RESEARCH ARTICLE

FINE-TUNING PRE-TRAINED LANGUAGE MODELS FOR VIETNAMESE: A COMPARATIVE STUDY OF FULL FINE-TUNING AND LORA

Luu Van Nhat Hao, Dinh Hung, Dinh Minh Hoa, Nguyen Van Xanh

Ho Chi Minh City University of Foreign Languages and Information Technology, Vietnam

22DH110913@st.huflit.edu.vn, hungd@huflit.edu.vn, hoadm@huflit.edu.vn, xanhnv@huflit.edu.vn

ABSTRACT—Pre-trained language models have brought significant improvements to Vietnamese natural language processing tasks. However, full fine-tuning of these large models remains resource-intensive and poses challenges for settings with limited computational capacity. This paper presents a comparative study between full fine-tuning and Low-Rank Adaptation – a parameter-efficient fine-tuning method, focusing on the trade-off between model performance and resource usage. Experiments are conducted on two core NLP tasks – sentiment analysis and named entity recognition – using benchmark Vietnamese datasets and pre-trained models such as PhoBERT, BARTpho and ViT5. The results show that LoRA achieves comparable accuracy to full fine-tuning while significantly reducing training cost, especially in transformer-based architectures. These findings suggest that LoRA is a viable and efficient alternative to full fine-tuning for fine-tuning Vietnamese PLMs in low-resource environments. Our work provides practical insights and experimental benchmarks to support informed decision-making in selecting fine-tuning strategies for Vietnamese NLP applications.

Keywords—Vietnamese Language Models, Fine-tuning, LoRA, Sentiment Analysis, Named Entity Recognition.

I. INTRODUCTION

In recent years, pre-trained language models (PLMs) have become foundational to a wide range of natural language processing (NLP) applications, driving substantial progress in tasks such as sentiment analysis, named entity recognition, and machine translation. For Vietnamese, models like PhoBERT, BARTpho, and ViT5 have demonstrated remarkable effectiveness when fine-tuned for specific tasks, particularly in scenarios where available data is limited.

Full fine-tuning (FFT) is the most widely used approach, where all model parameters are updated during training. Although this method can deliver high performance, it requires substantial computational resources and significant memory capacity, which poses challenges in environments with limited infrastructure-such as educational institutions, small research organizations, or small and medium-sized enterprises. These constraints have driven the development of parameter-efficient fine-tuning (PEFT) techniques, aiming to preserve model performance while minimizing computational costs. Among recent advances, Low-Rank Adaptation (LoRA) stands out as a prominent method, enabling the fine-tuning of large models by introducing additional low-rank parameters into specific layers without altering the original model architecture. LoRA is regarded as highly promising for its ability to conserve resources while still achieving accuracy comparable to FFT.

This paper presents an empirical study comparing two fine-tuning approaches, FFT and LoRA, for adapting Vietnamese language models to two common tasks: sentiment analysis and named entity recognition. We employ established pre-trained models, including PhoBERT, BARTpho, ViT5, and PhoBERT-v2, alongside two standard benchmark datasets: UIT-VSFC and PhoNER_COVID19. The study focuses on evaluating model performance from two perspectives: accuracy and resource consumption during training.

Through quantitative results and comparative analysis, we aim to provide practical insights for the Vietnamese NLP research and development community, especially in contexts where cost optimization remains a critical consideration.

II. RELATED WORK

The rapid advancement of PLMs has marked a significant breakthrough in NLP, particularly in tasks such as sentiment analysis and named entity recognition. For Vietnamese, numerous studies have demonstrated the superior effectiveness of PLMs like PhoBERT [1], BARTpho [2] and ViT5 [3] when fully fine-tuned on domain-specific datasets. For instance, PhoBERT has achieved strong results in both text classification and sequence labeling tasks, while ViT5 extends the capabilities of Vietnamese NLP to a broader range of text generation problems.

