ViPubmedDeBERTa: A Pre-trained Model for Vietnamese Biomedical Text

Manh Tran-Tien, Huu-Loi Le, Dang Nhat Minh, T. Tran Khang,
Huy-The Vu, Minh-Tien Nguyen[*]
Faculty of Information Technology,
Hung Yen University of Technology and Education Hung Yen, Vietnam.
{manhtt.079, lehuuloi.cs, trantrungkhang.cs, dangnhatminh.cs}@gmail.com
{thevh, tienm}@utehy.edu.vn

Abstract

Pre-trained language models have exhibited impressive efficacy within the medical domain. However, their applications to low-resource languages, such as Vietnamese, are still encumbered by the availability of such models. To fill the gap, this paper introduces a new pre-trained language representation model, named **ViPubmedDeBERTa**. The model is pre-trained by using 20 million high-quality Vietnamese biomedical abstracts translated from PubMed based on the weight of ViDeBERTa. The model is available with two versions: ViPubmedDeBERTa$_{\text{xsmall}}$ with 22 million parameters and ViPubmedDeBERTa$_{\text{base}}$ with 86 million parameters. The performance of ViPubmedDeBERTa is evaluated on five benchmark datasets with three biomedical tasks: text classification, named entity recognition, and natural language inference. Experimental results highlight that, despite having significantly fewer parameters, ViPubmedDeBERTa achieves promising results compared to recent state-of-the-art pre-trained models in Vietnamese. The ViPubmedDeBERTa model is available at: https://github.com/manhtt-079/vipubmed-deberta

1 Introduction

Representation learning plays an important role in Natural Language Processing (NLP) (Bengio et al., 2013; Liu et al., 2023). This is because it is the first step for mapping input sequences to intermediate representation for NLP tasks. The learning process of NLP models is beneficial from good representation while poor representation leads to low accuracy. Prior representation learning approaches include non-neural (Ando et al., 2005; Blitzer et al., 2006) to neural methods (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2017). The emergence of Transformer (Vaswani et al., 2017) boosts the development of pre-trained models (PLMs). BERT (Devlin et al., 2019) is perhaps the first pre-trained model that can be adapted for a wide range of NLP tasks. The training of BERT opens a new direction for pre-training representation models by using self-learning, which allows the training process to use a massive amount of data without human annotation. After pre-training, by using the transfer learning technique, BERT has shown promising results on a wide range of NLP downstream tasks in the GLUE benchmark (Wang et al., 2018). The success of BERT leverages the investigation of PLMs in both English (Radford and Narasimhan, 2018; Liu et al., 2019) or other languages, e.g., Vietnamese (Nguyen and Nguyen, 2020; Tran et al., 2023; Minh et al., 2022). Recently, generative pre-trained large language models (LLMs) have shown the paradigm shift of NLP in which all tasks can be formulated as text generation (Radford and Narasimhan, 2018; Brown et al., 2020; Gururangan et al., 2020; Du et al., 2021).

Representation learning has also received attention in the medical domain. The methods range from publishing medical embeddings (Sharma and Daniel Jr, 2019; Alsentzer et al., 2019) to releasing PLMs for the medical domain (Huang et al., 2019; Lee et al., 2020; Gu et al., 2021; raj Kanakarajan et al., 2021; Yasunaga et al., 2022). The pre-training mainly adopts two main methods: (i) using the masked language model (Huang et al., 2019; Lee et al., 2020; Yasunaga et al., 2022) and (ii) using the generator-discriminator mechanism (raj Kanakarajan et al., 2021). The pre-training either utilizes the weight from PLMs for common domains (Huang et al., 2019; Lee et al., 2020; Gu et al., 2021; Yasunaga et al., 2022) or trains PLMs from scratch by using domain-specific corpora (raj Kanakarajan et al., 2021). PLMs have shown promising results for medical downstream tasks. For Vietnamese, there exist several PLMs created for specific tasks. Table 1 summarizes the information of PLMs for Vietnamese. As we can observe,

[*] Corresponding Author.

Table 1: Pre-trained models for Vietnamese. *Rep* is representation and *gen* stands for generation.

