VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

Team: AI/Computer Vision

VietAI system research

Supervisor: Assoc. Prof. Nguyen Dinh Thuc
Students: Thanh Nguyen
Vinh To
Hoang Phan
Hung Do
Khanh Pham

Contents

1 Introduction

VietAI is a prestigious organization dedicated to advancing artificial intelligence (AI) education and research in Vietnam. Founded in 2017, VietAI aims to nurture a strong AI community and equip individuals with the skills and knowledge required to excel in this rapidly evolving field.

At its core, VietAI offers comprehensive AI training programs, workshops, and events designed to cater to individuals from various backgrounds, including students, researchers, professionals, and enthusiasts. The organization collaborates with leading experts and practitioners to deliver high-quality educational content that covers a wide range of AI topics, including machine learning, deep learning, natural language processing, computer vision, and more.

VietAI fosters a vibrant community by providing networking opportunities, mentorship programs, and a platform for individuals to connect, collaborate, and share ideas. The organization hosts regular meetups, hackathons, and competitions that encourage participants to apply their AI knowledge and solve real-world problems.

Moreover, VietAI actively engages in AI research and development to push the boundaries of innovation in Vietnam. By fostering partnerships with academia, industry, and government institutions, VietAI strives to contribute to the growth of AI technology and its ethical implementation.

VietAI has gained recognition for its efforts in promoting AI education and research. The organization has a network of experienced instructors, researchers, and industry professionals who serve as mentors and advisors, providing guidance and expertise to individuals and teams within the community.

Overall, VietAI plays a crucial role in elevating Vietnam's AI ecosystem by empowering individuals with AI skills, fostering collaboration and innovation, and contributing to the advancement of AI technology in the country.

2 Teaching

VietAI offers comprehensive and practical-focused, online courses in the domain of Machine Learning, each lasts for 3-10 weeks. After finishing a course, participants are equipped with a solid mathematical and programming foundation of the subject, as well as hands-on experience by completing quizzes and programming tasks assigned throughout the course. Additionally, at the end of every course, there are guest lecturers

who share their working and researching experiences, providing valuable insights.

Currently, there are 9 online courses available on the official website, which can be categorized into foundation and topic-focused courses:

- **Foundation courses:** These courses cover fundamental knowledge related to the subject, as well as programming techniques for implementing algorithms and models.
  - **Foundation of Machine Learning & Foundation of Machine Learning 05, Foundation of Machine Learning 06:** review concepts in Linear Algebra and Probability and Statistics, Python programming techniques; introduce some Machine Learning libraries (Numpy, Matplotlib, Pandas, TensorFlow), cloud-based development environment Google Colaboratory; introduce the concept of Regression, Binary and Multi-class Classification, and Neural Networks.
  - **Foundation of Data Science 01:** provides an introduction to the Machine Learning pipeline, Exploratory Data Analysis techniques, feature engineering and feature selection, Machine Learning models, and Machine Learning problems.
  - **Foundation of Deep Learning:** provides comprehensive and depp knowledge of Multilayer Perceptron, Convolutional Neural Networks, Recurrent Neural Network & Seq2Seq Model, Attention Mechanism, and Transformer.
- **Topic-focused courses:** These courses concentrate on specific problems, techniques, and state-of-the-art models in Computer Vision and Natural Language Processing.
  - **Advances in Computer Vision:** introduces visual-text multimodal, covers object detection, object classification, image segmentation, and with state-of-the-art models.
  - **Advances in Natural Language Processing, Advances in Natural Language Processing 03:** focuses on Recurrent Neural Networks, Long short-term Memory, Attention Mechanism, and Transformer, Multilingual language Model, Natural Language Generation, and more.

– **ChatGPT/Bard for Everyone:** provides an introduction to Large Language Models, Bard & ChatGPT, and Prompt Engineering and their applications.

By enrolling in these courses, learners can gain in-depth knowledge and practical skills in the field of Machine Learning, and stay up-to-date with the latest advancements in Computer Vision and Natural Language Processing.

