ViDeBERTa: A powerful pre-trained language model for Vietnamese

Cong Dao Tran * [1], Nhut Huy Pham * [1], Anh Nguyen [2], Truong Son Hy † [3], and Tu Vu [4]
FPT Software AI Center [1], Microsoft [2], University of California San Diego [3], University of Massachusetts Amherst [4]
Co-first author *, Correspondent to tshy@ucsd.edu †

Abstract

Although many successful pre-trained language models based on Transformer have been widely proposed for the English language, there are still few pre-trained models for Vietnamese, a low-resource language, that perform good results on downstream tasks, especially Question answering. This paper presents ViDeBERTa, a new pre-trained monolingual language model for Vietnamese, with three versions - ViDeBERTa$_{xsmall}$, ViDeBERTa$_{base}$, and ViDeBERTa$_{large}$, which are pre-trained on a large-scale corpus of high-quality and diverse Vietnamese texts using DeBERTa architecture. We fine-tune and evaluate our model on three important natural language downstream tasks, Part-of-speech tagging, Named-entity recognition, and Question answering. The empirical results demonstrate that ViDeBERTa, with far fewer parameters, surpasses the previous state-of-the-art models on multiple Vietnamese-specific natural language understanding tasks. Notably, ViDeBERTa$_{base}$ with 86M parameters, which is only about 23% of PhoBERT$_{large}$ with 370M parameters, still performs the same or better results than the previous state-of-the-art model. Our ViDeBERTa models are available at: https://github.com/HySonLab/ViDeBERTa.

Motivations and Contributions

Motivations

- Previous pre-trained language models, based on BERT (Devlin et al., 2019) architecture, were trained on relatively small Vietnamese datasets, while PLMs can be significantly improved by using more pre-training data.
- Recently, DeBERTa (He et al., 2020, 2021) architecture using several novel techniques can significantly outperform BERT.
- Question Answering, including Machine Reading Comprehension and Open-domain Question Answering, is an impactful task, but few pre-trained language models for Vietnamese produce efficient results.

Contributions

- We present an improved large-scale pre-trained language model, namely ViDeBERTa, for Vietnamese based on DeBERTa architecture and pre-training techniques.
- We conduct extensive experiments to verify the performance of our pre-trained model compared to previous models in terms of Vietnamese language modeling.
- We release our model as an effective pre-trained model for Vietnamese NLP applications and research.

ViDeBERTa model

How we trained ViDeBERTa?

Pre-training data We use a large corpus CC100 Dataset of 138GB uncompressed texts as a pre-training dataset. We perform word and sentence segmentation using a Vietnamese toolkit PyVi on the pre-training dataset. After that, we use a pre-trained SentencePiece tokenizer from DeBERTaV3 to segment these sentences with sub-word units, which have a vocabulary of 128K sub-word types.

Model architecture Our model follows the DeBERTaV3 (He et al., 2021) architecture, which is trained using the self-supervised learning objectives of MLM and RTD task and a new weight-sharing Gradient-Disentangled Embedding Sharing (GDES) to enhance the performance of the model. We present three versions: ViDeBERTa$_{xsmall}$, ViDeBERTa$_{base}$, and ViDeBERTa$_{large}$ with 22M, 86M, and 304M backbone parameters, respectively.

Optimization We use Adam as the optimizer with weight decay and use a global batch size of 8,192 across 32 A100 GPUs (80GB each) and a peak learning rate of 6e-4 for both ViDeBERTa$_{xsmall}$ and ViDeBERTa$_{base}$, while peak learning rate of 3e-4 was used for ViDeBERTa$_{large}$.

POS tagging and NER tasks

For POS tagging and NER tasks, we use standard benchmarks of the VLSP POS tagging dataset and the PhoNER dataset. A linear layer for prediction is appended on top of our model architecture (the last Transformer layer). We then use Adam to optimize our model for fine-tuning with a fixed learning rate of 1e-5 and batch size of 16.

ID	Form	POS
1	Tôi $_{I}$	PRON
2	là $_{am}$	VERB
3	sinh_viên $_{student}$	NOUN
4	Đại_học $_{university}$	NOUN
5	Công_nghệ $_{technology}$	NOUN