PLMs	#params	Max length	Domain	Task
PhoBERT$_{\text{base}}$ (Nguyen and Nguyen, 2020)	135M	256	Wiki, news	rep
PhoBERT$_{\text{large}}$ (Nguyen and Nguyen, 2020)	370M	256	Wiki, news	rep
ViHealthBERT$_{\text{base-word}}$ (Minh et al., 2022)	135M	256	Healthcare	rep
ViDeBERTa$_{\text{base}}$ (Tran et al., 2023)	86M	512	Wiki, News	rep
ViT5$_{\text{base}}$ (Phan et al., 2022b)	220M	1024	Wiki	gen
ViT5$_{\text{large}}$ (Phan et al., 2022b)	866M	1024	Wiki	gen
BARTPho$_{\text{base-word}}$ (Tran et al., 2021)	150M	1024	Wiki	gen
BARTPho$_{\text{large-word}}$ (Tran et al., 2021)	420M	1024	Wiki	gen
ViPubmedT5 (Phan et al., 2023)	220M	1024	Biomedical	gen
ViPubmedDeBERTa$_{\text{xsmall}}$ (Ours)	22M	512	Biomedical	rep
ViPubmedDeBERTa$_{\text{base}}$ (Ours)	86M	512	Biomedical	rep

PLMs are designed for two main tasks: representation and text generation. Among contextual PLMs, only ViHealthBERT (Minh et al., 2022) was pre-trained for healthcare and medical tasks. Other PLMs of representation were designed for common domains that may not be inappropriate for medical tasks. For generation PLMs, only ViPubmedT5 (Phan et al., 2023) was trained for the biomedical domain but it is appropriate for text-generation tasks. In practical applications, the adaptation of PLMs from general domains to the medical domain is feasible. However, a knowledge gap exists among different domains, which may constrain the performance of PLMs in downstream tasks (Huang et al., 2019; Lee et al., 2020; Gu et al., 2021; raj Kanakarajan et al., 2021; Yasunaga et al., 2022; Minh et al., 2022; Phan et al., 2023). Our research emphasizes the significance of focusing on Vietnamese biomedical texts, as it addresses a growing demand within a rapidly evolving field. The current language models available for this domain are often quite limited (only ViHealthBERT (Minh et al., 2022) and ViPubmedT5 (Phan et al., 2023)), and our study showcases the potential of leveraging a large scale translated biomedical dataset to improve the quality of PLMs for the biomedical domain. We posit that the utilization of PLMs specifically designed for biomedical Vietnamese text can yield improved performance in biomedical tasks.

This paper introduces a new PLM designed for the medical domain in Vietnamese. To do that, the model is continuously trained from the ViDeBERTa model (Tran et al., 2023) which is pre-trained by using Wikipedia and news data. To adapt the model to the medical domain, we use 20 million high-quality abstracts of translated medical text from English to Vietnamese (Phan et al., 2023). The original ViDeBERTa model is continuously pre-trained by using the Vietnamese medical data using the masked language model and a weight-sharing mechanism. This results two new pre-trained models: ViPubmedDeBERTa$_{\text{xsmall}}$ and ViPubmedDeBERTa$_{\text{base}}$. Experimental results on Vietnamese medical tasks show the efficiency of the PLM. This paper makes two main contributions.

- It introduces a new PLM for the medical domain in Vietnamese. The model includes the small and base versions. The new PLM enriches representation models for Vietnamese and leverages the investigation of medical tasks in Vietnamese. The code of pre-training the model is also accessible.

- It validates the efficiency of the model by comparing it to strong PLMs in Vietnamese. Experimental results on five benchmark datasets show that the PLM achieves promising results compared to other state-of-the-art PLMs on a wide range of medical-related tasks.

2 Related Work

Pre-trained models In the early stage of pre-trained language models, CBOW/Skip-Gram (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) were proposed to produce word representation, which can capture the semantic meanings of words. However, these models are context-free and fail to capture higher-level concepts in context. To deal with this problem, contextual pre-trained models were introduced such as CoVe (McCann et al., 2017) and ELMo (Peters et al., 2018). After the

Transformer (Vaswani et al., 2017) was introduced, there have been many architectures were proposed for example, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019). They were trained with different language modeling types such as the masked language model (Devlin et al., 2019), permuted language modeling (Mohamad Zamani et al., 2022), denoising (Lewis et al., 2019), and contrastive learning (Clark et al., 2019). These transformer-based models have been shown as powerful tools for contextual representation.

Pre-trained models for medical text Apart from models pre-trained on the general domain, representation learning has recently received attention in the medical domain. The models proposed for either producing medical medical embeddings (Sharma and Daniel Jr, 2019; Alsentzer et al., 2019) or releasing PLMs for medical text (Huang et al., 2019; Lee et al., 2020; Gu et al., 2021; raj Kanakarajan et al., 2021; Yasunaga et al., 2022). These models often adopt two language modeling methods including the masked language model (Huang et al., 2019; Lee et al., 2020; Yasunaga et al., 2022) and contrastive learning (raj Kanakarajan et al., 2021). Since they are pre-trained models for specific domains, continuous pre-training from trained models in general domains is widely used (Huang et al., 2019; Lee et al., 2020; Gu et al., 2021; Yasunaga et al., 2022). In addition, some of them were trained from scratch using domain-specific corpora (raj Kanakarajan et al., 2021) with promising results.

