Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding

Bolei Ma, Ercong Nie, Helmut Schmid, Hinrich Schütze

Ludwig Maximilian University of Munich

bolei.ma@lmu.de

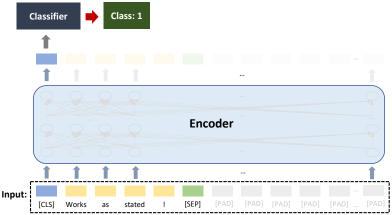
@ KONVENS 2023, Ingolstadt

September 19, 2023

Outline

Research Subject and Motivation

- Prompt-Based Finetuning (ProFiT) vs. Vanilla Finetuning



(a) Vanilla finetuning (b) Prompt-based finetuning

Figure 1: The comparison of vanilla finetuning and prompt-based finetuning. [CLS], [SEP], [MASK], [PAD] are special tokens in the encoder vocabulary. The verbalizer is a function mapping from the task label set to a subset of the encoder vocabulary. Input tokens in blue represent the prompt pattern.

Research Subject and Motivation

- Prompt-based learning has recently emerged as a notable advancement, surpassing regular finetuning approaches (Liu et al., 2023).
- A detailed investigation of zero-shot[1] cross-lingual transfer performance of prompt-based learning on NLU has not yet been carried out.
- It is interesting to further analyze the underlying linguistic factors which could affect the zero-shot cross-lingual performance of prompt-based learning

[1] In our work, “zero-shot” in “zero-shot cross-lingual tranfer” refers to the number of target language training data, i.e., no target language data is provided.

Research Questions

- **RQ1**: Does prompt-based finetuning outperform vanilla finetuning in the zero-shot cross-lingual transfer performance in different NLU tasks?
- **RQ2**: Is prompt-based finetuning always better than vanilla finetuning?
- **RQ3**: What underlying factors could affect the cross-lingual performance of prompt-based finetuning?

Outline

Method: ProFiT Pipeline

- **Training**: A fixed prompt pattern $P(X)$ in the source language transforms the input text X into a cloze-style question with a mask token. A verbalizer is used to map the original labels onto words. The sentence classification task of vanilla finetuning is changed into a masked token prediction task.
- **Inference**: In the cross-lingual setting, we simply apply the same functions P and v to the target language examples without further modifications.

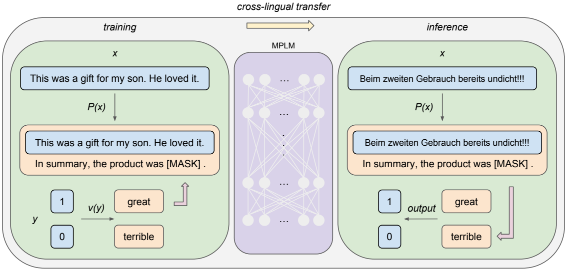


Figure 2: ProFiT pipeline of training and cross-lingual transfer with examples. X is an input sentence and $P(X)$ denotes the prompt pattern which reformulates the input into a prompt. $v(y)$ is the verbalizer which maps each class label y onto a word from the source language vocabulary.

Outline

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Datasets

- In order to investigate the performance on diverse NLU tasks, three representative different classification tasks on NLU are selected for evaluation:
  - Multi-class sentiment analysis task on **Amazon product reviews** (Keung et al., 2020) in 6 languages,
  - Binary paraphrase identification task on **PAWS-X** (Yang et al., 2019) in 7 languages, and
  - Multi-class natural language inference task on **XNLI** (Conneau et al., 2018) in 15 languages.

Prompt Design for the Datasets

- **Amazon Reviews Dataset**:
  - $P(X) = X \circ$ "All in all, it was [MASK]."
  - $v(1) =$ "terrible", $v(2) =$ "bad", $v(3) =$ "ok", $v(4) =$ "good", $v(5) =$ "great"
- **PAWS-X**:
  - $P(X_1, X_2) = X_1 \circ$ "? [MASK], " $\circ X_2$
  - $v(0) =$ "Wrong", $v(1) =$ "Right"
- **XNLI**:
  - $P(X_1, X_2) = X_1 \circ$ "? [MASK], " $\circ X_2$
  - $v(0) =$ "Yes", $v(1) =$ "Maybe", $v(2) =$ "No"

