BUI CHI MINH

Ha Noi, Viet Nam
(+84) 886870997 ⋄ buichiminh.cntt@gmail.com

EDUCATION

VNU University of Engineering and Technology *2023 - now*
MSc. Computer Science
Current GPA : 3.54
Relevant courses : Natural Language Processing using Deep Learning (4.00/4.00) , Modern issues of Computer science (4.00/4.00), AI Specialization (4.00/4.00)

Posts and Telecommunications Institute of Technology *2015 - 2020*
BSc. Computer Science
Relevant course: Artificial Intelligent (4.00/4.00), Data Structure (4.00/4.00), Programming Languages (4.00/4.00)

EXPERIENCE

NLP Engineer - Viettel Group *Jun 2022 -now*

1. News Categories Classification
- Developed and optimized the model for CPU servers with improved overall accuracy to 85% and reduced latency by more than 30% compared with older models.
- Automated the labeling phase, significantly reducing financial and labor costs, and integrated the application into the website system.
- Designed a recommendation model focusing on delivering personalized news for private customers.
Outcome: Successfully integrated into the Reputa system, supporting 200+ companies for social listening purposes.

2. End-to-End QA System
- Designed deep learning models for an end-to-end question answering system, achieving a top-3 in ZALO AI Challenge.
- Leveraged GPUs and optimized code for cross-training and parallel training.
- Implemented bi-encoder and evaluated cross-encoder for the retrieval phase.
- Built a Docker services and integrated in K8S.

3. Court Judgement Prediction and Explanation
- Used large sparse transformer models (Longformer) combined with lexical models to enhance information and improve task accuracy.
- Achieved first place in competitions with a scientific paper accepted at SemEval (ACL).

4. Research on LLM
- Experimented with and deployed a advanced RAG pipeline, including fine-tuning LLM models (e.g., Llama series, Qwen series) with LoRA, qLoRA and augmenting data for specific use cases (e.g., legal QA), using LangChain and custom libraries.
- Deployed and optimized inference speed for Llama and decoder-based LLMs using vLLM and exLlama.
- Leveraged SFT and DPO for Llama3 series to achieved SOTA in VMLU (Vietnamese LLM benchmark) using crawled and synthetic data from top models.

5. Further research and implementations
- Developed a domain-centric Knowledge Graph, optimized using the GraphRAG approach, and submitted findings to SIGIR 2025.
- Built and integrated services into Kubernetes (K8s) for scalable and efficient production deployment.

Data Scientist, Vingroup *Jun 2021 - Jun 2022*

1. Tracking human workforce using vband

- Worked with Human Activity Recognition (HAR) problems to helped the manager tracking employees workforce. (https://news.microsoft.com/vi-vn/2021/09/06/vantix-ung-dung- tri-tue-nhan-tao-nang-cao-nang-suat-lao-dong/)
- Used various timeseries model and approached to improved the accurate of the predictions. (Worked with RNN, CNN, ensemble models and XGBoost)
- Worked with GPS correction problems for tracking worker in the construction (applied in VinHomes construction).
- Used Kalman-Filtering algorithms to improved the roadmap for GPS and also reduced the GPS flashing, flicking problems.
- Built a FastAPI demo for map tracking and dockerized the application into production.

Product: The application has served all the VinPearls and some VinHomes projects.

AI Engineer, Viettel Group *Jan 2020 - Jun 2021*

1. Duplicate news and video recommendation

- Designed an algorithm for video recommendation within the stipulated time. Used and applied tensorflow as main application and flask for api. Worked with a 5-members research team for accomplishing this.
- Increased the average time per devices by 5-10% and the total amounts of time by 13% by A/B testing.
- Detected duplicate news cover 95% accuracy over a thousand news per day. This helped the moderators to reduced 80% working hours to check the duplicate ones.

Product: The features has applied in Mocha (app for Android and IOS) and Netnews.vn

2. Auto tuning firewall project

- Developed a system that detect anomaly user behaviors for auto tuning WAF system.
- Built a system that collected the data and detect bad requests to reduce false positive from traditional WAF (Web ApplicationFirewall).
- Built a profile prototype and a model to tracked IP to rank users rating.
- Reduced the human workforce in monitoring with an acceptable accuracy.
- Applied quantization to the model to reduce the latency (support around 100 mils user requests per day).
- Developed a DNS monitor module (tracking DNS to alert redirect site) using cron job.

Products: The application has integrated into the WAF and served more than 100 millions requests per day and denied around 100 to 1000 IP attackers per day.

AWARDS AND SCHOLARSHIPS

Professional Machine Learning Engineer Certification , Google Cloud *2024*

AI City Challenge 2024. , CVPR Workshop *2024*

Awards: Top 4 Public

Task: Challenge Track 2: Traffic Safety Description and Analysis. (team VAI)

SemEval-2023 Task 6: LegalEval , ACL *2023*

Awards: Top-1 ranking in the subtask C1 (Legal Judgment Prediction)
and top-2 ranking in the subtask C2 (Court Judgment Prediction & Explanation).

Zalo AI Challenge 2022. , Zalo *2022*

Awards: Impressive Solution, Top 3/103.

Task: End-to-end Question Answering, finding an answer for
each question in Vietnamese Wikipedia Corpus.

BKAV Online CTF competitions , BKAV *2020*

Awards: Top 5 leaderboard for CTF.

Scholarship for Global Problems Based Learning (gPBL 2019) , Shibaura University *2019*

PTIT CTF , PTIT *2018,2019*

Awards: Third prize and second prize

PTIT ACM/ICPC , PTIT *2018,2019*

Awards: 2 Consolation prizes

SKILLS

Machine Learning

- Proficiency: Natural Language Processing, Deep Learning, Computer Vision
- Intermediate: Data analytics, visualization

Programming Languages

- Proficiency: Python
- Intermediate: Java, bash

Tools and Frameworks

- Proficiency: Pytorch, Git, Docker
- Intermediate: Tensorflow, Keras, Flask, FastAPI

Database & Cloud

- Proficiency: MySQL
- Intermediate: MongoDB, GCP
- Basic: AWS, Azure

Others

- K8S, Crontab, Ubuntu, MLOps

LANGUAGES

- English (Intermediate), Vietnamese (Native)

PUBLICATION

1. Thanh Dat Hoang, Chi Minh Bui, Khac Hoai Nam Bui. "Viettel-AI at SemEval-2023 Task 6: Legal Document Understanding with Longformer for Court Judgment Prediction with Explanation". Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023).

2. Thanh-Do Nguyen, Chi Minh Bui, Thi-Hai-Yen Vuong and Xuan-Hieu Phan. "Passage-based BM25 Hard Negatives: A Simple and Effective Negative Sampling Strategy For Dense Retrieval". Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation. (PACLIC 37)