Pre-trained models for Vietnamese text There have been several models that were pre-trained for Vietnamese (Nguyen and Nguyen, 2020; Tran et al., 2021; Minh et al., 2022; Phan et al., 2022b). PhoBERT (Nguyen and Nguyen, 2020) is the first attempt and archives remarkable results on several downstream tasks. Recently, it has been updated to the second version trained with more training data.[1] While most of them focus on the general domain, ViHealthBERT (Minh et al., 2022) and ViPubmedT5 (Phan et al., 2022b) were created for working with biomedical text. ViHealthBERT was pre-trained on over 130M sentences covering health, medical, and general domains. To deal with the issue of lacking biomedical text data in Vietnamese - a low-resource language, ViPubmedT5 (Phan et al., 2022b) leveraged the state-of-the-art translation architectures to build the ViPubmed dataset translated from 20M English PubMed abstracts. This dataset was then used to continuously pre-train ViT5 (Phan et al., 2022b). Inspired by this work, we introduce ViPubmedDeBERTa. Contrary to ViPubmedT5 trained for text generation tasks, our model is trained for representation learning tasks. It relies on ViDeBERTa (Tran et al., 2023) and is trained using the masked language model and weight-sharing mechanisms. This allows the model to produce more fine-grained representations of medical documents in Vietnamese.

3 Pre-traing ViPubmedDeBERTa

This section describes the architecture, pre-training data, training strategy, and optimization setup for ViPubmedDeBERTa.

Data The ViPubmedDeBERTa was pre-trained with two main data sources. The first data source implicitly comes from the 138GB CC100 dataset as we continuously pre-trained the previous ViDeBERTa model (Tran et al., 2023). The second data source consists of 20 million Vietnamese biomedical abstracts obtained via a state-of-the-art large-scale translation system for English-Vietnamese translation in the biomedical domain (Phan et al., 2022b), based on the PubMed[2] dataset. PubMed is a high-quality dataset of biomedical research, including up-to-date information of COVID-19 diseases and vaccines, collected from sources such as life science publications, medical journals, and published online e-books. The 20 million translated abstracts were used to pre-train ViPubmedT5 (Phan et al., 2022b), an encoder-decoder model, which has demonstrated a low level of bias and errors in its translation outputs. This model has achieved state-of-the-art results in medical and clinical contexts, demonstrating its effectiveness in translating biomedical knowledge into Vietnamese. A fuzzy de-duplication, targeting documents with high overlap, was conducted at the document level to enhance the quality of the second data source. To do that, we employed locality-sensitive hashing with a threshold of 0.9 to remove abstracts that overlap over 90%. This process results in the average reduction of the dataset's size by 3%. We used Pyvi[3] - a Vietnamese toolkit to perform word segmentation on the pre-training dataset (the second data source) for compatibility with our tokenizer.

[1] https://github.com/VinAIResearch/PhoBERT

[2] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline

[3] https://pypi.org/project/pyvi/

Model architecture The architecture of ViPubmedDeBERTa relies on ViDeBERTa (Tran et al., 2023). ViDeBERTa, a recently developed pre-trained monolingual language model for the Vietnamese language, leverages the extensive CC100 dataset, which composes 138GB of uncompressed texts derived from web crawls of monolingual data sources (Conneau et al., 2019). ViDeBERTa is built upon the architecture of DeBERTaV3 (He et al., 2021). The model undergoes training using self-supervised learning objectives, specifically Masked Language Modeling (MLM) and Relation-aware Token Discrimination (RTD), aiming to optimize its performance. Furthermore, we employ Gradient Disentangled Embedding Sharing (GDES) to further enhance the overall efficacy of the model. The ViDeBERTa was selected because it is the smallest PLM (86M parameters) (Table 1) but achieves promising results for NLP-related tasks (Tran et al., 2023). The small number of parameters helps to reduce the request for heavy computing resources and speed up the pre-training and fine-tuning processes.