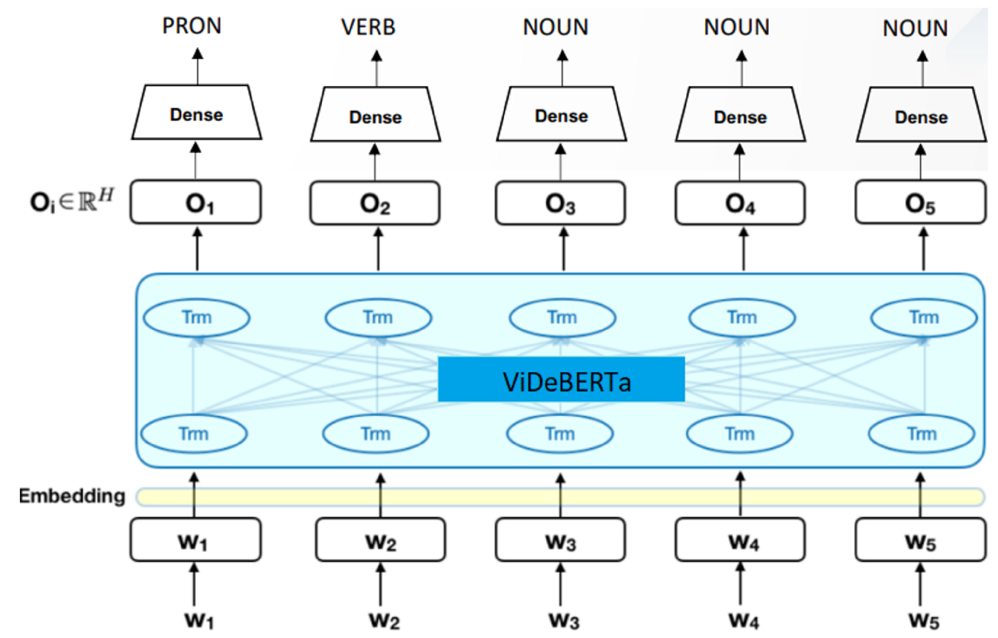

Figure 1: An illustration of the Vietnamese POS tagging task and the fine-tuning of our ViDeBERTa for this task.

Question Answering tasks

We evaluate our model on two main tasks for Question Answering: Machine Reading Comprehension (MRC) and Open-domain Question Answering (ODQA). We use the ViQuAD corpus for assessing these tasks.

Machine reading comprehension

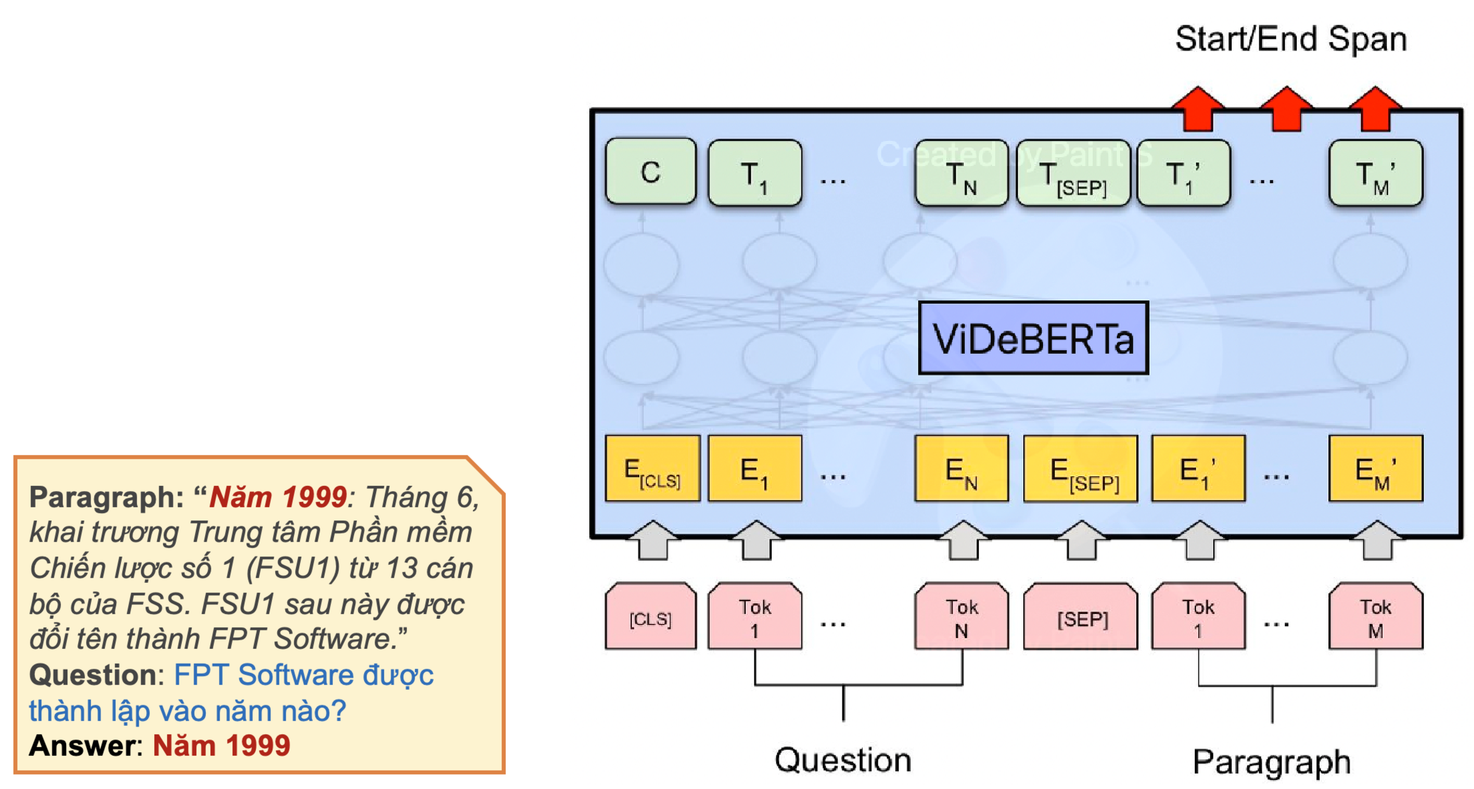

Figure 2: An illustration of the Vietnamese Machine Reading Comprehension task and the fine-tuning of our ViDeBERTa for this task.

