WEI WU

Reinforcement Learning, Generative AI, AI search
(+86)155-0127-7588 ◇ lxww301@gmail.com

WORK EXPERIENCE

TikTok Inc., San Jose *07/2024 Present*
Tech Lead, TikTok AI Search

- **TikTok LLM Pretraining:** Contributed to the development of *Thoth*, a 7B-parameter large language model (LLM) for TikTok. Designed and implemented innovative KV cache reduction techniques, achieving a **32x throughput improvement** over comparable open-source 7B models. Enhanced context-length extension methods, extending the context length to **128k**.
- **TikTok LLM Alignment:** Led the alignment process for *Thoth* using an improved knowledge distillation approach combined with RLHF-based fine-tuning. The aligned model outperformed all other open-source 7B models in **multilingual benchmarks**.
- **Search RAG Development:** Played a key role in TikTok's Search RAG system by managing the training pipeline for the summary generation model. Utilized advanced answer-planning techniques to create high-quality demonstration and preference datasets. Improved the models ability to deliver well-structured results, significantly boosting **CTR for top results** and **average card stay time**.

Bytedance, Beijing *04/2020 07/2024*
Generative AI Researcher, Seed Alignment

- **Reward Modeling:** Developed process reward models for mathematical and general reasoning tasks and introduced **reward model merging** techniques to address reward hacking challenges.
- **Reinforcement Learning:** Contributed to data selection and sampling strategies for RL training. Played a key role in the development of *VeRL*, Bytedance's open-source RL training framework.

Machine Learning Engineer, Search

- **Douyin Visual Search:** Developed early versions of recall models for Douyin Visual Search and led the creation of fine-ranking models, enhancing search precision and user experience.
- **Toutiao Search Pre-training:** Designed a funnel transformer-based pretraining model to enable long-range query-document interactions for Toutiao's Precise Question Answering (PQA) system. Implemented a **multi-task distillation system** to compress 23 PQA-related models into one, improving efficiency and reducing complexity.
- **TikTok Search Pre-training:** Developed TikTok's first multilingual Search BERT model, applied across fine-ranking relevance, vector-based recall, and query recommendation systems. Successfully launched in all TikTok Search markets (except P0 countries).

ShannonAI, Beijing *07/2019 02/2020*
Machine Learning Engineer

- Pretrained a Chinese BERT model tailored for financial applications using crawled financial data on Google Cloud TPU Pods.
- Designed and implemented an **end-to-end trainable financial information extraction framework** to extract insights efficiently from financial documents.

EDUCATION

Peking University, Beijing, China *09/01/2016 - 07/01/2019*
MS.E. in Computer Science; Advisor: Prof. Houfeng Wang
School of Electronics Engineering and Computer Science
Research focus: Answer Selection and Sentence Representation

Nanjing University, Nanjing, China *09/01/2012 - 07/01/2016*
B.E. in Microelectronics; Advisor: Prof. Chenglei Peng
School of Electronic Science and Engineering

PUBLICATIONS

Coreference Resolution as Query-based Span Prediction
Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, Jiwei Li
ACL 2020, Poster

Description based text classification with reinforcement learning
Duo Chai, **Wei Wu**, Qinghong Han, Fei Wu, Jiwei Li
ICML 2020, Poster

Glyce: Glyph-vectors for Chinese Character Representations
Wei Wu, Yuxian Meng, Qinghong Han, Muyu Li, Xiaoya Li, Ping Nie, Xiaofei Sun, Jiwei Li
NeurIPS 2019, Poster

Towards Comprehensive Description Generation from Factual Attribute-value Tables
Tianyu Liu, Fuli Luo, Pengcheng Yang, **Wei Wu**, Baobao Chang, Zhifang Sui
ACL 2019, Poster

Question Condensing Networks for Answer Selection in Community Question Answering
Wei Wu, Xu Sun, Houfeng Wang
ACL 2018, Poster

Phrase-level Self-attention Networks for Universal Sentence Encoding
Wei Wu, Houfeng Wang, Tianyu Liu, Shuming Ma
EMNLP 2018, Oral

SGM: Sequence Generation Model for Multi-label Classification
Pengcheng Yang, Xu Sun, Wei Li, **Wei Wu**, Houfeng Wang
COLING 2018, Best Paper Award

Bi-directional Gated Memory Networks for Answer Selection
Wei Wu, Houfeng Wang, Sujian Li
CCL 2017, Poster

Icorating: A Deep-learning System for Scam ICO Identification
Shuqing Bian, Zhenpeng Deng, Fei Li, Will Monroe, Peng Shi, Zijun Sun, **Wei Wu**, Sikuang Wang, William Yang Wang, Arianna Yuan, Tianwei Zhang, Jiwei Li

HONORS AND AWARDS

Special academic scholarship in Peking University, 2018
Special academic scholarship in Peking University, 2017
First Prize in Student's Platform for Innovation and Entrepreneurship Training Program, 2015
First Prize in Jiangsu Province on National Undergraduate Electronics Design Contest, 2014