To enhance the model's performance, we fine-tuned ViDeBERTa using a high-quality corpus of biomedical texts by using the MLM. This is because MLM is appropriate for representing the meaning of tokens based on their neighbors. We did not employ the next sentence prediction as BERT (Devlin et al., 2019) because it does not show improvements for contextual representation (Liu et al., 2019). To improve both training efficiency and the quality of the pre-trained model, the pre-training process uses a weight-sharing mechanism called Gradient-Disentangled Embedding Sharing (Clark et al., 2019). After pre-training, ViPubmedDeBERTa exists two versions: ViPubmedDeBERTa$_{\text{xsmall}}$ with 22M parameters and ViPubmedDeBERTa$_{\text{base}}$ with 86M parameters, respectively. It facilitates the fine-tuning of the model into downstream tasks with two options for computing resources. Our experimental results show that a vocabulary size of 128k covers almost all subwords, so we leverage the pre-trained token weights of the ViDeBERTa model and do not retrain the new tokenizer. This tokenizer integrates word and sentence segmentation, employing the Vietnamese toolkit PyVi. Additionally, it is pre-trained with a SentencePiece tokenizer (Kudo and Richardson, 2018) to facilitate the segmentation of sentences into sub-word units. This tokenizer results in a vocabulary of 128K sub-word types, which enhances the diversity of representation.

Pre-training We used the last checkpoint of ViDeBERTa (Tran et al., 2023) and continuously pre-trained our ViPubmedDeBERTa model from the checkpoint with 20 million medical abstracts from ViPubmed (Phan et al., 2022a). Our model was trained on a single A100 GPU (40GB), with a batch size of 24 and a gradient accumulation step of 4 (resulting in a global batch size of 96). The ViPubmedDeBERTa$_{\text{xsmall}}$ version was pre-trained in 220k steps in 5 days and the ViPubmedDeBERTa$_{\text{base}}$ version was pre-trained in 400k steps in 9 days with the model's peak learning rate of 1e-4. To balance the representation and training speed, the maximum sequence length was fixed to 512 tokens. To evaluate the model's performance during pre-training, we regularly assess its efficacy on the validation set at specified training steps. Early stopping is employed if the validation loss does not decrease after five evaluations.

4 Downstream Medical Datasets

This section describes benchmark medical datasets for fine-tuning the pre-trained ViPubmedDeBERTa model. The datasets include three medical tasks, namely, text classification, Named Entity Recognition (NER), and Natural Language Inference (NLI). Table 2 summarizes the information of the datasets.

Table 2: Statistics of the benchmark datasets. *"IC"* means intent classification in the ViMQ dataset.

Datasets	#train	#val.	#test	#class/ner
acrDrAid	4,000	523	1,030	2
ViMQ IC	7,000	1,000	1,000	4
COVID-19 NER	5,027	2,000	3,000	10
ViMQ NER	7,000	1,000	1,000	3
ViMedNLI	11,232	1,395	1,422	2

4.1 Classification

The classification task includes two smaller tasks: binary classification with the **acrDrAid** dataset (Minh et al., 2022) and multiclass classification with the **ViMQ** dataset (Huy et al., 2023).

Binary classification The binary classification task uses **acrDrAid** (Minh et al., 2022) - a dataset designed for Vietnamese Acronym Disambiguation (AD). It contains radiology reports obtained from the Vinmec hospital[4] in Vietnam. The primary objective of the dataset is to ascertain the accurate identification of acronym expansion within the context of the provided radiology reports. The dataset includes 135 acronyms and 424 expansion texts annotated by three radiologist experts.

Multiclass classification The task of multiclass classification uses **ViMQ** (Huy et al., 2023) - a Vietnamese medical questions dataset crawled from online consultation sections between patients and doctors from `www.vinmec.com`. The ViMQ includes 9,000 samples and is annotated for two tasks: Intent Classification (IC) and Name Entity Recognition (NER). The IC problem includes four labels: Diagnosis, Severity, Treatment, and Cause. The Diagnosis consists of questions relating to the identification of symptoms or diseases. The Severity contains questions relating to the conditions or grade of an illness. The Treatment composes questions relating to the medical procedures for an illness. The Cause includes questions relating to factors of a symptom or disease. The NER set is used for the NER task presented in the next section.

4.2 Named Entity Recognition

COVID-19 NER (Truong et al., 2021) is the first manually annotated domain-specific dataset in Vietnamese. This dataset contains 10 entity types with the aim of extracting key information related to COVID-19 patients. It includes a total of 10,027 samples that are divided into train/valid/test samples with the amount of 5,027/200/3,000, respectively. We used the word-level version as the same as PhoBERT (Nguyen and Nguyen, 2020).