Open-domain Question Answering For ODQA, we propose a new framework ViDeBERTa-QA, that uses a BM25 as a retriever and ViDeBERTa as a text reader.

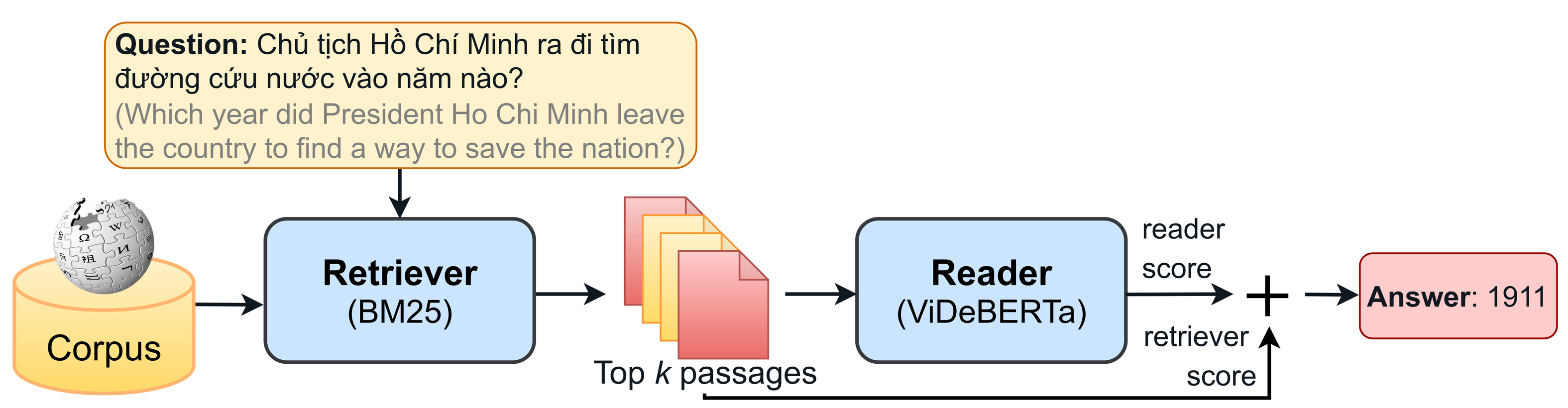

Figure 3: An overview of ViDeBERTa-QA framework for Vietnamese Open-domain Question Answering.

Experimental Results

POS tagging, NER, and MRC tasks: Our model produces significantly better results than the baselines and achieves new SOTA performance on these tasks.

Model	POS	NER	MRC
	Acc.	F_1	F_1
XLM-R$_{base}$	96.2†	–	82.0‡
XLM-R$_{large}$	96.3†	93.8*	87.0‡
PhoBERT$_{base}$	96.7†	94.2*	80.1
PhoBERT$_{large}$	96.8†	94.5*	83.5
ViT5$_{base1024-length}$	–	94.5*	–
ViT5$_{large1024-length}$	–	93.8*	–
ViDeBERTa$_{xsmall}$	96.4	93.6	81.3
ViDeBERTa$_{base}$	96.8	94.5	85.7
ViDeBERTa$_{large}$	**97.2**	**95.3**	**89.9**

ODQA task: ViDeBERTa-QA achieves better scores than the previous models, including BERTsini, DrQA, and SOTA XLMRQA at the top k passages, selected by retrievers, is 10 and 20.

Model	Top k selected passages			
	1	5	10	20
DrQA	37.86	37.86	37.86	37.86
BERTserini	55.55	58.30	57.98	58.09
XLMRQA	**61.83**	**64.99**	64.49	64.49
ViDeBERTa$_{xsmall}$	52.76	56.24	56.93	57.40
ViDeBERTa$_{base}$	58.55	61.37	61.89	62.43
ViDeBERTa$_{large}$	61.23	63.57	**64.89**	**65.34**

Discussion

- ViDeBERTa$_{base}$ (86M) with fewer parameters but still perform slightly better than XLM-R$_{large}$ and competitively the same as the previous SOTA PhoBERT$_{large}$. The possible reasons are: *i) our model inherits the robustness of DeBERTaV3 architecture and pre-training techniques, demonstrating superior performance; ii) using more high-quality pre-training data (138GB) can help ViDeBERTa significantly improve its performance compared to PhoBERT (20GB).*
- ViDeBERTa outperforms PhoBERT by a large margin. Our models are more scalable than PhoBERT for long contexts since *PhoBERT set a maximum length of 256 subword tokens for both versions, while ViDeBERTa set a larger one of 512.*
- The results obtained by ViDeBERTa-QA on ODQA also suggest that our framework achieves the best performance with large top k passages selected by the retriever (i.e., $k = 10, 20$).

Reference

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.