ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Rethink the Effectiveness of Text Data Augmentation: An Empirical Analysis

Zhengxiang Shi and Aldo Lipani

University College London
Gower St, London - United Kingdom

Abstract. In recent years, language models (LMs) have made remarkable progress in advancing the field of natural language processing (NLP). However, the impact of data augmentation (DA) techniques on the fine-tuning (FT) performance of these LMs has been a topic of ongoing debate. In this study, we evaluate the effectiveness of three different FT methods in conjugation with back-translation across an array of 7 diverse NLP tasks, including classification and regression types, covering single-sentence and sentence-pair tasks. Contrary to prior assumptions that DA does not contribute to the enhancement of LMs' FT performance, our findings reveal that continued pre-training on augmented data can effectively improve the FT performance of the downstream tasks. In the most favourable case, continued pre-training improves the performance of FT by more than 10% in the few-shot learning setting. Our finding highlights the potential of DA as a powerful tool for bolstering LMs' performance.[1]

1 Introduction

In recent years, the development of LMs has revolutionized the field of NLP [1, 2], leading to remarkable progress in a range of downstream tasks, including text classification [3, 4], information retrieval [5, 6, 7, 8], and multi-modalities [9, 10, 11]. While LMs have shown impressive performance in many tasks, there has been a debate over the effectiveness of simple data augmentation (DA) techniques, such as back-translation, for improving the FT performance.

Previous research [12] evaluated DA techniques, such as Back-Translation, suggesting that these prevalent task-agnostic DA yields limited and inconsistent improvements for pre-trained LMs [13] in many basic classification tasks. Additionally, [14] contended that most previous augmentation methods offer only marginal gains and are generally ineffective, pointing out that DA often leads to unstable performance and can trigger a failure mode, characterized by severe performance drops or fluctuations.

In this study, we provide an empirical study to re-evaluate the effectiveness of text DA with two state-of-the-art prompt-based FT approaches [15, 16], as well as the conventional `CLS`-based FT [13], as shown in Figure 1(a,b). We perform experiments on seven distinct NLP tasks, including classification and regression tasks that involve single sentences and sentence pairs, to assess the efficacy of DA. Our findings contest the previously held belief that DA does not enhance LMs' FT performance. We discover that continued pre-training LMs on augmented

[1] The code is available at `https://github.com/ZhengxiangShi/PowerfulPromptFT`.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

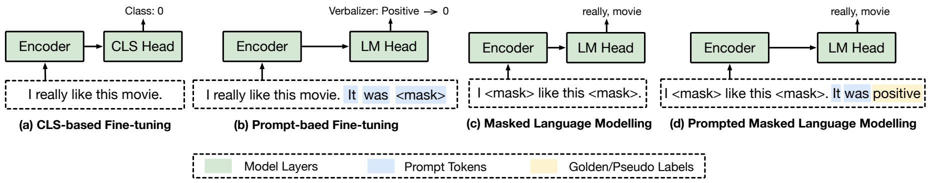

Fig. 1: The overview of CLS-based and prompt-based FT, along with their corresponding continued pre-training objectives.

data can largely improve the performance of FT approaches, offering an efficient alternative for enhancing model performance in practical applications.

2 Related Work

Prompt-based methods. In recent years, the exploration of prompt-based approaches has been conducted to enhance FT performance. PET/iPET [17] adapted the CLS-based FT [13] by presenting it as a masked language modelling problem, which is considered better suited for pre-training objectives, as illustrated in Figure 1. Subsequent studies further refined the application of templates and label words through automatic search mechanisms [15] or soft prompts that can be updated independently of any words [16].

Continued Pre-training. Earlier research, such as [12, 14], has questioned the efficacy of simple DA in improving FT performance for downstream tasks. Prior studies [18, 19] have demonstrated the effectiveness of continued pre-training on improving the model performance even with hundreds of unlabelled examples. However, the effectiveness of continued pre-training on the back translation augmented data is unclear in the context of few-shot learning.

3 Background

In this section, we provide a brief overview of FT approaches and their respective continued pre-training methods. Figure 1(a) illustrates the conventional CLS-based FT [13], which trains the output vector of the [CLS] token using an additional head layer. Further pre-training on task-related texts (see Figure 1c) before CLS-based FT typically leads to improved model performance [20, 18].