ViMQ NER is the NER part of the ViQM dataset (Huy et al., 2023). This set consists of three entity categories: symptom-disease, medicine, and medical procedure with 13,253, 2,000, and 979 NERs, respectively. The set is split into train/dev/test with the number of samples as 7,000/1,000/1,000 and only has the word-level version.

4.3 Natural Language Inference

ViMedNLI (Phan et al., 2022a) is a Vietnamese medical natural language inference dataset (ViMedNLI) that was translated from MedNLI (Romanov and Shivade, 2018) and refined with biomedical experts. Given a premise sentence and a hypothesis sentence, the relation of two sentences falls into one of the labels: entailment, contradiction, and neutral.

5 Fine-tuning ViPubmedDeBERTa for Vietnamese Medical Tasks

This section describes the fine-tuning process of ViPubmedDeBERTa for Vietnamese medical downstream tasks. To balance the quality and the efficiency of the fine-tuning process, the maximum sequence length of 512 and executed reword-segmentation were used for all tasks.

5.1 Classification

As mentioned, the classification task contains two smaller tasks: binary classification and multiclass classification. The fine-tuning process of classification is described in the next sections.

Binary classification The binary classification is to classify an acronym text into one of two classes: acronym or non-acronym. Based on the provided code by (Minh et al., 2022) for the **acrDrAid** dataset, we employed transfer learning to optimize the effectiveness of our models. In particular, the input sequence was presented as a single sentence, wherein the final hidden vector denoted as $C \in \mathbb{R}^H$ associated with the first input token `[CLS]` is selected. This vector is then concatenated with the mean representation of acronym tokens in the sentence represented by $A \in \mathbb{R}^{L \times H}$, which serves as the comprehensive sentence representation. The new components introduced during this process are two linear layers, each equipped with a weight matrix: $W_1 \in \mathbb{R}^{2H \times K}$ and $W_2 \in \mathbb{R}^{K \times 1}$, respectively. Where H is the hidden model dimension, L is the length of the acronym tokenized text, and K is the intermediate dimension. The outputs were computed with a standard classification loss to predict whether the acronym text matched the label or not. The model was fine-tuned in 100 epochs with a learning rate of 3e-5, a batch size of 32, and a fixed random seed. To assess the model's performance during the fine-tuning phase, we evaluated its effectiveness at fixed training steps on the validation set by using the Macro-F1 score as the early stopping criteria. The best model checkpoint was then evaluated on the test set to report the final score.

Multiclass classification The pre-trained ViPubmedDeBERTa model was also fine-tuned for the multiclass classification task on the intent classification part of **ViMQ** (Huy et al., 2023). The input is

[4] https://vinmec.com/

a sequence and the model has to predict the classes of the input. To obtain an aggregate representation, we utilized the first or last hidden vector denoted as $C \in \mathbb{R}^{H}$ extracted from the pre-trained model, specifically the vector associated with the special token `[CLS]`. Additionally, we introduced an architectural modification by incorporating an extra linear layer on top of the pre-trained model. This linear layer is parameterized by a weight matrix $W \in \mathbb{R}^{H \times K}$, where H is the hidden dimension of the model and K represents the number of labels involved in the classification task. Subsequently, we proceed with the fine-tuning of our models. The fine-tuning process was employed in 15 epochs with a learning rate of 3e-5 and a batch size of 16.

5.2 Named Entity Recognition

The pre-trained ViPubmedDeBERTa model was fine-tuned for the medical NER task. The fine-tuning process uses two datasets: **COVID-19** (Truong et al., 2021) and **ViMQ NER** (Huy et al., 2023) (the NER part of VIMQ). For fine-tuning, an additional linear layer was stacked on top of ViPubmedDeBERTa with a weight matrix of $W \in \mathbb{R}^{H \times N}$, where H is the hidden model dimension and N is the number of NER types. The input sequence was structured as a single sentence and the model utilized the final hidden state vectors $C \in \mathbb{R}^{L \times H}$ (L is the length of the sequence), which serves as a comprehensive representation for each token of the input sentence. The matrix C was then fed into the aforementioned linear layer W for the sequence labeling of each input token.

The fine-tuning process employed the cross-entropy loss the measure the difference between predicted tags and gold labels. The fine-tuning process was conducted in 10 epochs with a learning rate of 2e-5 and a batch size of 16.

5.3 Natural Language Inference

The pre-trained ViPubmedDeBERTa model was also fine-tuned for the NLI task on the **ViMedNLI** dataset (Phan et al., 2022a). To do that, we formulated the input as a sentence pair comprising a premise (the first sentence) concatenated with a hypothesis (the second sentence) by using the special token `[SEP]`. The pair was then passed through the pre-trained model to extract features by using the representation of the `[CLS]` token. For classification, we introduced a new linear layer on top of the pre-trained model, with a weight matrix $W \in \mathbb{R}^{H \times 1}$, where H denotes the hidden dimension of the pre-trained model. The fine-tuning process was done in 20 epochs with a learning rate of 2e-5 and a batch size of 16.

