

# Long Phan

☎ (216) 309 8789

✉ lnp26@case.edu

🐙 [github.com/justinphan3110](https://github.com/justinphan3110)

🌐 [linkedin.com/in/long-phan-3110/](https://www.linkedin.com/in/long-phan-3110/)

## Education

**Bachelor of Science in Computer Science**, Case Western Reserve University

*Expected December 2022*

- **Concentration:** Natural Language Processing - Deep Learning

## Publications

- Chinh Ngo, TH Trieu, Hieu Tran, **Long Phan**, Hieu Nguyen, Minh-Thang Luong. "More Pretraining or Supervised Data: the Curious Case of Low-Resource Machine Translation." *EMNLP 2022, submitted*
- Thomas Wang, Teven Le Scao, ... , **Long Phan**, ... , Yacine Jernite. "The BigScience Corpus A 1.6TB Composite Multilingual Dataset" *NeurIPS 2022, submitted*
- **Long Phan**, Hieu Tran, Hieu Nguyen, Trinh H. Trieu. "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation." *NAACL SRW, 2022*
- **Long Phan**, Hieu Tran, Daniel Le, Hieu Nguyen, James Anibal, Alec Peltekian, Yanfang Ye. "CoText: Multi-task Learning with Code-Text Transformer." *ACL NLP4Prog, 2021*
- Hieu Tran, **Long Phan**, James Anibal, Binh T. Nguyen, Truong-Son Nguyen. "SPBERT: An Efficient Pre-training BERT on SPARQL Queries for Question Answering over Knowledge Graphs." *ICONIP, 2021*
- **Long Phan**, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, Grégoire Altan-Bonnet. "SciFive: a text-to-text transformer model for biomedical literature." *Bioinformatics, 2021, Submitted*
- Hieu Nguyen, **Long Phan**, Hieu Tran. "VieSum: How Robust Are Transformer-based Models on Vietnamese Summarization?." *arXiv preprint arXiv: 2110.04257*
- Hieu Tran, Cuong V. Dinh, **Long Phan**, Son T. Nguyen. "Hierarchical Transformer Encoders for Vietnamese Spelling Correction." *IEA/AIE, 2021*
- Trung-Hieu Tran, **Long Phan**, Truong-Son Nguyen, Tien-Huy Nguyen. "Leveraging Transfer Learning for Reliable Intelligence Identification on Vietnamese SNSs (ReINTEL)." *VLSP, 2020*

## Experiences

**Twitter**, AI Research Intern

*May 2022 - present*

- Working on Machine Learning Ads Generating project to create value and improve Twitter's internal product
- Developing a Sequential Recommendation with Bidirectional Encoder Representations from Transformer models that leverages transfer learning to predict Tweets interactions

**Hugging Face**, Guest NLP Researcher - BigScience Project

*October 2021 - present*

- Collaborating with 600+ NLP Researchers around the world to develop a biggest Multilingual Dataset and Language Model for the [BigScience workshop](#)
- Leading the development of the Vietnamese Data co-host team contributing Vietnamese corpus, pipeline, and multiple pretrain language models (300M+ parameters) for the project

**VietAI**, NLP Researcher (Supervised by Minh-Thang Luong and Google Cloud Platform)

*October 2020 - present*

- VietAI is a largest non-profit AI organization in Vietnam led by Thang Luong and Thuc Vu to conduct research on multiple AI topics for the Vietnamese Research Communities.
- 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 English -Vietnamese Translation corpus (4.2M sentence-pairs) which is analyzed to have higher quality on an EMNLP 2021 study on different Transformer architecture settings.
- Gained attention from both academia and industry with 100+ stars on GitHub Repository

**Samsung Research America**, AI Research Intern

*August 2021 - December 2021*

- Developed new 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.

## Technical Skills

**Languages:** Python, Java, JavaScript

**Technologies and Libraries:** PyTorch, Tensorflow, Tensor2Tensor, HuggingFace, Transformers, JAX, FLAX, Node.js, Flask, React, MongoDB, GCP, AWS, Docker, Spring Boot, Kafka, Elastic Search, Redis, Zookeeper