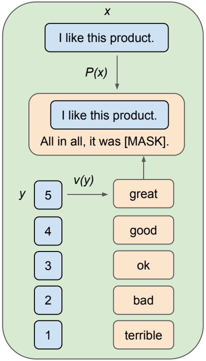


Figure 3: A prompt example for Amazon Dataset

Multilingual Models

- Multilingual BERT model (Devlin et al., 2019) "`bert-base-multilingual-cased`" (M)
- XLM-R model (Conneau et al., 2020) "`xlm-roberta-base`" (X)

Outline

Main Results

- Overall, ProFiT outperforms the Vanilla baseline with both mBERT and XLM-R models on all three classification tasks. [2]

	Amazon	PAWS-X	XNLI	Avg.
Vanilla-mBERT	42.97	80.24	65.05	62.75
ProFiT-mBERT	**43.98**	**82.16**	**65.79**	**63.98**
Vanilla-XLM-R	54.56	82.51	73.61	70.22
ProFiT-XLM-R	**54.66**	**82.73**	**73.82**	**70.40**

Table 1: Overview of results

[2] To avoid random effects on training, we trained each experiment with 5 different random seeds and take the average results.

Main Results

- While the overall performance of ProFiT is better than Vanilla for all three tasks in both mBERT and XLM-R settings, slight differences between languages can be noticed. → In Section 5, we will therefore further investigate how language factors influence cross-lingual transfer performance.

Task	Model	en	ar	bg	de	el	es	fr	hi	ja	ko	ru	sw	th	tr	ur	vi	zh	avg.
Amazon	Vanilla-M	58.92	-	-	45.69	-	48.02	47.45	-	35.07	-	-	-	-	-	-	-	**38.63**	42.97
	ProFiT-M	**59.05**	-	-	**46.66**	-	**49.30**	**48.38**	-	**37.31**	-	-	-	-	-	-	-	38.26	**43.98**
	Vanilla-X	59.61	-	-	**60.14**	-	55.24	55.66	-	51.93	-	-	-	-	-	-	-	49.82	54.56
	ProFiT-X	**60.06**	-	-	59.60	-	**55.72**	**55.89**	-	**52.34**	-	-	-	-	-	-	-	**49.75**	**54.66**
PAWS-X	Vanilla-M	93.85	-	-	84.94	-	87.11	86.55	-	73.39	72.44	-	-	-	-	-	-	77.01	80.24
	ProFiT-M	**94.21**	-	-	**86.06**	-	**88.17**	**87.91**	-	**75.79**	**75.82**	-	-	-	-	-	-	**79.22**	**82.16**
	Vanilla-X	94.33	-	-	86.92	-	88.55	89.04	-	**76.07**	74.71	-	-	-	-	-	-	79.75	82.51
	ProFiT-X	**94.90**	-	-	**87.06**	-	**88.87**	**88.86**	-	75.53	**75.40**	-	-	-	-	-	-	**80.63**	**82.73**
XNLI	Vanilla-M	82.57	65.12	68.97	71.40	66.30	74.22	73.68	60.02	-	-	68.95	50.24	53.15	62.02	57.96	69.80	68.91	65.05
	ProFiT-M	82.57	**65.55**	**69.47**	**71.57**	**67.43**	**75.10**	**74.57**	**60.57**	-	-	**69.55**	**51.13**	**54.58**	**62.64**	**58.04**	**70.74**	**70.08**	**65.79**
	Vanilla-X	84.91	**71.86**	77.78	76.86	75.96	79.25	78.21	69.92	-	-	**75.79**	**65.21**	72.02	73.12	66.07	74.71	73.72	73.61
	ProFiT-X	**84.97**	71.81	**77.92**	**77.35**	**76.11**	**79.31**	**78.75**	**70.10**	-	-	75.43	65.13	**72.39**	**73.23**	**66.95**	**75.05**	**73.92**	**73.82**

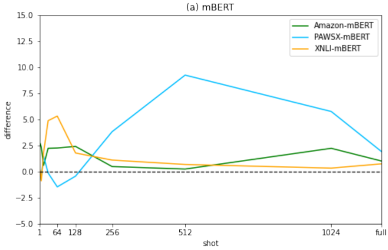
Table 2: Detailed cross-lingual performance results on three classification tasks. When calculating the average (avg.), due to the aim of zero-shot cross-lingual transfer, the performance results of the source language English are not taken into account. Model M stands for mBERT, and X for XLM-R.