6 Results and Discussion

6.1 Experimental Results

This section shows the comparison of our fine-tuned model to state-of-the-art (SOTA) PLMs in Vietnamese. The SOTA PLMs are as follows. **PhoBERT** (Nguyen and Nguyen, 2020) is a PLM trained by Wikipedia and news for a wide range of NLP tasks in Vietnamese. PhoBERT includes small and base versions. **ViHealthBERT** (Minh et al., 2022) is a PLM designed for medical tasks in Vietnamese. It has a base version on the word level. **ViDeBERTa** (Tran et al., 2023) is another PLM for NLP tasks in Vietnamese for general domains. It includes a base version. To ensure a fair comparison, we removed the additional layers (e.g., CRF for NER) used to stack the three PLMs and re-ran them on three medical tasks: classification, NER, and NLI. We did not compare our model to ViT5 (Phan et al., 2022b) or BARTPho (Tran et al., 2021) due to our different purpose in representation learning. The evaluation metrics of each task were followed from original studies (Nguyen and Nguyen, 2020; Minh et al., 2022; Tran et al., 2023; Phan et al., 2023).

Classification Tables 3 and 4 show the performance of PLMs for binary and multiclass classification. The results show two important points. First, our PLM is significantly better than ViDeBERTa (Tran et al., 2023), the base model used to train our PLM. It confirms the contribution of domain adaptation with medical data. By fine-tuning ViDeBERTa with medical data, the proposed model can adapt well to the medical domain and shows promising results for the classification downstream task. Second, for binary classification on the **acrDrAid** dataset, the proposed ViPubmedDeBERTa model consistently outputs better performance than other PLMs across all evaluation metrics, achieving an improvement of +1.78%, +4.56%, and +3.47% in terms of Macro-Precision, Macro-Recall, and Macro-F1 respectively, in comparison to the previous state-of-the-art model ViHealthBERT (Minh et al., 2022). For multiclass classification on the **ViMQ** dataset, the results in Table 4 share the same trend as Table 3. The proposed PLM is the best followed by ViHealthBERT.

Table 3: The performance of binary classification on **acrDrAid**. **Mac** is the Macro metric. **P** is precision and **R** means recall. **Bold** values are the best and underline values are the second best used for all tables.

Model	Mac-P	Mac-R	Mac-F1
PhoBERT$_{\text{base}}$	91.97	74.81	82.51
PhoBERT$_{\text{large}}$	91.50	70.38	79.56
ViHealthBERT$_{\text{base}}$	93.20	75.62	83.49
ViDeBERTa$_{\text{base}}$	87.38	62.74	73.04
Ours$_{\text{xsmall}}$	**94.98**	**80.18**	**86.96**
Ours$_{\text{base}}$	92.67	78.92	85.24

Table 4: The performance of multiclass classification on **VIMQ**. **Acc**: accuracy, **Mic**: micro, and **Mac**: macro.

Model	Acc	Mic-F1	Mac-F1
PhoBERT$_{\text{base}}$	89.30	89.30	90.33
PhoBERT$_{\text{large}}$	89.30	89.30	90.33
ViHealthBERT$_{\text{base}}$	90.10	90.10	90.85
ViDeBERTa$_{\text{base}}$	86.00	86.00	86.46
Ours$_{\text{xsmall}}$	**90.60**	**90.60**	**91.26**
Ours$_{\text{base}}$	90.50	90.50	91.19

Other PLMs also achieve promising results. Interestingly, the small version of our model is competitive with the base model even though the number of parameters is much smaller (22M vs. 86M). It shows that with appropriate data for fine-tuning, small PLMs can achieve promising results (similar to PhoBERT (Nguyen and Nguyen, 2020)).

NER Table 5 reports the comparison of the proposed model and strong PLMs for NER in Vietnamese. Similar to classification, we can observe that the proposed model is better than the base model, ViDeBERTa, especially on the NER part of ViMQ. For the **COVID-19** dataset, the results show that ViPubmedDeBERTa (the base version) achieves improvements for all metrics compared to other PLMs. The small version of ViPubmedDeBERTa is also competitive with BERT$_{\text{large}}$ even though this version has the smallest number of parameters (22M vs. 370M of BERT$_{\text{large}}$). ViHealthBERT also obtains promising results which are very close to the scores of our model.

