PHOGPT:
GENERATIVE PRE-TRAINING FOR VIETNAMESE

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, Hung Bui
VinAI Research, Hanoi, Vietnam
`{v.datnq9, v.linhnt140, v.chitb, v.dungnn28, v.dinhpq2, v.hungbh1}@vinai.io`

ABSTRACT

We open-source a state-of-the-art 4B-parameter generative model series for Vietnamese, which includes the base pre-trained monolingual model PhoGPT-4B and its chat variant, PhoGPT-4B-Chat. The base model, PhoGPT-4B, with exactly 3.7B parameters, is pre-trained from scratch on a Vietnamese corpus of 102B tokens, with an 8192 context length, employing a vocabulary of 20480 token types. The chat variant, PhoGPT-4B-Chat, is the modeling output obtained by fine-tuning PhoGPT-4B on a dataset of 70K instructional prompts and their responses, along with an additional 290K conversations. We demonstrate its strong performance compared to previous closed-source and open-source 7B-parameter models. Our PhoGPT models are available at: `https://github.com/VinAIResearch/PhoGPT`

1 INTRODUCTION

Undoubtedly, the success of large language models (LLMs), particularly decoder-only transformer-based generative models such as ChatGPT/GPT-4,[1] LLaMA/LLaMA-2 (Touvron et al., 2023a;b), Mistral (Jiang et al., 2023) and Falcon (Almazrouei et al., 2023), stands out as one of the most significant achievements in recent AI research and development. However, that success has largely been limited to English.

For Vietnamese, we now release the 4B-parameter base pre-trained monolingual model, PhoGPT-4B, along with its chat variant, PhoGPT-4B-Chat, as open-source. We pre-train the base model PhoGPT-4B from scratch on a Vietnamese corpus of 102B tokens with an 8192 context length. Here, it employs a Vietnamese-specific byte-level BPE tokenizer with a vocabulary of 20480 tokens. We further fine-tune the base model on a dataset of 70K instructional prompts and their responses, along with an additional 290K conversations, resulting in the chat variant, PhoGPT-4B-Chat. We demonstrate its strong performance compared to previous closed-source and open-source 7B-parameter models.

Our goal is to provide comprehensive and powerful LLMs for Vietnamese, facilitating future research and applications in generative Vietnamese NLP. Our PhoGPT can be used with popular libraries such as "transformers" (Wolf et al., 2020) and "vllm" (Kwon et al., 2023).

2 PHOGPT

2.1 PHOGPT-4B: MODEL ARCHITECTURE AND PRE-TRAINING

PhoGPT-4B is a Transformer decoder-based model (Brown et al., 2020; Vaswani et al., 2017), which incorporates (Triton) flash attention (Dao et al., 2022) and ALiBi (Press et al., 2022) for context length extrapolation. We train a Vietnamese-specific byte-level BPE tokenizer with a vocabulary of 20480 tokens using the "tokenizers" library.[2] In addition, we use a "max_seq_len" of 8192, "d_model" of 3072, "n_heads" of 24 and "n_layers" of 32, resulting in a model size of 3.7B parameters ($\sim$4B).

---

[1] `https://openai.com/blog/chatgpt`
[2] `https://github.com/huggingface/tokenizers`

Utilizing the Mosaicml "llm-foundry" library (Team, 2023),[3] we pre-train PhoGPT-4B from scratch on a 482GB deduplicated and cleaned pre-training corpus of Vietnamese texts ($\sim$102B tokens). Our pre-training Vietnamese corpus consists of:

- 1GB of Wikipedia texts (version 20/05/2023);
- 1.5GB of medical-related texts crawled from a wide range of publicly available and medical domain-specific websites such as medical journals and universities;[4]
- 3GB of publicly available books spanning a range of genres;
- 12GB of legal data crawled from `thuvienphapluat.vn` and `lawnet.vn`;
- a 40GB variant of the "binhvq" news corpus (version 21/05/2021);[5]
- an 88GB variant of the Vietnamese OSCAR-2301 subset;[6]
- a 336GB variant of the Vietnamese mC4 subset.[7]

2.2 PhoGPT-4B-Chat: Supervised fine-tuning

We then fine-tune the base pre-trained PhoGPT-4B using a dataset consisting of 70K instructional prompts and their responses, along with an additional 290K conversations, constructed by concatenating the following sources:

- 500 instructional prompt and response pairs for poem writing, 500 for essay writing, 500 for spelling correction, 500 for single-document summarization and 1000 for context-based question answering;
- 67K instructional prompt and response pairs from the Vietnamese subset of Bactrian-X (Li et al., 2023);
- 20K Vietnamese-translated ChatAlpaca conversations;[8]
- 40K Vietnamese-translated ShareGPT conversations (without code and mathematics);[9]
- 230K Vietnamese-translated UltraChat conversations (Ding et al., 2023);[10]

The resulting fine-tuned model is named PhoGPT-4B-Chat.

3 Evaluation

3.1 Baselines

We compare PhoGPT-4B-Chat with the closed-source models Gemini Pro in Bard and ChatGPT GPT-3.5-turbo, as well as other open-source 7B-parameter models, including:[11]

- Vistral-7B-Chat is the modeling output obtained by continually pre-training Mistral-7B (Jiang et al., 2023) on a diverse corpus of Vietnamese texts and then performing supervised fine-tuning using a diverse instructional and conversational dataset.[12]

---

[3] `https://github.com/mosaicml/llm-foundry`

[4] The content extracted from these sources contains no private data about the patients.