Few-Shot Ablations

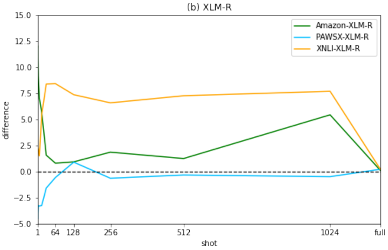
- Previous studies show that the prompt framework is more effective than finetuning when training data is scarce (Zhao and Schütze, 2021; Qi et al., 2022).
- We investigate how the performance changes as the number of training samples K increases in few-shot settings.
- The training data is randomly sampled with $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ shots per class from the English training data.

Few-Shot Ablations

- Results of few-shot ablations show that prompt-based finetuning exhibits greater advantages in most few-shot scenarios, with different performance patterns dependent on task types:



(a) mBERT



(b) XLM-R

Figure 4: Performance difference between ProFiT and Vanilla in different few-shot settings and full training on three tasks with both mBERT and XLM-R models.

Outline

Language Features

- We analyze the following language factors that could impact the cross-lingual performance:
    - **Target Languages Size (Size)**: The pretraining corpus size of the target languages is measured by the log_2 of the number of articles in Wikipedia.
    - Language Similarity:
        - **Typological & Phylogenetic Similarity ($\mathbf{Sim_1}$)**: Following the LANG2VEC approach (Littell et al., 2017), which provides information-rich vector representations of languages from different linguistic and ethnological perspectives, We adopt five linguistic categories: syntax (SYN), phonology (PHO), phonological inventory (INV), language family (FAM), and geography (GEO).
        - **Lexical Similarity ($\mathbf{Sim_2}$)**: The lexical similarity metric is based on a mean normalized pairwise Levenshtein distance matrix from ASJP (Wichmann et al., 2022). Two dimensionality reduction methods are employed: Uniform Manifold Approximation and Projection (*UMAP*) (McInnes et al., 2018) and Singular Value Decomposition (*SVD*) (Stewart, 1993).

Language Features and Task Performance

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

lang	Typological & Phylogenetic Sim.						Lexical Sim.			Size	Task Performance					
	SYN	PHO	INV	FAM	GEO	**Sim$_1$**	*UMAP*	*SVD*	**Sim$_2$**		amazon-M	amazon-X	pawsx-M	pawsx-X	xnli-M	xnli-X
ar	65.47	70.06	75.88	0.00	97.04	**61.69**	-1.90	4.87	**1.49**	**20.20**	-	-	-	-	65.55	71.81
bg	78.78	90.45	70.02	13.61	99.01	**70.38**	8.65	33.21	**20.93**	**18.15**	-	-	-	-	69.47	77.92
de	79.05	83.62	77.62	54.43	99.76	**78.90**	83.42	76.83	**80.13**	**21.42**	46.66	59.60	86.06	87.06	71.57	77.35
el	73.19	95.35	64.75	14.91	98.95	**69.43**	1.24	24.81	**13.03**	**17.76**	-	-	-	-	67.43	76.11
es	84.97	85.81	64.99	9.62	99.59	**69.00**	1.61	28.30	**14.96**	**20.83**	49.30	55.72	88.17	88.87	75.10	79.31
fr	76.83	75.26	73.64	9.62	99.93	**67.06**	1.34	31.76	**16.55**	**21.27**	48.38	55.89	87.91	88.86	74.57	78.75
hi	58.79	85.81	76.53	12.60	91.10	**64.97**	1.20	21.11	**11.16**	**17.26**	-	-	-	-	60.57	70.10
ja	49.63	64.44	65.92	0.00	85.65	**53.13**	-	-	-	**20.39**	37.31	52.34	75.79	75.53	-	-
ko	55.66	74.62	71.04	0.00	86.93	**57.65**	-0.22	12.42	**6.10**	**19.28**	-	-	75.82	75.40	-	-
ru	75.74	90.45	63.17	16.67	95.81	**68.37**	8.63	32.60	**20.62**	**20.87**	-	-	-	-	69.55	75.43
sw	42.26	90.91	76.16	0.00	91.50	**60.17**	-9.05	-7.18	**-8.12**	**16.23**	-	-	-	-	51.13	65.13
th	65.20	81.82	78.88	0.00	85.25	**62.23**	-0.21	3.82	**1.81**	**17.25**	-	-	-	-	54.58	72.39
tr	43.36	85.81	68.49	0.00	98.25	**59.18**	-7.80	-1.56	**-4.68**	**19.00**	-	-	-	-	62.64	73.23
ur	50.01	0.00	71.56	12.60	92.54	**45.34**	1.35	24.92	**13.14**	**17.54**	-	-	-	-	58.04	66.95
vi	64.92	78.33	74.76	0.00	85.25	**60.65**	0.86	-18.50	**-8.82**	**20.29**	-	-	-	-	70.74	75.05
zh	73.49	78.33	74.91	0.00	88.42	**63.03**	-	-	-	**20.37**	38.26	49.75	79.22	80.63	70.08	73.92

Table 3: Overview of language features and task performance with ProFiT for correlation analysis.

