Hieu Nguyen

Los Angeles, CA, 91324 | (323)-974-3478 | LinkedIn | Portfolio | hieuhocnlp@gmail.com | github.com/hieuchi911

Education

University of Southern California **G05/2024**

Master of Science in Computer Science specialized in AI | GPA: 3.76/4.0

- *Relevant courses: ML, Applied NLP, AI, Deep learning and its applications, Algorithms, Web Development*

International University - Vietnam National University HCM City (IU) **Sep 2017–Sep 2021**

Bachelor of Engineering in Information Technology | GPA: 3.54/4.0 *Full Scholarship*

Research Experience

Deep Learning Research Engineer May 2024

USC Information Sciences Institute - On the effects of Knowledge Distillation on LLM hallucinations

- Set up token level knowledge distillation (KD) of Llama-3-8B with Llama-3-70B as teachers and evaluate on hallucination benchmarks (HaluEval, XSum, CNN Dailymail) in H100 clusters.
- Design robust automatic hallucination evaluation pipeline with internal metrics (lookback ratio) and external metrics (rougeL, factual consistency from LLM hallucination detector).
- **Consistent improvements (up to 6%)** on the hallucination evaluation pipeline.

Research Assistant Mar 2023–Dec 2023

Deep USC Research Group - Silent bug detection with ABTorch

- Set up 4000 parallel ablation experiments of training/inference (with quantization, LoRA adapters, gradient checkpointing/accumulating, etc.) of bug-injected transformers on a GPU cluster.
- Unit tested and code reviewed Ablator (a framework for efficient scaling of thousands of deep learning experiments).
- Contributed to Ablator's tutorials (prototype, scale experiments for ablation studies and HPO.) and documentation.

Study and Develop Generative Model Chatbots

- Built the HRED+C-VAE model (hierarchical conditional **Variational Auto-Encoder**) for domain-controllable dialog.
- The model was able to control response generation and achieved similar evaluation results. View project here

Professional Experience

Backend Engineer Dec 2021–Jul 2022

CT Group Vietnam Joint Stock Company

- Deployed to production linux server RESTful APIs with docker Rasa assistants for customer support.
- Set up a multi-modal virtual assistant pipeline of docker microservices and integrated with Meta apps.

Projects - Competitions

Towards LLM-based robot control

- DDP multi-nodes finetune VideoLlaMA vision language model to generate estimation of human pose, height, moving velocity (for downstream robots control), taking into account the visual context.

Text editing modeling

- Identified architectural bugs of SmolLM-135M, ran SFT for text editing, further **improved by 2%** with **DPO**, **IPO**.

A **synthetic dataset** for Malicious Content Detection in Political Settings

- Fallacy Generation from political speeches with quantized Mistral7B, few-shot prompting and consistency filtering, improved the data scarcity with the synthetic **LOGICPOLITICS dataset 12,489 new data points**

EduSummarise the mind maps creator - **Runner up at TrojanHacks Spring 2024**

- Leveraged GPT3 to generate notes from lecture transcripts, based on which to extract relations between entities.
- Construct knowledge graph from the relations and visualize as mind maps. View our solution here.

Machine Learning and Deep learning mini projects

- Classification of sentiment, POS tagging, NER, with Naive Bayes, SVM, RNNs, BERT, etc.
- Implemented from scratch popular ML/DL models: tree based models, neural networks, CNNs, GNNs, Transformers.

Teaching Experience

International University - Vietnam National University HCM City **2021**
Teaching Assistant, Introduction to Data Mining

- Evaluated course examinations, written assignments, and weekly quizzes
- Hold discussion sessions for homeworks and assignments

Awards and Scholarships

Recipient of Vingroup STEM Scholarship for Master's and Ph.D. study in the US **2022**
$126,000 for 32 selected scholars in the whole country

Accepted candidate MOST GASE Global Talent Internship in Taiwan (canceled) **2020**
1 out of 4 Vietnamese that got selected

Full Scholarship from International University- Vietnam National University HCMC **2017**

Technical Skills

- Engineering: **Docker**, **High Performance Computing Clusters, SLURM job scheduler, large-scale model parallel training**; Python, Object-oriented programming, bash/shell Scripting, **GCP**, Linux, **Git**, CI/CD, Flask, SQL
- Large Language Models: huggingface libraries (**transformers, datasets, PEFT, accelerate**), **vLLM**, **instruction fine-tuning**, **LM evaluation harness**, knowledge distillation; **DPO, IPO for alignment**
- Machine Learning: **PyTorch**, **deepspeed**, **DP, DDP, FSDP**; VSCode debug distributed execution; Weight and Biases; EDA (**Pandas, Matplotlib, seaborn**), Scikit-learn, Pandas, Numpy, Matplotlib, Optuna, XGBoost, Selenium