Yun-Cheng (Joe) Wang

Mountain View, CA | joewang622@gmail.com | Google Scholar | joewang622

yunchengwang.github.io

About Me

I am a **machine learning** researcher and engineer devoted to developing **efficient**, **lightweight**, and **scalable** systems, with expertise in **knowledge graphs (KGs)**, **natural language processing (NLP)**, **large language models (LLMs)**, **information retrieval**, and **on-device AI**.

With a strong commitment to impactful research, I have led collaborations across industry and academia, developing AIML solutions that serve **millions of users**. I am also dedicated to mentoring and fostering innovation in teams, driving forward both individual growth and **cutting-edge AIML technologies**.

Technical Skills

Programming Languages: Python, C++, Java, SQL, SPARQL, Bash

Big Data & ML: Hadoop, PySpark, PyTorch, TensorFlow, XGBoost, LightGBM, FastText, Transformers, Pandas

Education

University of Southern California – Los Angeles, CA Jan 2021 - Dec 2023
Ph.D. in Electrical Engineering

- Dissertation: Green Knowledge Graph Completion and Scalable Generative Content Delivery

University of Southern California – Los Angeles, CA Aug 2018 - Dec 2019
M.S. in Electrical Engineering

- Relevant Coursework: Pattern Recognition, Multimedia Compression, Convex Optimization

National Taiwan University – Taipei, Taiwan Sep 2014 - Jun 2018
B.S. in Electrical Engineering

- Relevant Coursework: Digital Speech Processing, Machine Learning Foundations, Artificial Intelligence

Professional Experience

Research Scientist, Yahoo, Inc. – Mountain View, CA Jan 2024 – Present

- Developed machine learning solutions with state-of-the-art technologies for query understanding and whole-page optimization in Yahoo Search.
- Designed metrics and experiments to analyze problems and measure results.
- Optimized model size and reduced serving latency to serve over 900 million monthly active users.

Research Intern, Yahoo, Inc. – Remote Jun 2023 – Aug 2023

- Innovated a fact ranking mechanism to generate knowledge-grounded entity descriptions.
- Curated a high-quality data-to-text dataset containing 20K examples using LLMs and KG fact-checking.

Data Science Intern, Taboola, Inc. – Los Angeles, CA Jun 2019 – Aug 2019

- Discovered trending topics in news articles through network analysis.
- The topic graph was incrementally updated based on over 20K daily articles.

Research Projects

Decoupling Semantics and Syntax in Language Models Aug 2023 - Present

- Developed lightweight and trustworthy models for domain-specific language generation, e.g., bio-medical.
- Adopted knowledge graphs to capture semantic patterns in natural language.

- Modualized language models through a bottom-up manner to enhance efficiency and interpretability.

Multi-modality Alignment, *Sponsored by Army Research Lab (ARL)* Aug 2023 - Present

- Leveraged the embedding space to connect different modalities for multi-modal reasoning.
- Extracted human-object interactions using spatial and latent features with hierarchical classifiers.
- Devised an alignment module in a joint embedding space for text-to-image and image-to-text retrieval.

Scalable Generative AI Services under Edge-Cloud Computing Jan 2023 - Oct 2023

- Analyzed the memory, computation, and network requirements to deploy GenAI services, e.g., ChatGPT, across different scales.
- Estimated the latency for GenAI services under different communication frameworks.
- Identified considerations when designing GenAI systems with better efficiency, computation offloading, and privacy.

Efficient Reasoning on KGs using Lightweight Models Jan 2021 - Oct 2023

- This project aimed at predicting missing information, including entity types and relations, in knowledge graphs using lightweight models.
- Leveraged feature pruning to achieve parameter efficiency and SOTA performance in low dimensions.
- Proposed novel modeling of entity types to improve expressiveness while retaining scalability to large KGs.
- Innovated an asynchronous KGE learning framework to improve performance on both link prediction and entity type prediction tasks.
- Overall, inference FLOPs were reduced 100 times, and the number of parameters was reduced 15 times.

Selected Publications

[1] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "AsyncET: Asynchronous Representation Learning for Knowledge Graph Entity Typing," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[2] **Yun-Cheng Wang**, Jintang Xue, Chengwei Wei, C.-C. Jay Kuo, "An Overview on Generative AI at Scale with Edge-Cloud Computing," *IEEE Open Journal of the Communications Society*, 2023.

[3] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "GreenKGC: A Lightweight Knowledge Graph Completion Method," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[4] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Compounding Geometric Operations for Knowledge Graph Completion," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[5] Zhanxuan Mei, **Yun-Cheng Wang**, Xingze He, C.-C. Jay Kuo, "GreenBIQA: A Lightweight Blind Image Quality Assessment Method," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2022. **(Top 10% Paper Award)**

[6] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "KGBoost: A Classification-Based Knowledge Base Completion Method with Negative Sampling," *Pattern Recognition Letters*, 2022.

[7] Fenxiao Chen, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Graph Representation Learning: A Survey," *APSIPA Transactions on Signal and Information Processing*, 2020. **(2024 Sadaoki Furui Prize Paper Award Nominee)**

[8] Bin Wang, Angela Wang, Fenxiao Chen, **Yun-Cheng Wang**, C.-C. Jay Kuo, "Evaluating Word Embedding Models: Methods and Experimental Results," *APSIPA Transactions on Signal and Information Processing*, 2019. **(2022 Sadaoki Furui Prize Paper Award)**

Academic Services

Conference Reviewer: KDD, EMNLP, ACL Rolling Review (ARR), COLING, ACML, ECML, ICASSP

Journal Reviewer: IEEE/ACM Transactions on Audio, Speech and Language Processing (T-ASL), IEEE Internet of Things Magazine (IoTM), IEEE Transactions on Artificial Intelligence (TAI)