3 Research

With the desire to contribute to promoting science in our country, VietAI connects students and members of VietAI with mentors, which are world-class experts in AI, to do cutting-edge research. VietAI has made many contributions to the field of natural language processing, with a focus on Vietnamese. Here is an overview of some of the organization's most notable works in chronological order. These works have helped VietAI become a prestigious organization that provides many values to the AI community in Vietnam as well as applications in many fields.

3.1 ViT5: Pre-trained Text-to-Text Transformer for Vietnamese Language Generation

In 2021, BARTpho [9], a large pre-trained sequence-to-sequence model for Vietnamese that inherits the BART style [4], demonstrated the effectiveness of pre-trained language models on Vietnamese abstractive summarization. However, some past works have shown that the T5 architecture [13] might outperform BART in some aspects (e.g., [11]). Inspired by this, VietAI proposed ViT5 [12], a pre-trained Transformer-based encoder-decoder model for the Vietnamese language. ViT5 is trained on a large corpus of high-quality and diverse Vietnamese texts using T5-style self-supervised pre-training. ViT5 was benchmarked on two downstream text generation tasks: abstractive text summarization and named entity recognition (NER).

Being fine-tuned on two summarization datasets: Wikilingua [3] and Vietnews [8], and one NER dataset: PhoNER [15], ViT5 is compared with other pre-trained language models on both multilingual and monolingual corpora. Their experiments showed that ViT5 achieved state-of-the-art results on summarization in both the Wikilingua and Vietnews corpora, and competitive results on NER on the PhoNER.

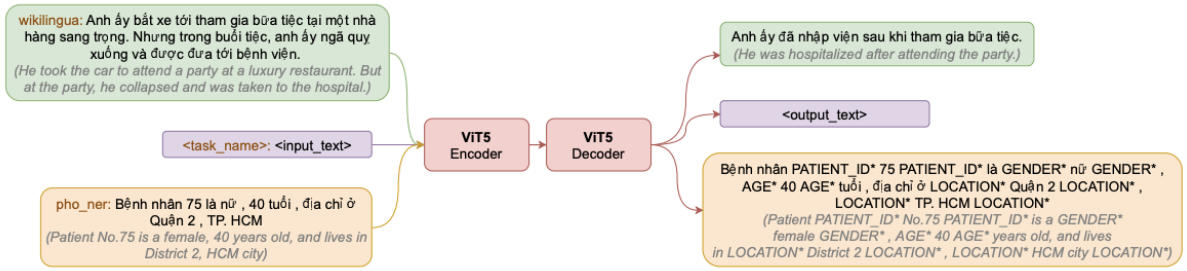

Figure 1: An overview of ViT5 encoder-decoder architecture, with input-output examples of two downstream tasks. For Named Entity Recognition, the decoder reconstructs the sentence with inserted Entity tags

3.2 MTet: Multi-domain Translation for English and Vietnamese

3.2.1 MTet: Multi-domain Translation for English and Vietnamese

In 2022, VietAI introduced MTet (Multi-domain Translation for English and Vietnamese) [6] - an extensive and publicly accessible parallel corpus designed for English-Vietnamese translation. This resource comprises 4.2 million high quality sentence pairs for training purposes, accompanied by a multi-domain test set meticulously curated by the Vietnamese research community. Through the combination of previous works in English-Vietnamese translation, the overall parallel dataset has now expanded significantly, reaching a substantial 6.2 million sentence pairs. In comparison to the current dataset, this dataset stands out for its substantial size and remarkable diversity. It incorporates technical and impactful domains that have been largely overlooked until now, such as law and biomedical data.