For sentiment analysis tasks, the study by Nguyen et al [1] demonstrates that fine-tuning PLMs like PhoBERT on the UIT-VSFC dataset significantly improves accuracy compared to traditional models. Meanwhile, BARTpho and ViT5, with their encoder-decoder architectures, exhibit superior capabilities in processing semantically complex texts. In named entity recognition tasks, models such as PhoBERT and XLM-R have been fine-tuned and evaluated on multiple Vietnamese datasets, including VLSP 2016 and PhoNER_COVID19, demonstrating strong generalization abilities under resource-constrained conditions.

However, fine-tuning the entire model demands substantial computational resources, particularly in terms of memory and training time, especially for large-scale models. This challenge has driven the development of PEFT methods [4], among which LoRA stands out [5]. LoRA operates by inserting low-rank matrices into the attention layers of the model, significantly reducing the number of trainable parameters while maintaining high performance. Recent studies [6] [7] have demonstrated LoRA's effectiveness not only for large language models (LLMs) but also in multilingual and low-resource contexts.

Although numerous studies have evaluated LoRA on English or multilingual models, research specifically applying this method to Vietnamese remains limited. Some recent works have begun to explore LoRA in tasks such as summarization, text classification, and Vietnamese chatbots [8] [9], however, there has not yet been a systematic comparison of the effectiveness of FFT and LoRA on tasks like sentiment analysis and named entity recognition. This gap serves as the motivation for the present study, which aims to provide a comprehensive quantitative assessment of both performance and resource cost between these two fine-tuning strategies, using representative Vietnamese models and benchmark datasets. The findings are intended to offer reference data for practical NLP applications in resource-constrained environments, supporting more informed decision-making in model deployment.

III. METHODOLOGY

A. DATASETS

This study focuses on two core tasks in Vietnamese natural language processing: sentiment analysis and named entity recognition. For sentiment analysis, the UIT-VSFC dataset [10]. is employed, containing 16,175 labeled student feedback samples categorized as positive, neutral, or negative. The dataset is divided into training (11,426 samples), validation (1,583 samples), and test sets (3,166 samples). Recent experiments using a hybrid PhoBERT-CNN-LSTM model achieved 93.24% accuracy and a 92.92% F1-score on UIT-VSFC, highlighting both the dataset's complexity and reliability [11].

For named entity recognition, the PhoNER_COVID19 [12] dataset is utilized, comprising 10,000 sentences extracted from over 35,000 Vietnamese COVID-19 articles. These sentences are manually annotated with six entity types: Disease, Symptom, Drug/Vaccine, Organization, Location, and Time. The dataset is split into training (7,000 sentences), validation (1,500 sentences), and test sets (1,500 sentences). Evaluations on PhoNER_COVID19 show that an XLM-RoBERTa-BiLSTM-CRF model achieves a 91.2% F1-score, outperforming the PhoBERT-CRF baseline (89.7% F1-score). These results underscore the dataset's specialization in healthcare while emphasizing the challenges of processing complex compound words and semantic diversity in Vietnamese.

B. VIETNAMESE PRE-TRAINED MODELS

In this study, three Vietnamese-specific PLMs – PhoBERT, BARTpho, and ViT5-were selected for experimentation. Each model represents a distinct approach within NLP, ranging from semantic representation to text generation and text-to-text transformation.

PhoBERT is the first monolingual language model specifically designed for Vietnamese, developed based on the RoBERTa architecture and pre-trained on a 20GB corpus of Vietnamese text collected from Wikipedia and news sources. Its preprocessing pipeline employs word segmentation using RDRSegmenter in combination with Byte-Pair Encoding (BPE), enabling the model to effectively adapt to the linguistic characteristics of Vietnamese. Unlike the original BERT, PhoBERT removes the Next Sentence Prediction objective and retains only the Masked Language Modeling task, thereby enhancing the learning of semantic representations. PhoBERT is available in two versions: the base-v2 variant, which has 12 Transformer layers, a hidden size of 768, and 12 attention heads; and the large version, featuring 24 Transformer layers, a hidden size of 1024, and 16 attention heads. Both versions have demonstrated superior performance across various Vietnamese NLP tasks-including text classification, sentiment analysis, and named entity recognition, especially when compared with multilingual models or traditional approaches.

