TAIREN PIAO

Email: tairenpiao@gmail.com ⋄ **Website:** piaotairen.com
LinkedIn: https://www.linkedin.com/in/tairenpiao/

ABOUT ME

I am a Research Engineer at Nota, and my work is mainly focus on AI model acceleration using quantization techniques. Before joining Nota I was an MLE at Xiaomi worked on LLM acceleration and large-scale recommender system optimization. I got my master's degree at Seoul National University, advised by Prof. U Kang, where I did research related to deep learning model compression.

EXPERIENCE

Research Engineer, *Nota, NetsPresso* *Jul. 2023 - Present*
Research on DNNs quantization, deploy various DNN models to various chips and edge devices, and develop key methods for NetsPresso.

Machine Learning Engineer, *Xiaomi, AI Lab* *Oct. 2021 - Jun. 2023*
Xiaomi LLM acceleration using quantization. Before that, I process million-level data and develop the entire ML pipeline for the recommendation system. Optimize the baseline model using ML techniques to improve both accuracy and efficiency.

Research Assistant, *SNU, Data Mining Lab* *Aug. 2019 - Aug. 2021*
Conducted deep learning model compression research, including pruning, quantization, etc. Published an SCI-E paper related to mixed precision quantization method on BERT.

EDUCATION

Seoul National University *Aug. 2019 - Aug. 2021*
M.S. in Computer Science and Engineering
Advisor: *Prof. U Kang*

Harbin Engineering University *Aug. 2015 - Jun. 2019*
B.Eng. in Computer Science and Technology

PUBLICATIONS

[1] SensiMix: Sensitivity-Aware 8-bit Index & 1-bit Value Mixed Precision Quantization for BERT Compression
Tairen Piao, Ikhyun Cho, and U Kang
PLOS ONE (SCIE Journal, 2022)

PROJECTS

Xiaomi

1. **Products Recommendation.** The goal is to discover the high-potential customers who are interested in purchasing products at Xiaomi mall and offer coupons to some of the top-scoring users to increase GMV (Gross Merchandise Volume). My role involves building the entire MLOps flow, optimizing the model, and measuring the performances of different models by doing AB tests. The highlights are as follows:

  - **Million-Level Data Feature Engineering.** Using Spark to process the raw features of users and items and doing feature engineering including feature cleaning, pre-processing, and selection.
  - **AutoML (NAS and HPO).** I designed a DARTs-based NAS method to search for a better recommendation model (search space: generally used CTR prediction modules) and applied Random Search-based HPO to optimize the hyper-parameters, which gains 1M dollar income improvement.

2. **AI Global Advertising.** The goal is to improve the Ads' eCPM (effective Cost Per Mile) and model efficiency of Xiaomi Ad system to improve the customer experience and gain business growth. I mainly focused on optimizing the baseline features and models in Xiaomi global Ad system. The highlights are as follows:
   - **Layer-wise Knowledge Distillation (KD).** To overcome QPS (Query Per Second) bottlenecks of servers caused by the large model size, I applied layer-wise KD to shrink the model size, which reduces half of the model serving time and achieves 5% higher eCPM.
   - **Model Optimization.** I designed a multi-task model combined with a context-aware embedding enhancing method to improve the performance. Besides, I applied different CTR calibration methods to different Ad slots to improve the final eCPM. Overall, the optimized model gains 10% eCPM improvement.

Data Mining Lab @ SNU

1. **BERT Model Compression.** The goal is to compress the pre-trained BERT model to a lightweight one while maintaining its accuracy. We proposed *SENSIMIX* that effectively applies 8-bit and 1-bit mixed precision quantization to the sensitive and insensitive parts of BERT, maximizing the compression rate while minimizing the accuracy drop. We also proposed three novel 1-bit training methods to minimize the accuracy drop and apply XNOR-Count GEMM for 1-bit quantization parts of the model to accelerate the inference speed on Turing NVIDIA GPUs. Experiments show that *SENSIMIX* reduced the original BERT model size by a factor of 8× and shrinking the inference time by around 80% without noticeable accuracy drop.
   - **Mixed-precision Quantization.** For more specific methodology and experimental results, please check the paper. To make the compressed model inference on real edge devices, I deployed the quantized model to Android phones based on PyTorch Mobile.
   - **Other Compression Methods.** I applied various pruning, KD, and factorization methods on the BERT model, and also achieved good accuracy and inference speed.

TEACHING EXPERIENCE

Teaching Asssistant

- SK-Univ, SK . . . . *Aug 2020*
- Data Structures (M1522.000900), SNU . . . . *Fall 2020*
- Introduction to Data Mining (M1522.001400), SNU . . . . *Spring 2020*

PATENTS

1. Tairen Piao, “Layer-Wise Knowledge Distillation Method for Compressing CTR Prediction Models.”, CN-Registration (2022)

2. Tairen Piao, “Auto Feature Selection Method for CTR Prediction Models based on Power Law Data Distribution.”, CN-Registration (2022)

3. Tairen Piao, Ikhyun Cho, and U Kang, “Quantization Method For Transformer Encoder Layer based on the Sensitivity of the Parameter and Apparatus Thereof”, KR-Registration No. 10-2020-0183411 (2020)

SKILLS

Programming Language: C, C++, Python, CUDA, Java, SQL, Shell
Frameworks and Tools: PyTorch, TensorFlow, TFLite, Pandas, Spark, Matplotlib, Git, Docker, TensorRT, onnx
Language: Korean (Advanced), English (Advanced), Mandarin (Native)