[5] `https://github.com/binhvq/news-corpus`

[6] `https://huggingface.co/datasets/oscar-corpus/OSCAR-2301`

[7] `https://huggingface.co/datasets/allenai/c4/tree/mC4_3.1.0`

[8] `https://github.com/cascip/ChatAlpaca`

[9] `https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered`

[10] `https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k` (including both train & test sets)

[11] Note that BLOOMZ-7B1-mt, an instruction following variant of BLOOM/BLOOMZ-7B1 fine-tuned with 1.7M Vietnamese examples, does not produce informative textual responses using the greedy search decoding method for the Vietnamese-translated Vicuna questions. BLOOMZ-7B1-mt is thus excluded from our evaluation.

[12] `https://huggingface.co/Viet-Mistral/Vistral-7B-Chat`

Model	All truthful questions	Vietnam-specific
PhoGPT-4B-Chat	38.2 (76 / 199)	40.8 (60 / 147)
Vistral-7B-Chat	41.2 (82 / 199)	42.9 (63 / 147)
Bard Gemini Pro	39.7 (79 / 199)	34.7 (51 / 147)
ChatGPT GPT-3.5-turbo	28.6 (57 / 199)	22.4 (33 / 147)
VBD-LLaMA2-7B-50b-Chat	16.6 (33 / 199)	10.9 (16 / 147)
VinaLLaMA-7B-Chat	11.6 (23 / 199)	8.2 (12 / 147)

Table 1: Obtained results.

- VBD-LLaMA2-7B-50b-Chat is the modeling output obtained by continually pre-training LLaMA-2-7B (Touvron et al., 2023b) on a data combination of 40B Vietnamese tokens and 16B English tokens and then performing supervised fine-tuning using 2M instructional and conversational samples.[13]

- VinaLLaMA-7B-Chat is the modeling output obtained by continually pre-training LLaMA-2-7B on a data combination of 100B English tokens, 230B Vietnamese tokens (from books and news), and 500B automatically generated Vietnamese tokens, and then performing supervised fine-tuning using 1M instructional and conversational samples.[14]

3.2 EXPERIMENTAL SETUP

Our current empirical study only employs the Vietnamese truthful question-answering dataset ViT-ruthfulQA (Nguyen et al., 2023), comprising 199 questions.[15] Each of the 199 questions is fed into 6 experimental models to generate responses, which are then anonymously shuffled. Here, we utilize the greedy search decoding method, which is more suitable for LLM comparison (Lin & Chen, 2023). Two annotators independently assess each generated response on whether the response is correct or not. A response is annotated as "correct" only if it does not contain any hallucinated information. We host a discussion session with the annotators to resolve annotation conflicts.

3.3 MAIN RESULTS

Table 1 presents the obtained accuracy results, showing that overall, our 4B-parameter PhoGPT-4B-Chat is strongly competitive compared to the closed-source Gemini Pro and the open-source 7B-parameter Vistral-7B-Chat. In addition, PhoGPT-4B-Chat produces substantially better scores than the closed-source GPT-3.5-turbo and the other open-source 7B-parameter models VBD-LLaMA2-7B-50b-Chat and VinaLLaMA-7B-Chat. Furthermore, when it comes to 147 of the 199 questions that specifically ask for information related to Vietnam, PhoGPT-4B-Chat obtains better accuracy than Gemini Pro.

Note that the obtained results in Table 1 are based on a single evaluation task and do not necessarily reflect real-world application performance. In future work, we will evaluate PhoGPT on a variety of real-world application tasks.

4 CONCLUSION

We have introduced state-of-the-art open-source 4B-parameter LLMs for Vietnamese, including the base pre-trained PhoGPT-4B and its chat variant, PhoGPT-4B-Chat. We hope that these models will foster future research and applications of Vietnamese LLMs.

---

[13]https://huggingface.co/LR-AI-Labs/vbd-llama2-7B-50b-chat

[14]https://huggingface.co/vilm/vinallama-7b-chat

[15]The original dataset consists of 213 questions; however, 14 of them are not questions or are unclear, e.g. "Tên gọi nào không phải là một loại trái cây Việt Nam?" (Which name is not a Vietnamese fruit?). Therefore, we remove them, resulting in a final evaluation set of 199 questions.

LIMITATIONS

PhoGPT has certain limitations. For example, it is not good at tasks involving reasoning, coding or mathematics. PhoGPT may generate harmful, hate speech, biased responses, or answer unsafe questions. Users should be cautious when interacting with PhoGPT that can produce factually incorrect output.

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Co-jocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Proceedings of NeurIPS*, 2020.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Proceedings of NeurIPS*, 2022.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint*, arXiv:2310.06825, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of SIGOPS*, 2023.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-X : A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation. *arXiv preprint*, arXiv:2305.15011, 2023.

Yen-Ting Lin and Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In Yun-Nung Chen and Abhinav Rastogi (eds.), *Proceedings of NLP4ConvAI*, pp. 47–58, 2023.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 754–764, 2023.

Ofir Press, Noah Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *Proceedings of ICLR*, 2022.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*, arXiv:2302.13971, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*, arXiv:2307.09288, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Proceedings of NIPS*, pp. 5998–6008, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP: System Demonstrations*, pp. 38–45, 2020.