BARTpho was developed based on the BART large architecture as a sequence-to-sequence generative model, featuring a symmetric encoder-decoder structure with 12 layers in each component. A notable distinction of BARTpho is its provision of two preprocessing schemes: a word-based version and a syllable-based version. In the syllable-based variant, words are segmented into discrete syllables (for example, “nghiên cứu” becomes “nghiên | cứu”), enabling the model to better capture morphological features and more effectively address word boundary ambiguities, a major challenge in Vietnamese language processing.

ViT5 is the first text-to-text model specifically designed for Vietnamese, built upon the T5 architecture with an encoder-decoder mechanism and trained on a large-scale corpus of 120GB Vietnamese text sourced from news articles, books, and Wikipedia. This model is optimized for natural language generation tasks such as summarization, translation, and question answering. A notable feature of ViT5 is its ability to efficiently process long text passages exceeding 512 tokens, enabled by dynamic text segmentation techniques, which extends its applicability to tasks requiring deep and continuous contextual understanding.

Table 1 - Overview of the Models

	PhoBERT (base-v2)	**PhoBERT** (large)	**BARTpho** (word)	**ViT5** (base)
Architecture	RoBERTa-base	RoBERTa-large	BART (seq2seq)	T5 (text-to-text)
Parameters (M)	135	370	420	222
Max-length (tokens)	256	256	-	256 or 1024

Overall, these three models represent the principal directions in Vietnamese NLP: PhoBERT emphasizes semantic and syntactic representation through its Transformer encoder architecture; BARTpho leverages the strengths of a multi-task generative model with an encoder-decoder structure; and ViT5 is oriented toward text-to-text transformation, offering robust language generation capabilities and scalability for long input sequences. Comparing the effectiveness of these models on sentiment analysis and named entity recognition tasks provides a more comprehensive perspective on the practical applicability of each architecture within the Vietnamese context. The architecture and parameter count of the models are summarized in Table-1.

C. FINE-TUNE TECHNIQUES

Fine-tuning pre-trained LLMs is a critical step in adapting them to specific tasks within the Vietnamese context. In this study, two widely used techniques are applied: FFT and LoRA, representing distinct approaches in adjustment scope and computational cost.

With FFT, all model weights are updated during the learning process. Specifically, parameters $\theta \in \mathbb{R}^{d \times k}$ are optimized to minimize the loss function $\mathcal{L}(\theta)$, according to the equation:

$$\theta_{\mathbf{new}} = \theta_{\mathbf{pre-trained}} - \eta \nabla_{\theta} \mathcal{L}(\theta) \qquad (1)$$

Where η is the learning rate. This method has the potential to achieve optimal performance since the entire model is synchronously adapted to the target task. However, FFT demands substantial computational resources, with GPU memory usage reaching three to four times the size of the original model, making it impractical for deployment in resource-constrained environments. To address this limitation, LoRA is introduced as a more lightweight alternative. Instead of updating all weights, LoRA keeps the original parameters frozen and inserts additional trainable low-rank matrices into the Attention layers of the Transformer architecture. Given a weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA represents the update ΔW as the product of two smaller matrices (see Figure 1), computed as Equation 2:

$$\Delta W = BA \qquad (2)$$

With $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ where $r \ll min(d, k)$. The number of trainable parameters is reduced from dk to $r(d + k)$. For example, with $d = 768, k = 768$; FFT requires 589K parameters, while LoRA only needs 12K parameters when $r = 8$. This mechanism reduces GPU memory usage by 70–90% and accelerates training by 25–40% compared to FFT, while also enabling multi-task integration through the swapping of the matrix A and B.