Correlation Analysis

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- On XNLI, significant correlations can be found especially with the typological & phylogenetic similarity and target language size.
- On PAWS-X and Amazon, more insignificant correlations with the proposed factors have been found, which could be due to the limited number of languages in their test data: PAWS-X and Amazon only contain 7 and 6 languages respectively, while XNLI has 15 different languages.

Task	Model	Stat.	Sim_1		Sim_2		Size	
			corr.	*p*	*corr.*	*p*	*corr.*	*p*
Amazon	ProFiT-M	P	0.73	0.16*	-0.95	0.21*	0.81	0.09*
		S	0.70	0.19*	-1.00	0.00	0.50	0.39*
	ProFiT-X	P	0.80	0.10*	1.00	0.01	0.92	0.03
		S	0.80	0.10*	1.00	0.00	1.00	1e-24
PAWS-X	ProFiT-M	P	0.82	0.05	0.31	0.69*	0.82	0.04
		S	0.83	0.04	0.20	0.80*	0.60	0.21*
	ProFiT-X	P	0.83	0.04	0.34	0.66*	0.84	0.04
		S	0.77	0.07*	0.20	0.80*	0.71	0.11*
XNLI	ProFiT-M	P	0.57	0.03	0.43	0.14*	0.86	9e-05
		S	0.59	0.03	0.53	0.06*	0.90	1e-05
	ProFiT-X	P	0.72	4e-03	0.43	0.14*	0.70	5e-03
		S	0.77	1e-03	0.63	0.02	0.72	4e-03

Table 4: Correlations between task performance and language similarities (Sim_1 & Sim_2) and target language size (Size), based on Pearson (P) and Spearman (S) test. Insignificant results with a p value > 0.05 are marked with *.

Correlation Analysis

- To sum up, language similarity and size are two factors that could impact the cross-lingual performance in our study, and we find more significant correlations when the test set contains a larger amount of languages.

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Outline

Conclusion

- ProFiT outperforms vanilla finetuning in zero-shot cross-lingual transfer performance on the three sentence classification tasks – multi-class sentiment classification, binary paraphrase identification, and multi-class natural language inference.
- The performance improvement of ProFiT is generally more obvious in few-shot scenarios.
- The similarity of the source and target language and the size of the target language pretraining data impact the cross-lingual transfer performance of ProFiT, especially on a big dataset with a variety of test languages.

Future Work

- **Different Tasks**: including question answering, parsing, knowledge probing, generation, etc.
- **Prompt Engineering**: Future work should pay more attention to methods that automatically apply a suitable prompt for finetuning. Also dynamic prompt applications could be taken into account, for the purpose of looking for a best-performing prompt.
- **Linguistic Insights**: Further research on the correlation of language features and model performance could be conducted, with more features and languages, as well as the impact of different prompt designs on the language features.

Thanks for your attention.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for

Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018.

Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Gilbert W Stewart. 1993. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566.

Søren Wichmann, Eric W. Holman, and Cecil H. 2022. The asjp database. Version 20.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Appendix

Appendix

ProFiT: Formal Description

- Let $D = \{(X_1, y_1), ..., (X_n, y_n)\}$ denote training examples, $y_1, ..., y_n$ class labels, $P(.)$ the prompt pattern, and $v(.)$ the verbalizer.
- The pretrained language model M with trainable parameters θ performs masked token prediction and returns the probabilities $p = M(P(X), \theta)$ of all candidate words for the masked token in $P(X)$.
- We predict the class $\hat{y}$ whose verbalizer $v(\hat{y})$ received the highest probability from model M:

$$\hat{y} = \arg\max_{y \in Y} p(v(y)) \tag{1}$$

- We finetune the parameters θ of model M by minimizing the cross-entropy loss function ℓ on D:

$$\hat{\theta} = \arg\max_{\theta} \sum_{(X,y) \in D} \ell(v(y), M(P(X), \theta)) \tag{2}$$