When does Continuous Learning for BERT make sense?

Zachary Yang[†,*]
[†] McGill University

Keywords: Natural Language Processing, Continuous Learning, AI Applications

1. Background

Natural Language Processing (NLP) is a well-established field that has been researched since the 1940s and encompasses many tasks (e.g. Named Entity Recognition (NER), text summarization and question-answering) [1]. The introduction of attention and transformer architecture by [2] revolutionized NLP and further inspired the development of popular pre-trained language models (PLMs) such as BERT [3], RoBERTa [4] and GPT-3 [5]. Pretrained language models (PLMs) gained popularity from its versatility where huge corpora of unlabeled text can be used to pre-train the models, i.e. train the models to predict masked words in a self-supervised manner. These models can then be fine-tuned, i.e. training the model on a specific dataset, for any downstream NLP tasks. PLMs are favored over traditional methods as they significantly perform better with reduced preprocessing effort.

However, some researchers have raised concerns about the out-of-vocabulary issue, where 1) the language used in pre-training data may differ from the specific domain, and 2) the model will lack the latest language trends or knowledge not found in the pre-training data. Continuous learning for PLMs is a new field aimed at addressing these concerns. While most researchers agree that pre-training models in specific languages is necessary in the settings of different languages (French, Chinese, Indian, Persian, etc)[1] or multilingual case (Bert-Multilingual [3], XLM-RoBERTa [6], T5 [7]), there is debate over whether further pre-training on domain-specific data is required. Some fields such as legal [8] and medical [9], have shown improved performance with further pre-training on domain-specific data.

2. Research Objectives

The primary objective of this research is to establish a comprehensive framework for NLP professionals to address crucial questions related to continuous learning for PLMs, including but not limited to:

(1) The extent to which PLMs can benefit from continuous learning in terms of pre-training and fine-tuning, as well as whether pre-training from scratch is the better choice
(2) The prerequisites for each method, including the methods for determining distribution shift, amount of data needed and amount of specialized knowledge needed.
(3) The trade-offs that exist among different methods, including the benefits and drawbacks of each approach in terms of performance, time cost, and other relevant factors.

3. Methodology

The research project is divided into two main parts:

[1] https://huggingface.co/models?sort=downloads

[*] zachary.yang@mail.mcgill.ca

(1) **Defining Distribution Shift**

The concept of continuous learning has been extensively studied in the domains of reinforcement learning and traditional machine learning models. Typically, distribution shift is determined by evaluating the loss on a newer dataset with known labels, and when this loss exceeds a predefined threshold, practitioners must decide whether to continue training the existing model or train a new model from scratch. However, in the case of with PLMs, the model is complicated by its two steps, the pre-training and the fine-tuning. Determining the distribution shift between different domains of the same language has traditionally been done based on intuition alone is subjective and not a reliable indicator. Therefore, a more concrete method is needed to measure the distribution shift.

Currently, two potential avenues for measuring distribution shift are being explored. The first avenue involves using the fact that PLMs are trained to predict masked tokens, and therefore, evaluating the loss of predicting the masked words on a specific domain dataset can serve as an indicator of the distribution shift. The second avenue involves using the trained tokenizer of the PLMs. For instance, in the case of the WordPiece Tokenizer [10] used in BERT, the number of new tokens found in the specific domain dataset could serve as an indicator. In contrast, for the BytePair Encoding Tokenizer [11] used in RoBERTa and GPT-3, the distribution of these tokens found in the pre-trained corpora and the specific domain dataset could serve as an indicator. However, it remains to be determined what the appropriate thresholds for these indicators are for different domains, such as social media (e.g., tweets, Reddit posts), news, legal, and biology. Furthermore, there is a need to study whether there are any consistency amongst different domains that can be found. Since PLMs are ultimately fine-tuned on the specific dataset, it is also important to determine whether these indicators are sufficient to differentiate between continuously pre-training and continuously fine-tuning the models.

(2) **Continuous Learning Methods**

There are currently numerous proposed methods for continuous pretraining, but conflicting advice exists regarding the amount of data required for this process. Some sources argue that a minimum of one gigabyte of data is necessary, while others suggest that pretraining on the downstream dataset alone may suffice [12]. This leads to several important questions: How much data from the specific domain is required? Should pre-training be limited to the specific domain, or should some combination of the original corpora and the specific domain be used? Another important consideration is the tokenizer. Should the original tokenizer be used, or should a new tokenizer be trained? Alternatively, can the original tokenizer be enhanced with domain-specific vocabulary [2], or should BERT be improved through methods such as ExBert[13]? When adding domain-specific vocabulary, at what point does the cost become too high relative to simply training a new tokenizer? For all of these methods, what are the trade-offs between performance gain, time required for both pre-training and fine-tuning, and changes in inference time? Finally, experiments involving only continuous finetuning are essential to determine whether continuous pre-training is necessary / worth at all.

[2] https://medium.com/@pierre_guillou/nlp-how-to-add-a-domain-specific-vocabulary-new-tokens-to-\a-subword-tokenizer-already-trained-33ab15613a41

Acknowledgements

We thank Dr. Reihaneh Rabbany (McGill University) and the Complex Data Lab for providing support, assistance and supervision[3].

References

[1] A. Roy. *Recent Trends in Named Entity Recognition (NER)*. 2021. DOI: 10.48550/ARXIV.2101.11420. URL: https://arxiv.org/abs/2101.11420.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. DOI: 10.48550/ARXIV.1907.11692. URL: https://arxiv.org/abs/1907.11692.

[5] T. B. Brown et al. *Language Models are Few-Shot Learners*. 2020. DOI: 10.48550/ARXIV.2005.14165. URL: https://arxiv.org/abs/2005.14165.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116 (2019). arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

[7] H. W. Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. DOI: 10.48550/ARXIV.2210.11416. URL: https://arxiv.org/abs/2210.11416.

[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. *LEGAL-BERT: The Muppets straight out of Law School*. 2020. DOI: 10.48550/ARXIV.2010.02559. URL: https://arxiv.org/abs/2010.02559.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2019). Ed. by J. Wren, pp. 1234–1240. DOI: 10.1093/bioinformatics/btz682. URL: https://doi.org/10.1093\%2Fbioinformatics\%2Fbtz682.

[10] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou. *Fast WordPiece Tokenization*. 2020. DOI: 10.48550/ARXIV.2012.15524. URL: https://arxiv.org/abs/2012.15524.

[11] R. Sennrich, B. Haddow, and A. Birch. *Neural Machine Translation of Rare Words with Subword Units*. 2015. DOI: 10.48550/ARXIV.1508.07909. URL: https://arxiv.org/abs/1508.07909.

[12] K. Krishna, S. Garg, J. P. Bigham, and Z. C. Lipton. *Downstream Datasets Make Surprisingly Good Pretraining Corpora*. 2022. DOI: 10.48550/ARXIV.2209.14389. URL: https://arxiv.org/abs/2209.14389.

[13] W. Tai, H. T. Kung, X. Dong, M. Comiter, and C.-F. Kuo. "exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1433–1439. DOI: 10.18653/v1/2020.findings-emnlp.129. URL: https://aclanthology.org/2020.findings-emnlp.129.

[3] As the supervisor of the author, Zachary Yang, I am aware of the acceptance of the paper titled When does Continuous Learning for BERT make sense? in the GSS track of 2023 Canadian AI, and approve that it can be published on the PubPub open access online publication platform