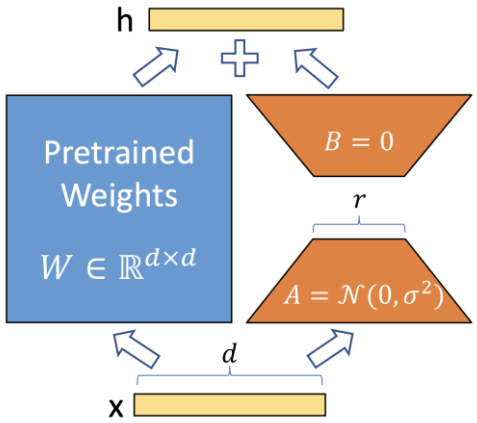

Figure 1. Weight Update Mechanism of the LoRA Technique [5]

The choice of hyperparameters in LoRA, specifically the rank (r) and the scaling factor (α), is also very important in terms of training efficiency and the quality of the final model [13]. Increasing the intrinsic dimension of the update matrices proxied through LoRA can result in improved performance, as indicated through recent work [14]. Optimizing the rank usually needs to involve a costly and brute-force search process. The choice of rank r influences the size of the trainable subspace in LoRA and essentially the model's representational ability. Experiments have shown that different modules can use different ranks, and these can potentially change dynamically during the fine-tuning process. Lower ranks diminish the number of trainable parameters but potentially restrict expressiveness, while larger ranks are more computationally expensive but can lead to improved performance. The α parameter, however, controls the degree to which the LoRA modifications impact the underlying model. The latest breakthroughs, including ALLoRA [15], have indicated that this hyperparameter can actually be dispensed with using adaptive learning rate techniques and therefore the need for careful tuning of α can potentially be substituted with automated methods. We use $r = 16$ and $\alpha = 64$ as implemented through hyperparameter tuning in this work to find a compromise between model fidelity and computation cost. This is consistent with recommendations favoring medium ranks to prevent overfitting while maintaining adequate representational capacity.

Overall, these two techniques embody the trade-off between performance and resource cost, enabling practitioners to select the most appropriate approach for each specific task and deployment environment.

D. EXPERIMENTAL SETUPS

In this study, the experiments are designed to evaluate both the effectiveness and resource costs of two model fine-tuning techniques – FFT and LoRA – applied to two representative tasks in Vietnamese NLP: sentiment analysis and named entity recognition. These tasks exemplify two principal approaches, namely classification and sequence labeling, and both require deep semantic features, making them suitable benchmarks for assessing the adaptability of large language models. The experimental procedures for each method follow the workflow illustrated in Figure 2.

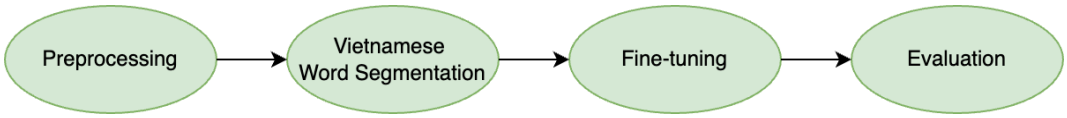

Figure 2 - Experimental Procedure

For the sentiment analysis task, both fine-tuning techniques were applied to four pre-trained model variants: PhoBERT-base-v2, PhoBERT-large, BARTpho-word, and ViT5-base. In contrast, the named entity recognition task was conducted only on the two PhoBERT variants-PhoBERT-base-v2 and PhoBERT-large-since this model architecture is better suited for sequence labeling and has demonstrated strong performance in previous studies. All training and evaluation procedures were carried out within the same hardware environment to ensure objective comparison between methods. The system setup included two *NVIDIA RTX A5000 GPUs* (each with 24GB VRAM), an *Intel Core™ i9-10900X processor*, 128GB RAM, and Ubuntu 20.04. This configuration is sufficiently powerful to meet the high computational demands of FFT, while also allowing for precise monitoring and measurement of GPU resource consumption throughout training.

In the LoRA experiments, the models were configured with a low-rank value of $r = 16$ and a scaling factor $\alpha = 64$. The insertion of low-rank matrices was limited to the query and value layers within the Attention component, following recent research recommendations to balance model performance and resource efficiency. LoRA was integrated using the PEFT library, enabling flexible application of this technique across different model architectures. For FFT, all model weights were updated during training, ensuring thorough adaptation but also requiring greater hardware resources.