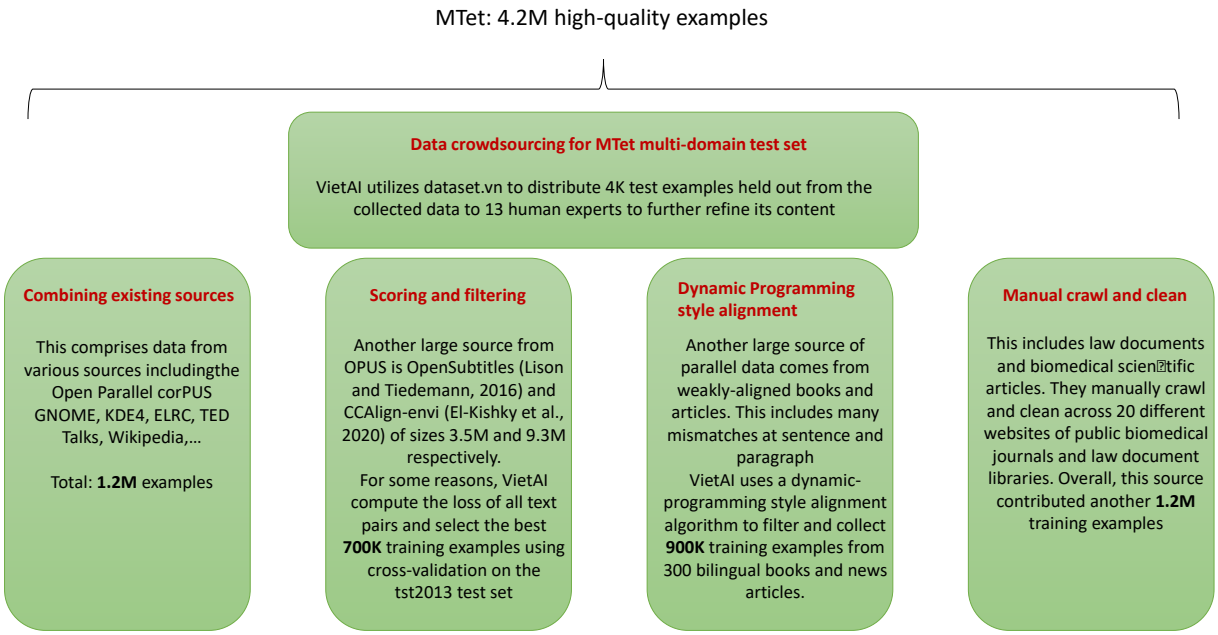

Figure 2: The Components of MTet

3.2.2 Benchmarking EnViT5 and MTet

In addition, VietAI also releases EnViT5 [6] for English and Vietnamese languages, the first pretrained Transformer-based encoder-decoder model for English-Vietnamese, which adopts the encoder-decoder architecture [16] initially introduced by (Vaswani et al., 2017)and the T5 framework [14] proposed by (Raffel et al., 2019). The CC100 Dataset [17] (Monolingual Datasets from Web Crawl Data) (Wenzek et al., 2020), which contains monolingual data for over 100 languages, is used for pre-training the model.

According their experimentation, they validate the quality of both MTet dataset [6]and the performance of their pretrained bilingual model, EnViT5 [6], for both English-to-Vietnamese and Vietnamese-to-English translation tasks. By training their EnViT5(base)[6] model on a combined dataset of MTet[6] and the released PhoMT [1], they have achieved state-of-the-art results in low-resource English-to-Vietnamese translation (45.47) and Vietnamese-to-English translation (40.57). Notably, their EnViT5[6] models surpass the performance of existing multilingual models like mBART[5] and M2M100 [2], all while maintaining a significantly smaller parameter size of 275 million compared to 448 million and 1.2 billion, respectively. This efficiency in parameters not only enables scalability in academia but also holds great promise for industry and community applications.

3.3 Enriching Biomedical Knowledge for Low-resource Language Through Large-Scale Translation

Due to limited resources of biomedical data in Vietnamese as well as in some other languages, VietAI has conducted a research and developed a model that can translate and produce both pretrained and supervised data in the biomedical domains. It enriches the biomedical data warehouse to take advantage of large pretrained models. Although this method is not as effective as self-collecting biomedical data by hiring scientists. However, because of limited human resources scientists, money, and many data sources are not accessible, the approach is less feasible.

