Xinyue (Sherry) Chen

+1 412-251-2831
sherryicss@gmail.com
San Jose, CA 95128
Google Scholar

EDUCATION

Carnegie Mellon University (CMU) — May 2023

- Master of Computational Data Science, School of Computer Science
- Selected Courses: Advanced NLP, Deep Learning, Deep Learning Systems (*ML framework with CUDA*)

Shanghai Jiao Tong University (SJTU) — June 2020

- Bachelor of Engineering in Computer Science and Technology

PUBLICATIONS

1. Revisiting the Role of Language Priors in Vision-Language Models
   in Proc. of ***ICML 2024*** | project
   Xinyue Chen*, Zhiqiu Lin*, Deepak Pathak, Pengchuan Zhang, Deva Ramanan
2. Hybrid Transducer and Attention based Encoder-Decoder Modeling for Speech-to-Text Tasks
   in Proc. of ***ACL 2023*** (Outstanding Paper)
   Yun Tang, Anna Y. Sun, Hirofumi Inaguma, **Xinyue Chen**, Ning Dong, Xutai Ma, Paden D. Tomasello, Juan Pino
3. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering.
   in Proc. of ***EMNLP 2020*** | paper | code
   Xinyue Chen*, Yanlin Feng*, Jun Yan, Bill Yuchen Lin, Xiang Ren
4. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning.
   in Proc. of ***EMNLP-IJCNLP 2019*** (Oral Presentation)
   Bill Yuchen Lin, **Xinyue Chen**, Jamin Chen, Xiang Ren

WORK EXPERIENCE

Machine Learning Engineer @ TikTok — Oct. 2023 - Present
Query & Content Understanding in TikTok Search — *San Jose, CA*

- **Building TikTok generative AI search**: a reliable, comprehensive and intelligent RAG-based response engine
  - Designed, developed, and deployed a sophisticated answer/response generation pipeline that delivers high-quality and relevant responses tailored to user search needs. This system seamlessly integrates in-app information and world knowledge while adapting response styles based on fine-grained query intent. The generated textual responses are integrated with strategically selected videos and are prominently displayed at the top of the search pages, enhancing search experience in user's quest for knowledge and information. As a result, we observed *significant increases in user search satisfaction metrics*.
  - Performing LLM post-training by integrating public data, in-house data, and domain knowledge, towards a powerful domain expert answer engine with the flexibility to adapt to evolving requirements.
- **Owner of query correction** for TikTok Search across all channels (general search and all vertical searches).
  - Enhancing the query correction pipeline, with a diverse array of models, including LLMs, compact neural models, and tree-based models, to deliver improved search experience for *multi-lingual* user interactions in over 20 languages. Employed various data mining techniques.
  - Designed and implemented a RAG strategy utilizing LLMs. With extensive experiments on training data and models, the final strategy significantly enhanced human evaluation metrics and A/B metrics.

Research Engineer Intern @ Meta (FAIR) — May 2022 - Aug. 2022
Transducer-Based Models for Simultaneous Speech Translation, Advisor: Juan Pino, Ning Dong — *Menlo Park, CA*

- Implemented high-quality transducer-based streaming speech translation systems based on cutting-edge papers from scratch in Fairseq library.
- Benchmarked and significantly improved the performance of the implemented systems both in terms of translation quality and streaming latency.

Applied Scientist Intern @ Amazon Web Services — Sept. 2020 - Aug. 2021
Task Selection for Multi-task Learning and Robustness in NLP, Advisor: Prof. He He

- Examined the relationship between the Fisher information matrix (FIM) and the training trajectories in multi-task learning framework and utilized FIM for task selection in NLP multi-task learning.
- Designed synthetic data to empirically study the debiasing effects of various training strategies and regularization methods and conducted experiments on natural language inference datasets (MNLI, QQP) to test the effect of these strategies on real-world data.

SKILLS

Programming languages: Python, C/C++, CUDA, Java, Verilog

Libraries: PyTorch, TensorFlow, MySQL, PySpark, NumPy, Pandas, Kubernetes

PROJECT EXPERIENCE

Compositionality Reasoning of Vision-Language Models Dec. 2022 - May 2023

Research Assistant, Advisor: Prof. Deva Ramanan, Pengchuan Zhang *CMU, Meta*

– Designed a method that leverages multi-modal generative models for image-text matching that requires compositionality understanding. The proposed method surpasses previous SOTAs on visio-linguistic compositionality benchmarks and serves as a quantitative diagnostic tool for unimodal bias of benchmarks.

Relational Reasoning for Natural Language Understanding July 2019 - May 2020

Research Assistant, Advisor: Prof. Xiang Ren *University of Southern California*

– Designed Graph Relation Network, a variant of GNNs capable of performing higher-order message passing over multi-relational knowledge graphs. This enables a system that incorporates LLMs and static knowledge graph, facilitating relational reasoning in natural language question answering tasks. The proposed model surpasses existing knowledge-augmented methods on CommonsenseQA and OpenbookQA datasets.

– Co-designed KagNet, a model that incorporates external knowledge graph with text for CommonsenseQA, a multiple-choice question answering dataset, and was able to achieve state-of-the-art performance.