Pengfei **He**

NLP Researcher · Machine Learning Engineer · Software Developer

3 years of NLP research & work experience focusing on key components including retrieval, generation, and instruction tuning.

☎ 425-772-7623 | ✉ hepengfe@uw.edu | ⌂ hepengfei.ml | hepengfe | hepengfe | hepengfe | Pengfei He

Education

University of Washington *Seattle, WA*

B.S. Applied Computational & Mathematical Sciences, Data Science & Statistics. Minor in Linguistics. *Sept. 2018 - June. 2021*

Work Experience

H2Lab at Paul G. Allen Computer Science School *Seattle, WA*

Part-Time Research Assistant, Mentor: Yizhong Wang *April. 2023 - Present*

- Designed and managed a distributed pipeline for model/data loading and training with **accelerate**, ensuring an interruption-resistant **project codebase**, and conducted **large-scale experiments**(70k+ gpu hours) on High Performance Cluster(HPC). The robustness of our codebase guarantees the accurate and successful completion of all experiments.
- Investigated and applied parameter efficient tuning methods such as LoRA and Adapter on the latest Large Language Model such as LLAMA, and trained them on instruction-tuning datasets, and analyzed the experiment results on wandb/tensorboard.
- Evaluated and deployed the optimal distributed parallel training methods such as **FSDP, DDP** and **deepspeed** for instruction tuning. Opting for the suitable training framework significantly accelerate the experimentation process.

Petuum Inc. *Sunnyvale, CA*

Machine Learning Engineer (NLP) *Aug. 2021 - April. 2023*

- Researched and implemented **Asynchronous Index Refresh** from **ANCE** based on forte project framework for QA applications, and it enables training the wikipedia passage embeddings on a cluster with low per-GPU-memory(12GB) and makes the **training significantly more efficient** than the traditional sequential pipeline of training and inference.
- Contributed to **open-source project forte** by adding **multi-modal data ingestion** and writing user-friendly **documentations** with a CI/CD pipeline. This involves following up issues, releasing new versions regularly, implementing test cases for new features. These contributions have significantly improved the usability of the repository.

H2Lab at Paul G. Allen Computer Science School *Seattle, WA*

Undergraduate Research Assistant, Mentor: Sewon Min, Aida Amini *Sept. 2020 - March. 2021*

- Adopted **sequence-to-sequence Transformer** models using Huggingface for **question answering systems** to generate a sequence of answers, and **pre-trained** them on NQ dataset and **fine-tuned** them on AmbigQA dataset to improve its performance on downstream tasks.
- Developed a **clustering-assisted question answering system** to address **question ambiguities** and to improve **answer diversity**, and our model achieved **higher recall** than the baseline model.
- Implemented **parallel model inference** on multiple GPU and utilized all CPU threads to prepare batch data, and it **speeds up the evaluation process by 4 times** and the whole training process significantly under 2-GPU settings.
- Worked closedly with the lab researchers and biologists, developing **Python scripts** to **parallelize the data preprocessing** of unstructured, document-level medical text from PubMed at a large scale. This formatted data serve as input for an end-to-end medical relation extraction system. The system's output aided our biologist collaborators in efficiently navigating and interpreting biomedical literature efficiently.

Projects

Survey of Spontaneous Emergent Discrete, Compositional and Point-Symmetrical Signals *Seattle, WA*

ACMS Honor Thesis Advisor: Shane Steinert-Threlkeld *Spring 2020 - Fall 2020*

- Implemented the cross-entropy loss function and adjusted model output accordingly for existing experiments using PyTorch.
- Analyzed clustering results of the intermediate layer of autoencoder under the new training conditions in Jupyter Notebooks and discovered a point symmetry phenomenon for min/max functions.

Courses

Deep Learning	CSE543 Deep Learning, CSE599I Generative Model
Machine Learning	CSE547 Machine Learning for Big Data, CSE546 Machine Learning
Natural Language Processing	CSE517 Natural Language Processing, CSE599D1 Multilingual NLP Seminar, LING572 Statistical NLP
Prescriptive Analytics	CSE542 Reinforcement Learning, CSE573 Artificial Intelligence
Data Analytics	CSE414 Database System, SOC225 Data & Society
Algorithm	CSE521 Advanced Algorithms, CSE373 Data Structure & Algorithm
	Course numbers above 500 represent **graduate levels*

Skills

Programming	Python, Huggingface, PyTorch, Multiprocessing, Linux, Java, Numpy, Spark, MapReduce, SQL
Distributed Training	Deepspeed, Accelerate, Ray, Slurm

Honors & Awards

2018~2021	**Dean's List**, Undergraduate academic scholarship over six quarters.	*Seattle, WA*
2021	**ACMS Honors Student**, Departmental Honors for students with academic excellence and an honor thesis.	*Seattle, WA*

Publication

2023	**Parameter Efficient Instruction Tuning: an Empirical Study**, 1st author
2022	**RAINIER: Reinforced Knowledge Introspector for Commonsense Question Answering** , 4th author