Both methods utilized the *AdamW* optimizer with a learning rate of 2×10^{-5} and a batch size of 32. Training conditions-including datasets, number of epochs, and hyperparameter settings-were kept consistent across models and methods to ensure uniformity and high comparability of the results obtained.

E. EVALUATION METRICS

To evaluate the effectiveness of the models on the two natural language processing tasks, this study adopts widely used metrics tailored to the characteristics of each task. Specifically, for sentiment analysis, Accuracy (3) and F1-score (4) are selected to comprehensively capture model performance in terms of absolute prediction correctness and the balance between recall and precision. The combination of these two metrics enables assessment of model stability, particularly in scenarios where label distribution may be imbalanced.

For named entity recognition, the study focuses on the weighted F1-score (5) a frequency-weighted variant of the F1-score to more accurately reflect labeling performance under class imbalance. The weighted F1-score is especially important in the context of named entity recognition, where certain entity types such as

"Organization," "Location," or "Person" may occur at very different frequencies, making unweighted averages potentially misleading for model evaluation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Prevision \times Recall}{Prevision + Recall} \quad (4)$$

Where:

$$Precision = \frac{TP}{TP + FP}\ ;\ Recall = \frac{TP}{TP + FN}$$

$$\text{Weighted F1} - \text{score} = \sum_{i=1}^{N} \frac{n_i}{n} \times F1_i \quad (5)$$

All evaluation metrics are computed on an independent test set, ensuring no overlap with training or fine-tuning data, thereby providing an accurate measure of the model's generalization ability. Consistent use of these criteria for each task also enhances the objectivity and reliability of comparisons between fine-tuning methods.

IV. RESULT AND DISCUSSION

This section presents and analyzes the experimental results to compare the performance of the two fine-tuning techniques: Full Fine-tuning and Low-Rank Adaptation. The evaluation metrics considered include Accuracy, F1-score, Weighted F1-score, number of trainable parameters, training time, and GPU memory usage. Results are analyzed separately for the two tasks: sentiment analysis and named entity recognition.

A. SENTIMENT ANALYSIS RESULTS

The results obtained from training with the two fine-tuning techniques – FFT and LoRA on four Vietnamese language models, namely PhoBERT-base-v2, PhoBERT-large, BARTpho-word, and ViT5-base, are summarized in Table 2. The evaluation metrics include accuracy, F1-score, the number of trainable parameters, average training time, and GPU memory usage.

Table 2 - Comparison of Performance and Resource Costs on the Sentiment Analysis Task

Models	Technique	Acc (%)	F1	Params	RAM (GB)	Time (min)
PhoBERT-base-v2	FFT	94.31	0.8349	135M	10.4	149
	LoRA	93.27	0.8147	1,5M	6.8	119
PhoBERT-large	FFT	93.94	0.8277	369M	43.4	367
	LoRA	92.45	0.8018	2.9M	24.1	338
BARTpho-word	FFT	94.19	0.7983	421M	47.7	361
	LoRA	93.27	0.7922	1.2M	30.1	307
ViT5-base	FFT	92.29	0.8069	220M	37.7	225
	LoRA	91.88	0.7827	1M	24.5	268

To further illustrate the trade-off between resource consumption and model performance, Figure 3a shows the GPU RAM required for training each model with the two different fine-tuning techniques, while Figure 3b presents the model performance (accuracy) for each fine-tuning method and model. It is evident that GPU RAM usage is significantly reduced, yet the decrease in performance compared to FFT is minimal.