For the NER part of **ViQM**, the trend is similar to the COVID-19 dataset, in which the base and small versions of our PLM obtain the best and second-best results. ViHealthBERT (Minh et al., 2022) follows our two versions with tiny margins. It shows that ViHealthBERT is a strong PLM for Vietnamese medical tasks.

Table 5: The NER performance on the **COVID-19** and **ViMQ** datasets.

Model	Acc	Mic-F1	Mac-F1
	COVID-19		
PhoBERT$_{\text{base}}$	97.59	93.26	92.01
PhoBERT$_{\text{large}}$	97.75	93.79	92.69
ViHealthBERT$_{\text{base}}$	98.21	93.95	92.85
ViDeBERTa$_{\text{base}}$	97.62	89.14	85.82
Ours$_{\text{xsmall}}$	98.81	94.62	92.98
Ours$_{\text{base}}$	**98.83**	**94.76**	**93.18**
	ViMQ-NER		
PhoBERT$_{\text{base}}$	91.67	78.18	76.91
PhoBERT$_{\text{large}}$	91.76	78.21	74.73
ViHealthBERT$_{\text{base}}$	91.93	78.26	76.25
ViDeBERTa$_{\text{base}}$	88.33	69.25	61.47
Ours$_{\text{xsmall}}$	91.99	80.08	77.34
Ours$_{\text{base}}$	**92.04**	**80.65**	**77.83**

NLI Table 6 shows the comparison of NLI between our model and other PLMs. We can observe that ViHealthBERT (Minh et al., 2022) is the best for all metrics. Our model follows ViHealthBERT with tiny margins. In addition, our models are still better than other strong PLMs such as PhoBERT and ViDeBERTa. Note that ViPubmedT5 (Phan et al., 2023) only reports an accuracy of 81.65% which is the best. However, we could not access the code for reporting other metrics. So the accuracy of ViPubmedT5 was not put into Table 6 to keep the consistency of the evaluation.

Table 6: The performance on the **ViMedNLI** dataset.

Model	Acc	Mic-F1	Mac-F1
PhoBERT$_{\text{base}}$	77.29	77.24	77.24
PhoBERT$_{\text{large}}$	77.49	77.49	77.50
ViHealthBERT$_{\text{base}}$	**78.19**	**78.19**	**78.2**
ViDeBERTa$_{\text{base}}$	66.80	66.80	66.92
Ours$_{\text{xsmall}}$	77.77	77.77	77.75
Ours$_{\text{base}}$	77.71	77.66	77.66

6.2 Discussion

The results of the three tasks shown in Tables 3, 4, 5, and 6 indicate two important points. First, our model is consistently better than ViDeBERTa, the base version used to train our models. It confirms the efficiency of the pre-training and fine-tuning processes in Sections 3 and 5. By using a massive high-quality amount of medical data, our models can adapt well to the medical domain. Second, our small version outputs competitive results compared to the base version, even though the number of parameters of the small version is much smaller than that of the base version (22M vs. 86M). A possible reason is that the base version needs more training steps for convergence. In addition, the final Transformer layer of the base version may not be optimized to capture the comprehensive syntactic structural information, as suggested by previous studies (Hewitt and Manning, 2019; Jawahar et al., 2019). In general, the two versions achieve promising results compared to SOTA PLMs in Vietnamese.

The results also show that the performance of the two versions (22M and 86M parameters) is better than that of PhoBERT$_{\text{large}}$ (with 370M parameters) and ViHealthBERT$_{\text{base-word}}$ (with 135M parameters) despite having a much smaller number of parameters. It shows that with appropriate data and pre-training strategies, a small PLM can be comparable with large PLMs. In practice, it helps to reduce the request for heavy computing resources for pre-training. For other PLMs with a similar number of parameters, i.e., ViDeBERTa (86M parameters), our pre-trained model is much better in terms of performance. Among the three tasks, ViPubmedDeBERTa is the best for classification and NER while it is comparable to ViHealthBERT for the NLI task. For classification and NER tasks, a possible reason is that ViPubmedDeBERTa effectively utilizes the robustness inherited from ViDeBERTa, a model based on the DeBERTaV3 architecture trained on a large-scale, high-quality Vietnamese dataset (138GB) (He et al., 2021). Additionally, the pre-training of ViPubmedDeBERTa on 20M Vietnamese biomedical abstracts, acquired through large-scale translation and refined by domain experts, contributes to improving the contextual representation of input tokens. It facilitates the fine-tuning of ViPubmedDeBERTa for downstream classification and NER tasks.