However, there exists a discrepancy between the pre-training objective and the CLS-based FT objective, leading to prompt-based research to enhance language model performance. Figure 1(b) demonstrates that prompt-based FT is designed as an MLM problem with the objective of predicting the masked token [17]. Specifically, the input text X is conditioned using a specific prompt template $\tilde{X} = \mathcal{T}(X)$, containing one special token, [MASK]. The prompt-based FT then connects the output vector related to the [MASK] token to a label word. The probability of predicting class $y \in \mathcal{Y}$ is calculated as follows:

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Dataset	$\|\mathcal{Y}\|$	L	#Train	#Test	Type	Labels (classification tasks)
SST-5	5	18	8,544	2,210	Sentiment	v. pos., positive, neutral, negative, v. neg.
MR	2	20	8,662	2,000	Sentiment	positive, negative
CR	2	19	1,775	2,000	Sentiment	positive, negative
MPQA	2	3	8,606	2,000	Opinion Polarity	positive, negative
Subj	2	23	8,000	2,000	Subjectivity	subjective, objective
TREC	6	10	5,452	500	Question cls.	abbr., entity, description, human, loc., num.
STS-B	$\mathcal{R}$	11/11	5,749	1,500	Sent. Similarity	-

Table 1: The datasets evaluated in this work. $|\mathcal{Y}|$: # of classes for classification tasks (with one exception: STS-B is a real-valued regression task over the interval $[0, 5]$). L: average # of words in input sentence(s).

$$p(y|X) = p(\texttt{[MASK]} = \mathcal{M}(y)|\tilde{X}), \tag{1}$$

where the verbalizer $\mathcal{M} : \mathcal{Y} \to \mathcal{V}$ maps the task label space to individual words in the vocabulary $\mathcal{V}$. Prompt-based FT can employ either hard or soft prompt templates $\mathcal{T}$, with label words possibly being part of the prompt templates as well [16]. Hard prompt templates [17] necessitate the careful design of prompts and label words for each task. However, the use of hard prompts was found to be sub-optimal and sensitive to prompt selection. Soft prompts [16] were proposed to utilize unused tokens from the vocabulary $\mathcal{V}$ or additional tokens as tunable embeddings for prompt templates, which can be directly trained with task-specific supervision. A recent study [19] proposed prompt-based continued pre-training prior to prompt-based FT to further enhance language model performance on downstream tasks, as depicted in Figure 1(d).

4 Experiments

In this section, we assess the impact of DA (*i.e.,* back translation) on all comparison methods. Additionally, we present datasets and baselines.

Datasets. Our study performs a comprehensive analysis of 7 NLP datasets, including classification and regression tasks. We derive 6 single-sentence tasks (SST-5 [21], MR [22], CR [23], MPQA [24], Subj [25], TREC [26]) and 1 sentence-pair English tasks (STS-B [27]), as shown in Table 1. According to [17, 15, 16, 19], we sample K-shot (K=16) per class from the full training set of each dataset.

Baselines. We train the K-shot examples using three different FT approaches, either incorporating back-translation as DA or not. The approaches are as follows: (1) "`CLS`-based FT": see Figure 1a; (2) "Prompt-based FT (hard)": FT with high-quality manual or auto-generated prompts and label words [17] (see Figure 1b). Please refer to Table 2 for the template details; and (3) "Prompt-based FT (soft)": FT with soft prompts using additional tokens for both templates and label words [16], where the same template is applied to all tasks (see Figure 1b). We use the SST-5 and STS-B templates for all single-sentence tasks and sentence pair tasks, respectively.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Task	Template	Label words
SST-5	<S_1> It was `[MASK]` .	v.positive: great, positive: good, neutral: okay, negative: bad, v.negative: terrible
MR	<S_1> It was `[MASK]` .	positive: great, negative: terrible
CR	<S_1> It was `[MASK]` .	positive: great, negative: terrible
MPQA	<S_1> is `[MASK]` .	positive: positive, negative: negative
Subj	<S_1> This is `[MASK]` .	subjective: subjective, objective: objective
TREC	`[MASK]` : <S_1>	abbreviation: Expression, entity: Entity, description: Description human: Human, location: Location, numeric: Number
STS-B	<S_1> `[MASK]` , <S_2>	y_u: Yes, y_l: No

Table 2: Templates and label words used for prompt-based FT.