To build the ViPubmed and ViMedNLI dataset. They fine-tuned the ViT5 model [12] on 7.2M bitexts corpus consisting of En-Vi datasets from MTet [6] and PhoMT [7] (6.2M) along with 1M pairs of En-Vi biomedical abstracts from the PubmedCorpus [10]. ViPubmedT5 model shows better results than the previous model (ViT5), achieving high results.

In this work, they released two datasets which use ViPubmedT5:

1. **ViMedNLI: A Natural Language Inference Dataset For The Vietnamese Clinical Domain**

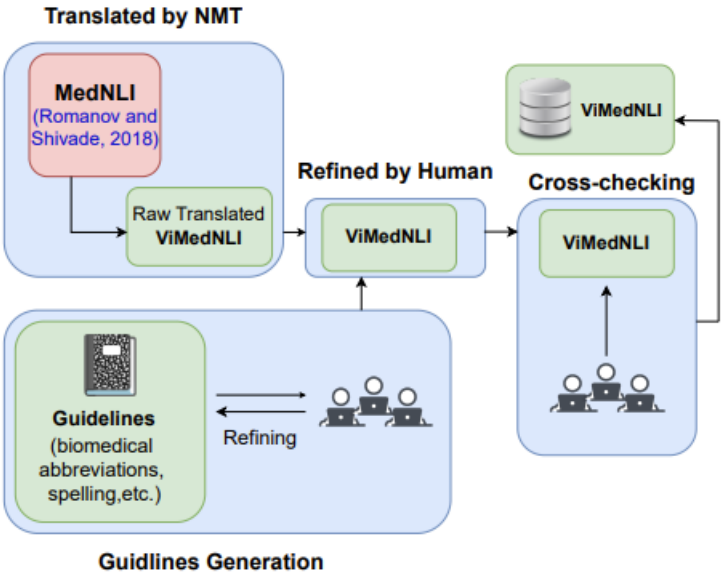

Figure 3: The Process of ViMedNLI Corpus Creation

ViMedNLI dataset is a dataset for natural language inference (NLI) in the biomedical domain for Vietnamese. The dataset is translated from MedNLI, an English

dataset for natural language inference in the medical domain. The ViMedNLI dataset uses an advanced translation model between English and Vietnamese, and is carefully edited by biomedical experts. The dataset can help develop natural language processing applications for the biomedical domain in Vietnamese, a low-resource language.

The process of creating the ViMedNLI is shown in Figure 3. First, they translate MedNLI dataset into Vietnamese by ViPubmedT5. After that, the biomedical experts with experience in natural language processing check and correct the spelling, grammar, vocabulary, and labels of the translated dataset. These experts also evaluated the quality of the dataset according to criteria such as correctness, naturalness, and diversity of language.

2. **ViPubmed: 20M Vietnamese Biomedical abstracts generated by large scale translation**

   Pubmed is a leading online database in the medical field. It was developed by the National Institutes of Health (NIH) and administered by the National Library of Medicine (NLM). The Pubmed provides access to the MEDLINE database which contains titles, abstracts, and metadata from medical literature since the 1970s. The dataset consists of more than 35 million biomedical abstracts from the literature collected from sources such as life science publications, medical journals, and published online e-books. This dataset is maintained and updated yearly to include more up-to-date biomedical documents [10].

4 Community

VietAI is a non-profit organization that aims to build a community of talented AI enthusiasts who can contribute to social development. VietAI believes that AI knowledge should be accessible to everyone, regardless of their background or location. That is why VietAI organizes various events, such as conferences, seminars and workshops, across Vietnam to spread awareness and education about AI.