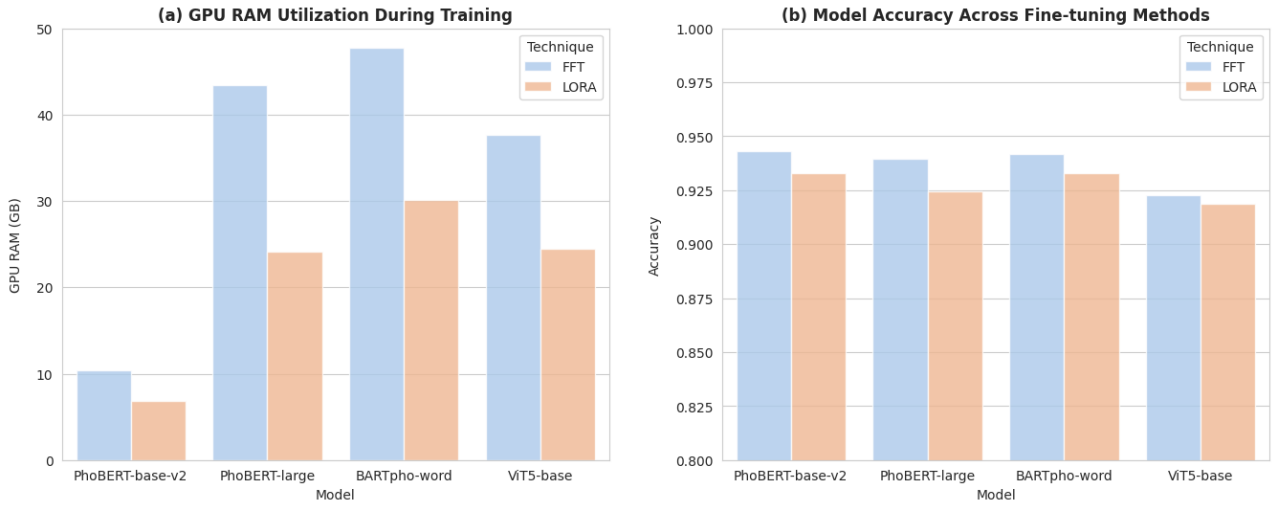

Figure 3 - Comparison of Performance and GPU RAM Usage Across Fine-Tuning Techniques

B. NAMED ENTITY RECOGNITION RESULTS

For the named entity recognition task, only the PhoBERT-base-v2 and PhoBERT-large models were utilized, as NER is a sequence labeling problem that requires token-level outputs. The encoder-decoder architectures of BARTpho and ViT5 are not well-suited for token-level labeling on small, partitioned datasets, whereas PhoBERT (based on RoBERTa) effectively supports methods such as CRF or BiLSTM-CRF for NER tasks. Table 3 summarizes the weighted F1-score results along with the corresponding resource costs for the two fine-tuning techniques.

Table 3 - Comparison of Performance and Resource Costs on the Named Entity Recognition Task

Models	Technique	F1	RAM (GB)	Time (min)
PhoBERT-base-v2	FFT	0.9786	19.2	228
	LoRA	0.9486	12.4	172
PhoBERT-large	FFT	0.9789	46.9	658
	LoRA	0.9212	27.7	488

C. DISCUSSION

Experimental evidence verifies that choosing $r = 16$ and $\alpha = 64$ strikes the proper trade-off between model performance and efficiency in terms of computation. A choice of rank $r = 16$ enables LoRA to capture significant variations in the parameter space effectively without overfitting, consistent with prior work suggesting that medium ranks tend to deliver the optimal trade-offs. Using the scaling factor $\alpha = 64$ assures the LoRA updates are strong enough to impact the baseline model significantly while maintaining training stability. This accords with prior evidence highlighting the significance of paying attention to adjusting the scaling factor to the optimal level. This LoRA setup greatly decreases the number of trainable parameters compared to full fine-tuning—down to 1–2% of the number from the model—yet produces similar performance, according to the aims of parameter-efficient fine-tuning.

The performance–resource trade-off curves clearly demonstrate the advantage of LoRA in scenarios where computational cost optimization is required. However, when resources are not a constraint, FFT continues to deliver the highest and most stable performance, particularly for named entity recognition. In addition, architectural differences among models have a marked impact on the effectiveness of each fine-tuning technique, indicating that the choice of model and method should be tailored to the specific objectives of the intended application.

These analyses not only clarify the trade-offs between efficiency and resource cost for the two fine-tuning approaches, but also provide a solid empirical foundation for drawing overall conclusions and suggest directions for future development, as discussed in the following section.