To compare the effectiveness of direct supervision learning on augmented data from back-translation [28], we use augmented data as the corpus for continued pre-training with a masked language modelling objective. Consequently, we train these three types of FT approaches from three different types of checkpoints to evaluate their relative effectiveness: (i) the off-the-shelf RoBERTa-Large checkpoint [13]; (ii) the task-adaptive pre-training (TAPT) checkpoint [20, 29] for `CLS`-based FT; and (iii) the prompt-based continued pre-training (PCP) checkpoint [19] for prompt-based FT.

Training Details. We perform a grid search for learning rates within the set {1e5, 2e-5, 5e-5} with a batch size of 8. We train the model for 1,000 steps, evaluate performance every 100 steps, and select the best model based on the evaluation set. We augment each example using English-German and English-Russian translations, resulting in two augmented examples per original example.

Dataset Evaluation Metrics	SST-5 (acc)	MR (acc)	CR (acc)	MPQA (acc)	Subj (acc)	TREC (acc)	STS-B (Pear.)
Majority (full)	23.1	50.0	50.0	50.0	50.0	18.8	-
`CLS`-based FT	$41.7_{1.3}$	$76.3_{3.2}$	$79.5_{3.8}$	$65.1_{12.6}$	$91.7_{0.4}$	$80.3_{5.8}$	$46.0_{16.3}$
+ BT	$40.8_{2.0}\downarrow$	$71.1_{5.7}\downarrow$	$78.9_{3.2}\downarrow$	$69.2_{4.3}\uparrow$	$91.0_{1.9}\downarrow$	$83.1_{9.1}\uparrow$	$51.5_{22.6}\uparrow$
+ TAPT	$41.9_{2.2}\uparrow$	$76.1_{7.1}\downarrow$	$85.3_{3.6}\uparrow$	$75.3_{5.0}\uparrow$	$91.8_{1.2}\uparrow$	$83.8_{6.4}\uparrow$	$41.9_{19.0}\downarrow$
Prompt-based FT (hard)	$46.7_{1.5}$	$86.2_{1.2}$	$90.7_{0.8}$	$80.8_{6.9}$	$91.0_{1.1}$	$84.7_{4.4}$	$67.7_{8.1}$
+ BT	$45.4_{2.2}\downarrow$	$85.5_{1.3}\downarrow$	$91.1_{0.4}\uparrow$	$82.8_{5.1}\uparrow$	$91.3_{1.0}\uparrow$	$86.1_{4.3}\uparrow$	$66.3_{7.1}\downarrow$
+ PCP	$49.1_{1.5}\uparrow$	$87.0_{1.4}\uparrow$	$91.3_{0.9}\uparrow$	$85.9_{1.9}\uparrow$	$91.5_{1.3}\uparrow$	$86.8_{3.9}\uparrow$	$70.1_{8.1}\uparrow$
Prompt-based FT (soft)	$48.0_{0.7}$	$86.8_{1.4}$	$90.8_{1.3}$	$81.2_{6.8}$	$90.3_{2.1}$	$83.0_{3.0}$	$63.7_{6.8}$
+ BT	$46.7_{0.9}\downarrow$	$86.1_{1.4}\downarrow$	$91.0_{0.9}\uparrow$	$82.9_{1.5}\uparrow$	$90.8_{1.0}\uparrow$	$85.8_{2.6}\uparrow$	$69.1_{8.4}\uparrow$
+ PCP	$49.9_{1.2}\uparrow$	$85.9_{1.4}\downarrow$	$91.7_{1.2}\uparrow$	$84.6_{2.0}\uparrow$	$91.4_{1.5}\uparrow$	$86.3_{2.3}\uparrow$	$69.6_{7.9}\uparrow$

Table 3: Test results using RoBERTa-large, where mean and standard deviation are reported over 5 seeds. Green and red arrows indicate the positive/negative changes with respect to the FT baselines that do not involve the back-translation. The best performance on each dataset is highlighted in blue.

Results. Table 3 presents the performance of three different FT approaches, which involve using augmented examples as either supervised or continued pre-training training instances. Our experimental results reveal two primary observations: (1) using augmented examples for continued pre-training (TAPT or PCP) typically results in greater improvements compared to using them in supervised learning, and (2) continued pre-training occasionally leads to considerable performance enhancements. We delve into these findings below.