4.1 VietAI summit

The VietAI summit 2018 was held on December 22nd, 2018 in Ho Chi Minh City, Vietnam. With topic "Emerging Trends In AI", it featured many projects on AI research

and adoption, such as AI applications in agriculture, healthcare, and voice transcription1. Some of the technical highlights of the summit were:

- Kenneth Tran, an engineer at Microsoft Research, introduced his project on AI applications in agriculture. He developed an algorithm that can operate indoor farms in an automatic and effective way, using AI to control the machines and to make decisions on the amount of water and light. He showed that his system can increase the productivity of cucumber crops significantly compared to previous methods1.
- Chuong Huynh, a product engineer at HasBrain and a new graduate of the University of Natural Science, National University of Ho Chi Minh City, presented his healthcare project on the diagnosis system on lung X-ray images by combining chest x-ray bone shadow exclusion with deep learning. This technology can provide an automatic diagnosis for patients through lung x-ray images with high accuracy.
- Do Truong and Minh Thang from VietAI introduced their four demo models on how AI can automatically transcribe Vietnamese with high accuracy. They used deep neural networks and speech recognition techniques to create systems that can recognize and convert speech to text in Vietnamese. This technology can improve the interaction between machines and humans in the Vietnamese language and has been utilized in smart homes1.

The VietAI summit 2019 was held on November 2nd, 2019 in Hanoi, Vietnam. It focused on "AI for the Future", exploring the most recent advances and challenges in AI technologies. Some of the technical highlights of the summit were:

- Dr. Thang Luong, co-founder of VietAI and research scientist at Google Brain, spoke about recent advances in language technologies. He discussed how natural language processing (NLP) can enable machines to understand and generate natural language, such as text and speech. He also shared some of his research projects on NLP, such as neural machine translation, text summarization, and question answering.
- Dr. Hung Bui, director of VinAI Research, spoke about computer vision and its applications. He explained how computer vision can enable machines to perceive and understand visual information, such as images and videos. He also shared

some of his research projects on computer vision, such as face recognition, object detection, and scene understanding.

- Dr. Quoc Le, co-founder of Google Brain and director of VinBrain, spoke about reinforcement learning and its applications. He described how reinforcement learning can enable machines to learn from their own actions and rewards, without explicit supervision or guidance. He also shared some of his research projects on reinforcement learning, such as AlphaGo, AlphaZero, and MuZero.
- Dr. Thuc Vu, co-founder of VietAI and CEO of Kambria, spoke about AI ethics and its implications. He discussed how AI can pose ethical challenges and risks, such as bias, fairness, privacy, and accountability. He also shared some of his initiatives on AI ethics, such as OpenAI, Partnership on AI, and Kambria Code of Ethics.

https://medium.com/kambria-network/vietai-summit-2018-recap-e091f67ce068
https://www.youtube.com/watch?v=KoqS6IkecQk

4.2 Conferences, seminars and workshops

5 Summary

Overall, VietAI system includes three fundamental parts: Teaching, Research and Community.

References

[1] Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation, 2021.

[2] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

[3] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *CoRR*, abs/2010.03093, 2020.

[4] Mike Lewis et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[5] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[6] Chinh Ngo, Trieu H Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. Mtet: Multi-domain translation for english and vietnamese. *arXiv preprint arXiv:2210.05610*, 2022.

[7] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.

[8] Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Hoai. Vnds: A vietnamese dataset for summarization. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 375–380, 2019.

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

[10] Long Phan, Tai Dang, Hieu Tran, Trieu Trinh, Vy Phan, Lam Chau, and Minh-Thang Luong. 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, 2023.

[11] Long Phan, Hieu Tran, Daniel Le, Hieu Nguyen, James Annibal, Alec Peltekian, and Yanfang Ye. Cotext: Multi-task learning with code-text transformer. In *Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021)*, pages 40–47, Online, 2021. Association for Computational Linguistics.

[12] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 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. Association for Computational Linguistics, 2022.

[13] Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[15] Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 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*, 2021.

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

[17] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.