✉ *lnp26@case.edu*
🌐 *longphan.ai*
github.com/justinphan3110
🎓 *Google Scholar*

Long Phan

Education

Bachelor of Science in Computer Science, *Case Western Reserve University* *August 2018 - December 2022*

- **Concentration:** Artificial Intelligence

Manuscripts and Publications

* denotes equal contribution

In review EACL "23	MTet: Multi-domain Translation for English and Vietnamese Chinh Ngo*, TH Trieu*, **Long Phan***, Hieu Tran*, Hieu Nguyen, Minh-Thang Luong https://arxiv.org/abs/2210.05610
In review EACL "23	Enriching Biomedical Knowledge for Low-Resource Languages via Large Scale Translation **Long Phan***, Tai Dang*, Hieu Tran*, Vy Phan, Lam D. Chau, TH Trieu https://arxiv.org/abs/2210.05598
In review Journal of Machine Learning Research	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model Teven Le Scao, Angela Fan, ... , **Long Phan**, ... , Thomas Wolf https://arxiv.org/abs/2211.05100
NeurIPS "22	The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset Thomas Wang, Teven Le Scao, ... , **Long Phan**, ... , Yacine Jernite https://openreview.net/forum?id=UoEw6KigkUn
NAACL SRW "22	ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation **Long Phan**, Hieu Tran, Hieu Nguyen, TH Trieu https://arxiv.org/abs/2205.06457
ACL NLP4Prog "21	CoTexT: Multi-task Learning with Code-Text Transformer **Long Phan**, Hieu Tran, Daniel Le, Hieu Nguyen, James Anibal, Alec Peltekian, Yanfang Ye https://arxiv.org/abs/2105.08645
PLOS Computational Biology Journal	HAL-X: Scalable hierarchical clustering for rapid and tunable single-cell analysis James Anibal, Alexandre G Day, Erol Bahadiroglu, Liam O'Neil, **Long Phan**, Alec Peltekian, Amir Erez, Mariana Kaplan, Grégoire Altan-Bonnet, Pankaj Mehta doi:10.1371/journal.pcbi.1010349
ICONIP "21	SPBERT: An Efficient Pre-training BERT on SPARQL Queries for Question Answering over Knowledge Graphs Hieu Tran, **Long Phan**, James Anibal, Binh T. Nguyen, Truong-Son Nguyen https://arxiv.org/abs/2106.09997
arXiv preprint	SciFive: a text-to-text transformer model for biomedical literature **Long Phan**, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, Grégoire Altan-Bonnet https://arxiv.org/abs/2106.03598
IEA/AIE "21	Hierarchical Transformer Encoders for Vietnamese Spelling Correction Hieu Tran, Cuong V. Dinh, **Long Phan**, Son T. Nguyen https://arxiv.org/abs/2105.13578

Research Experiences

Twitter, *AI Research Intern* *May 2022 - August 2022*

- Worked on Machine Learning Ads Generating project to create value and improve Twitter Ads's Candidate Sourcing
- Developed a Sequential Recommendation with Bidirectional Encoder Representations from Transformer models that leverages transfer learning to predict Ads Tweets interactions
- Achieved internal **state-of-the-art results by 5%** for Tweet Context Generation and Tweet Topic classification

Hugging Face, *Volunteer NLP Research Intern - BigScience Project* *October 2021 - May 2022*

- Collaborated with 300+ NLP Researchers worldwide to develop an open-science Multilingual Dataset ROOTS and Language Model BLOOM for the BigScience workshop
- Participated in the development of the Vietnamese Data co-host team contributing Vietnamese corpus and pipeline

VietAI, *NLP Researcher (Supervised by Minh-Thang Luong and Google TPU Research)* *October 2020 - present*

- VietAI is the largest non-profit AI organization in Vietnam led by Minh-Thang Luong to conduct research on multiple AI topics for the Vietnamese Research Community
- Implemented and built a state-of-the-art English-Vietnamese Machine Translation Model that outperformed Google Translate by more than 2% in BLEU Score
- Released the largest high-quality Engish - Vietnamese Translation MTet corpus (4.2M sentence-pairs) that gained attention from both academia and industry with 130+ stars on GitHub Repository
- Led the development of the first monolingual Transformer Encoder-Decoder Vietnamese model (ViT5) that achieved state-of-the-art results various on Vietnamese NLP tasks

Samsung Research America, *AI Research Intern* *August 2021 - December 2021*

- Developed novel Question-Answering models and NLP/NLU methods for use in the Bixby Intelligent Assistant
- Implemented a Siamese BERT-Networks with a custom Multiple Negative Ranking Loss and Context Attention in Intent Classification for over 100+ Bixby skills
- Achieved internal ***state-of-the-art by 3% in accuracy*** for the Intent Classification tasks for Bixby, Samsung's virtual assistant.

National Cancer Institute (NCI), *Deep Learning Research Special Volunteer* *October 2020 - August 2021*

- Developed a domain-specific text-to-text transfer transformer (SciFive) for biomedical NLP tasks
- Proposed a modified Encoder architecture for single-cell data that can be fine-tuned on a broad range of biological datasets
- Led a team of 6 undergraduate students to win the NIH STRIDES Codeathons

Case Western Reserve University, *Research Assistant (Supervised by Yanfang Ye)* *January 2021 - May 2021*

- Developed CoTexT, a transformer-based model for multi-task learning on both natural language and programming language
- Achieved state-of-the-art results on the Microsoft xGLUE benchmark datasets (focusing on code summarization, code generation, defect detection, code debugging)

Zalo, *AI Research Intern* *April 2020 - October 2020*

- Researched and developed a scalable system translating domain-specific natural language questions into Knowledge Graph's Query for **the first Vietnamese virtual assistant KiKi**
- Initiated the development of a Named-entity recognition (NER) module capturing and resolving date/time text into structured data which was deployed for 80 million users on Zalo, the biggest free message and call application in Vietnam
- Researched and integrated a distilled Hierarchical Transformer model for Vietnamese spelling correction on Laban Key, a Vietnamese Keyboard mobile app with 1 million users.

Academic Experiences

Reviewer for Scientific Data - Nature, TU @ COLING 2022

Teaching Assistant, *Case Western Reserve University*

- *CSDS 133 Introduction to Data Science and Engineering*, Spring & Fall 2022
- *CSDS 393 Software Engineering*, Fall 2022