For the NLI task, the average length of sequences in ViMedNLI is approximately 128 tokens which is quite short compared to classification and NER corpora. These 128 tokens are far from the maximum length of 512 designed for ViPubmedDeBERTa. It may reduce the context representation of input tokens. In contrast, ViHealthBERT was pre-trained with a maximum length of 256 which seems to be more appropriate for the ViMedNLI than our model. However, the gap between our model and ViHealthBERT for NLI is tiny even though the number of parameters of our model is much smaller than that of ViHealthBERT.

7 Conclusion

This paper introduces ViPubmedDeBERTa, a pre-trained language representation model specifically designed for Vietnamese biomedical tasks. The model was continuously trained by using ViDeBERTa with 20 million high-quality Vietnamese biomedical abstracts translated from PubMed. Experimental results on three biomedical tasks show three important points. First, ViPubmedDeBERTa is consistently better than ViDeBERTa, the base model used in the pre-training step. It confirms the contribution of biomedical data that helps to adapt the model to the medical domain. Second, ViPubmedDeBERTa obtains promising results compared to recent SOTA PLMs in Vietnamese even though our model is much smaller in terms of parameters. Finally, ViPubmedDeBERTa faces challenges with short input sequences. By publishing ViPubmedDeBERTa as a robust pre-trained model, we aim to facilitate future research and applications of NLP tasks in the healthcare and medical domains.

Although we have demonstrated ViPubmedDeBERTa's state-of-the-art performance across various NLP tasks for the Vietnamese medical domain, further analyses and ablation studies are required to gain a comprehensive understanding of the behavior of the model. The model needs further observation to show how good ViPubmedDeBERTa is to capture Vietnamese linguistic knowledge. In addition, we will continue to improve the quality of the base model by training with additional steps and plan to test ViPubmedDeBERTa for other NLP tasks in more diverse genres and scenarios. We leave these in-depth investigations as future work.

Ethics Statement

The data and models used in this study have no unethical applications or risky broader impacts. The data was widely used in the research community. It does not include any confidential or personal information of workers or companies. The models were accessed by using the provided GitHub links. There is no bias for the re-implementation that can affect the final results.

Acknowledgements

Computing resources are sponsored by Intelligent Integration Co., Ltd. (INT2), Vietnam.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11).

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the 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. *ArXiv*, abs/1810.04805.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Ta Duc Huy, Nguyen Anh Tu, Tran Hoang Vu, Nguyen Phuc Minh, Nguyen Phan, Trung H. Bui, and Steven Quoc Hung Truong. 2023. Vimq: A vietnamese medical question dataset for healthcare dialogue system development. *ArXiv*, abs/2304.14405.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Annual Meeting of the Association for Computational Linguistics*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *ArXiv*, abs/1808.06226.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

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

Yixin Liu, Yizhen Zheng, Daokun Zhang, Vincent CS Lee, and Shirui Pan. 2023. Beyond smoothing: Unsupervised graph representation learning with edge heterophily discriminating. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4516–4524.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NIPS*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. Vihealthbert: Pre-trained language models for vietnamese in health text mining. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 328–337.

Nur Azmina Mohamad Zamani, Jasy Suet Yan Liew, and Ahmad Muhyiddin Yusof. 2022. XLNET-GRU sentiment regression model for cryptocurrency news in English and Malay. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 36–42, Marseille, France. European Language Resources Association.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Long Phan, Tai Dang, Hieu Tran, Trieu Trinh, Vy Phan, Lam Chau, and Minh-Thang Luong. 2023. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3123–3134.

Long Phan, Tai Dang, Hieu Trung Tran, Trieu H. Trinh, Vy Phan, Lam D. Chau, and Minh-Thang Luong. 2022a. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H Trinh. 2022b. 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*, pages 136–142.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Alexey Romanov and Chaitanya P. Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Conference on Empirical Methods in Natural Language Processing*.

Shreyas Sharma and Ron Daniel Jr. 2019. Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. *arXiv preprint arXiv:1908.05760*.

Cong Dao Tran, Nhut Huy Pham, Anh-Việt Nguyên, Truong Son Hy, and Tu Vu. 2023. Videberta: A powerful pre-trained language model for vietnamese. *ArXiv*, abs/2301.10439.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: pre-trained sequence-to-sequence models for vietnamese. *arXiv preprint arXiv:2109.09701*.

Thinh Hung Truong, M. Dao, and Dat Quoc Nguyen. 2021. Covid-19 named entity recognition for vietnamese. In *North American Chapter of the Association for Computational Linguistics*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.