V. CONCLUSION

This study has focused on evaluating two fine-tuning strategies for large language models – FFT and LoRA in the context of Vietnamese, using two common tasks: sentiment analysis and named entity recognition. By

conducting experiments on several state-of-the-art pre-trained models, including PhoBERT, BARTpho, and ViT5, we provide a comprehensive perspective on the effectiveness, training cost, and practical applicability of each technique under real-world conditions.

The results indicate that LoRA, despite utilizing fewer trainable parameters and requiring significantly less GPU memory compared to FFT, still achieves nearly equivalent performance in terms of accuracy and F1-score. Notably, in resource-constrained settings-which are prevalent in academic labs, educational institutions, or small enterprises-LoRA stands out as a practical choice, enabling broader adoption of large language models without the need for costly infrastructure investment.

The main contribution of this research lies not only in supplying quantitative data on the trade-off between performance and resource cost for FFT and LoRA, but also in establishing LoRA as an effective solution for rapid deployment of Vietnamese models in real-world text analysis applications.

Looking ahead, promising directions include integrating LoRA into lightweight AI product pipelines, extending experiments to specialized domains such as healthcare, legal, or education, and combining this approach with other optimization techniques like quantization or pruning to maximize deployment efficiency on end-user devices.

VI. ACKNOWLEDGEMENT

This research is funded by Ho Chi Minh City University of Foreign Languages - Information Technology under grant number HSV2024-05.

VII. REFERENCES

[1] D. Q. Nguyen and A. Tuan Nguyen (2020). “PhoBERT: Pre-trained language models for Vietnamese,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. doi: 10.18653/v1/2020.findings-emnlp.92.

[2] N. L. Tran, D. Le, and D. Q. Nguyen (2022). “BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese,” in *Interspeech 2022*, ISCA, Sep. 2022, pp. 1751–1755. doi: 10.21437/Interspeech.2022-10177.

[3] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh (2022). “ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, D. Ippolito, L. H. Li, M. L. Pacheco, D. Chen, and N. Xue, Eds., Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 136–142. doi: 10.18653/v1/2022.naacl-srw.18.

[4] N. Ding *et al.* (2023). “Parameter-efficient fine-tuning of large-scale pre-trained language models,” *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023, doi: 10.1038/s42256-023-00626-4.

[5] E. J. Hu *et al.* (2022). “LoRA: Low-Rank Adaptation of Large Language Models,” in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[6] Y. Li, L. Song, and H. Hou (2024). “LoRAN: Improved Low-Rank Adaptation by a Non-Linear Transformation,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 3134–3143. doi: 10.18653/v1/2024.findings-emnlp.177.

[7] C. Zhang, J. Cheng, Y. Xu, and Q. Li (2024). “Parameter-efficient fine-tuning with controls,” in *Proceedings of the 41st International Conference on Machine Learning*, in ICML’24. Vienna, Austria: JMLR.org, 2024.

[8] S. Truong *et al.* (2024). “Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 2849–2900. doi: 10.18653/v1/2024.findings-naacl.182.

[9] V.-T. Doan, Q.-T. Truong, D.-V. Nguyen, V.-T. Nguyen, and T.-N. N. Luu (2023). “Efficient Finetuning Large Language Models For Vietnamese Chatbot,” in *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2023, pp. 1–6. doi: 10.1109/MAPR59823.2023.10288647.

[10] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen (2018). “UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis,” in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Ho Chi Minh City: IEEE, Nov. 2018, pp. 19–24. doi: 10.1109/KSE.2018.8573337.

[11] D. Nguyen *et al.* (2025). “A Hybrid PhoBERT-CNN-LSTM Model for Sentiment Analysis of Vietnamese Student Feedback,” *Asian J. Res. Comput. Sci.*, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:277828326

[12] T. H. Truong, M. H. Dao, and D. Q. Nguyen (2021). “COVID-19 Named Entity Recognition for Vietnamese,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Online: Association for Computational Linguistics, 2021, pp. 2146–2153. doi: 10.18653/v1/2021.naacl-main.173.

