の計算資源をより良く比較するため、表 3 に全ての訓練可能なパラメータをリストアップした。我々は PEFT ライブラリ [5]. の実装を使用し、訓練可能なパラメータ数を数えた。

1) https://huggingface.co/facebook/opt-125m
2) https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
3) https://github.com/princeton-nlp/SimCSE/blob/main/data/download_nli.sh
4) https://github.com/facebookresearch/SentEval
5) https://huggingface.co/docs/peft/package_reference/peft_model

Model	STS12	STS13	STS14	STS15	STS16	STS-B	STS-R	avg.
base size models (sup.)								
SBERT [4]	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE [3]	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
OPT †	7.47	9.48	8.30	19.63	22.45	7.40	24.91	14.23
OPT + PromptEOL (ICL) [11]	62.22	73.10	61.84	71.09	72.08	67.80	64.10	67.46
OPT + CL + < *t2e* > **(ours)**	69.87	79.55	73.92	83.91	79.06	80.90	74.17	77.34
7B size models (sup.)								
LLaMA2 †	22.30	30.92	27.10	38.92	52.95	33.66	42.54	35.48
LLaMA2 + CL [13]	78.39	89.95	84.80	88.50	86.04	87.86	81.11	85.24
LLaMA2 + CL + PromptEOL [11]	79.16	90.22	85.40	88.99	86.25	88.37	81.51	85.70
LLaMA2 + CL + < *t2e* > **(ours)**	72.45	86.00	80.57	84.60	83.40	85.96	80.63	81.94

表 2　7 つの STS タスクの結果。すべてのタスクに対して spearman correlation を報告した。† が付いているモデルは、微調整なしのモデルで、最後のトークンからの出力を文埋め込みとして使用していることを示している。ICT は、in-context-learning の略で、+CL は、モデルが CL 目標の下でトレーニングされていることを示している。

Model	Trainable Param	Total Param	Percentage
OPT-125M + SFT	125239296	125239296	100%
OPT-125M + < *t2e* >	768	125240064	0.00061%
LLaMA2 + LoRA	159907840	6898319360	2.32%
LLaMA2 + < *t2e* >	4096	6738415616	0.000061%

表 3　< *t2e* > と既存手法の訓練可能なパラメータの比較。

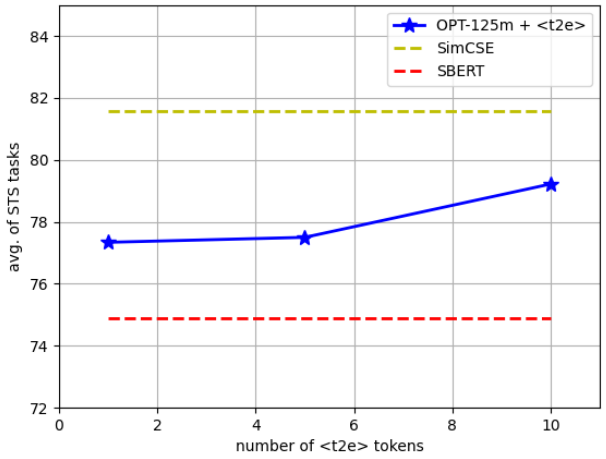

図 2　< *t2e* > トークンの数が STS タスクに及ぼす影響。

我々の手法は追加のトークンがただ 1 つだけなので、訓練可能なパラメータは常にモデルの隠れ状態 (hidden state) のサイズに等しいである。これらのパラメータは、それぞれ OPT-125M と LLaMA2-7B のすべてのパラメータの約 0.000061%および約 0.00061%しか占めていない。LoRA [27] を利用する LLaMA2+CL の場合、我々は [11] での記述に従って訓練可能なパラメータを計算し、その割合は約 2.32%である。したがって、我々の手法はパラメータの数を約 38,000 倍節約している。

実験結果部分では、我々は一つの < *t2e* > プラグイントークン (plug-in) を使用した結果を報告したが、訓練可能なトークンの数が STS タスクに与える影響を調査するため、より多くの < *t2e* > トークンを使用した実験も行った。その結果は図 2 に示されている。

訓練可能な < *t2e* > トークンの数が増えると、STS タスクの平均性能も向上する傾向が明確である。これは、より多くの訓練可能なパラメータを持つモデルがより多くの情報を記憶し、文を埋め込みにより良く変換できる可能性があるためである。

5　おわりに

本研究では、大規模言語モデルで文埋め込みを生成させる、単一学習可能な特殊トークン < *t2e* > を導入する新しい効率的な手法を提案した。この手法は、LLM を文埋め込みを生成するために適応させるもので、他の既存の手法とは異なり、非常に低いコストで LLM の事前学習済みの生成能力に影響を与えずに文脈情報を捉えることができる。文類似度タスク上の実験結果は、我々の手法がほとんどのパラメータを更新する手法と近い結果を達成できることが示された。

参考文献

[1] Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. Leveraging multi-lingual positive instances in contrastive learning to improve sentence embedding, 2023.

[2] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In **Conference on Empirical Methods in Natural Language Processing**, 2018.

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In **Conference on Empirical Methods in Natural Language Processing**, 2021.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Conference on Empirical Methods in Natural Language Processing**, 2019.

[5] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. **ArXiv**, Vol. abs/1807.03748, , 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **North American Chapter of the Association for Computational Linguistics**, 2019.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. **ArXiv**, Vol. abs/2302.13971, , 2023.

[8] Llama 2: Open foundation and fine-tuned chat models, 2023.

[9] OpenAI. Introducing chatgpt, 2022. https://openai.com/blog/chatgpt.

[10] Gpt-4 technical report, 2023.

[11] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models, 2023.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. **ArXiv**, Vol. abs/2005.14165, , 2020.

[13] Xianming Li and Jing Li. Angle-optimized text embeddings, 2023.

[14] Xianming Li and Jing Li. Deelm: Dependency-enhanced large language model for sentence embeddings, 2023.

[15] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yuechen Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. **ArXiv**, Vol. abs/2308.08747, , 2023.

[16] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. **ArXiv**, Vol. abs/2205.01068, , 2022.

[17] Daniel Matthew Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **International Workshop on Semantic Evaluation**, 2017.

[18] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In **North American Chapter of the Association for Computational Linguistics**, 2013.

[19] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **International Conference on Learning Representations**, 2013.

[20] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, **Proceedings of the 31st International Conference on Machine Learning**, Vol. 32 of **Proceedings of Machine Learning Research**, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[21] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[22] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

[23] Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for english. In **Conference on Empirical Methods in Natural Language Processing**, 2018.

[24] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. Benchmarking large language models for news summarization. **ArXiv**, Vol. abs/2301.13848, , 2023.

[25] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. **ArXiv**, Vol. abs/2306.13304, , 2023.

[26] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. **arXiv preprint arXiv:1803.05449**, 2018.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.