#1. Continued pre-training (TAPT or PCP) on three different FT approaches results in performance enhancements in 18 out of 21 cases, whereas using augmented data for supervised training leads to improvements in only 11 out of 21 cases. Furthermore, the average performance of FT with continued pre-training is 77.0% across all datasets and FT approaches, while the average performance of FT using supervised training on augmented data is approximately 75.5%. These results highlight the benefits of continued pre-training.

#2. In certain instances, conducting continued pre-training (TAPT or PCP) on LMs with augmented data before preceding the FT can lead to substantial improvements. Specifically, this approach enhances the performance of prompt-based FT (hard) from 46.7% to 49.1% on the SST-5 dataset and from 80.8% to 85.9% on the MPQA dataset. Notably, it boosts the performance of `CLS`-based FT from 65.1% to 75.3% on the MPQA dataset, resulting in an approximate 6% absolute value increase. These findings challenge the conclusions of prior research [14] suggesting that DA techniques yield only minor gains.

5 Conclusion

In conclusion, our study challenges the notion of data augmentation's limited impact on FT LMs in NLP tasks. We show that continued pre-training on augmented data can effectively improve model performance.

References

[1] Pin Ni, Yuming Li, Gangmin Li, and Victor Chang. Natural language understanding approaches based on joint task of intent detection and slot filling for iot voice interaction. *Neural Computing and Applications*, 2020.

[2] Pin Ni, Qiao Yuan, Raad Khraishi, Ramin Okhrati, Aldo Lipani, and Francesca Medda. Eigenvector-based graph neural network embeddings and trust rating prediction in bitcoin networks. ICAIF '22, 2022.

[3] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *AAAI 2022*.

[4] Zhengxiang Shi, Pin Ni, Meihui Wang, To Eun Kim, and Aldo Lipani. Attention-based ingredient parser. In *ESANN*, Bruges, Belgium, 2022.

[5] Hossein A. Rahmani, Mohammad Aliannejadi, Mitra Baratchi, and Fabio Crestani. Joint geographical and temporal modeling based on matrix factorization for point-of-interest recommendation. In *ECIR*. Springer, 2020.

[6] Xiao Fu and Aldo Lipani. Priming and actions: An analysis in conversational search systems. SIGIR, 2023.

[7] Xiao Fu, Emine Yilmaz, and Aldo Lipani. Evaluating the cranfield paradigm for conversational search systems. ICTIR, 2022.

[8] Zhengxiang Shi, Xi Wang, and Aldo Lipani. Self contrastive learning for session-based recommendation. *arXiv preprint arXiv:2306.01266*, 2023.

[9] Zhengxiang Shi, Yue Feng, and Aldo Lipani. Learning to execute actions or ask clarification questions. In *Findings of NAACL 2022*.

[10] Mariya Hendriksen, Maurits Bleeker, Svitlana Vakulenko, Nanne van Noord, Ernst Kuiper, and Maarten de Rijke. Extending clip for category-to-image retrieval in e-commerce. In *ECIR*, 2022.

[11] Zhengxiang Shi, Jerome Ramos, To Eun Kim, Xi Wang, Hossein A Rahmani, and Aldo Lipani. When and what to ask through world states and text instructions: Iglu nlp challenge solution. *NeurIPS IGLU Competition Workshop*, 2023.

[12] Shayne Longpre, Yu Wang, and Chris DuBois. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of EMNLP 2020*. ACL, 2020.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. FlipDA: Effective and robust data augmentation for few-shot learning. In *ACL*. ACL, May 2022.

[15] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, pages 3816–3830, Online, August 2021. ACL.

[16] Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. In *ICLR*, 2022.

[17] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *ACL*. ACL, April 2021.

[18] Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. Rethinking semi-supervised learning with language models. In *Findings of ACL 2023*, Toronto, Canada, 2023. Association for Computational Linguistics.

[19] Zhengxiang Shi and Aldo Lipani. Don't stop pretraining? make prompt-based fine-tuning powerful learner. In *Arxiv*, 2023.

[20] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, pages 8342–8360. ACL, July 2020.

[21] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *emnlp*, 2013.

[22] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *acl*, 2005.

[23] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[24] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3), 2005.

[25] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *acl*, 2004.

[26] Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In *SIGIR*, 2000.

[27] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. 2017.

[28] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT*, 2019.

[29] Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of EMNLP 2022*.