[13] Y. Mao, Z. Zhao, S. Ping, Y. Liu, and W. Ding"Enhancing Parameter Efficiency and Generalization in Large Models: A Regularized and Masked Low-Rank Adaptation Approach".

[14] R. Zhang, R. Qiang, S. A. Somayajula, and P. Xie (2024). "AutoLoRA: Automatically Tuning Matrix Ranks in Low-Rank Adaptation Based on Meta Learning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 5048–5060. doi: 10.18653/v1/2024.naacl-long.282.

[15] H. Huang and R. Balestriero (2024). *ALLoRA: Adaptive Learning Rate Mitigates LoRA Fatal Flaws*. 2024. doi: 10.48550/arXiv.2410.09692.

SO SÁNH PHƯƠNG PHÁP TINH CHỈNH MÔ HÌNH NGÔN NGỮ CHO TIẾNG VIỆT: FINE-TUNING TOÀN PHẦN VÀ LORA

Lưu Văn Nhật Hào, Đinh Hùng, Đinh Minh Hòa, Nguyễn Văn Xanh

TÓM TẮT — Các mô hình ngôn ngữ tiền huấn luyện đã mang lại những cải tiến đáng kể cho các bài toán xử lý ngôn ngữ tự nhiên tiếng Việt. Tuy nhiên, việc tinh chỉnh toàn phần (Full Fine-tuning – FFT) các mô hình lớn này đòi hỏi nhiều tài nguyên và gây khó khăn trong các môi trường có hạn chế về tính toán. Bài báo này trình bày một nghiên cứu so sánh giữa FFT và LoRA – một phương pháp tinh chỉnh hiệu quả về tham số – với trọng tâm là sự đánh đổi giữa hiệu năng mô hình và mức tiêu thụ tài nguyên. Thực nghiệm được tiến hành trên hai nhiệm vụ NLP cốt lõi là phân tích cảm xúc và nhận dạng thực thể có tên, sử dụng các bộ dữ liệu chuẩn cùng các mô hình tiền huấn luyện phổ biến như PhoBERT, BARTpho và ViT5. Kết quả cho thấy LoRA đạt độ chính xác tương đương với FFT trong khi giảm đáng kể chi phí huấn luyện, đặc biệt trên các kiến trúc Transformer. Những phát hiện này cho thấy LoRA là một lựa chọn khả thi và hiệu quả cho việc tinh chỉnh PLM tiếng Việt trong bối cảnh hạn chế tài nguyên. Công trình cung cấp những hiểu biết thực nghiệm và bộ kết quả tham chiếu hữu ích nhằm hỗ trợ việc lựa chọn chiến lược tinh chỉnh trong các ứng dụng NLP tiếng Việt.

Từ khóa — Mô hình ngôn ngữ tiếng Việt, Tinh chỉnh mô hình ngôn ngữ, LoRA, Phân tích cảm xúc, Nhận dạng thực thể có tên.

Mr. Nguyen Van Xanh received his M.Eng. in Geographic Information Systems (GIS) from Ho Chi Minh City University of Technology, Vietnam National University, in 2005. He is currently a lecturer at the Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT).

His main research interests include Intelligent Information Systems and Data Visualization.

Dr. Dinh Hung received his M.Sc. in Information Technology from the University of Science (VNU-HCM) in 2003, his M.B.A. from Lincoln University in 2014, and his Ph.D. from Lac Hong University in 2020. He is currently the Head of the E-Commerce Department at Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT).

His primary research interests include E-Commerce, Business Administration, and Information Technology Management.

Mr. Dinh Minh Hoa received his M.Sc. in Information Technology from Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT) in 2022. He is currently a lecturer at the Faculty of Information Technology, HUFLIT, and a member of the university's IDPS research group.

His primary research interests include Artificial Intelligence, Natural Language Processing, and Question Answering Systems.

Luu Van Nhat Hao is currently an undergraduate student majoring in Data Science within the Information Technology program at Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT).