

TAIREN PIAO

Email: tairenpioa@gmail.com ◇ **Website:** piaotairen.com

Linkedin: <https://www.linkedin.com/in/tairenpioa/>

EXPERIENCE

Machine Learning Engineer

Oct. 2021 - Present

Xiaomi AI Lab @ Xiaomi, Beijing, China

Key works: *AI Marketing, AI Advertising, Large-Scale Data Processing*

Research Assistant

Aug. 2019 - Aug. 2021

Data Mining Lab @ SNU, Seoul, Republic of Korea

Key works: *Deep Learning, Model Compression, NLP, Data Mining*

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

GPA: 3.8 / 4.0

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. **AI Marketing.** The goal is to discover the high-potential customers who desire to buy products at Xiaomi mall. We predict buy possibility score of users for different products and rank the user scores. Then, the Operation Managers operate marketing activities (e.g., push coupons) to some of the top people from these users, to improve GMV (Gross Merchandise Volume). As a Machine Learning Engineer, I build the entire MLOps flow and optimize the model (usually deep learning model) to improve the performance. After the model optimization, we measure the performances of different models by doing AB tests and push the better model online. The main technical works are as follows:
 - **Million-Level Data Feature Engineering.** Processing the million-level user data using Spark. For features, I create the raw features of users and items and do feature engineering including feature cleaning, feature pre-processing, and feature selection. Furthermore, to efficiently predict and rank the user scores, I also use Spark to make model inference.
 - **Model Optimization.** Understanding the business scene and optimizing the model using both technical and business knowledge. I applied AutoML methods on the model. For instance, I applied Random Search-based HPO to optimize the hyper-parameters and DARTs-based NAS to search the better model (search space: generally used CTR prediction modules) to improve the offline metrics, such as AUC and Recall.

2. **AI Global Advertising.** The goal is to optimize the features and models of Xiaomi global Ad system to improve the customer experience and gain business growth. Similar to AI Marketing, I process million-level data, do feature engineering, train and optimize the model, measure the performances by doing AB tests. I mainly focus on improving the Ads' eCPM (effective Cost Per Mile) by optimizing the baseline deep CTR prediction models, such as Wide & Deep, DeepFM, and DCN. The main technical works are as follows:

- **Model Compression.** There are QPS (Query Per Second) bottlenecks for servers when making inference using large model. To tackle the problem, I applied layer-wise KD (Knowledge Distillation) to shrink the model size while keeping its accuracy, and my model reduces half of the model serving time and even achieves higher eCPM.
- **Model Optimization.** To improve the performance, I applied entire space multi-task model and optimize the model by applying context-based Embedding methods. Besides, I applied different CTR calibration methods to different Ad slots to improve the final eCPM.

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. Heavy language models like BERT have problems of large memory cost and long inference time. To tackle this, we propose *SENSIMIX* that effectively applies 8-bit index quantization and 1-bit value quantization to the sensitive and insensitive parts of BERT, maximizing the compression rate while minimizing the accuracy drop. We also propose three novel 1-bit training methods to minimize the accuracy drop and apply FP16 general matrix multiplication (GEMM) and XNOR-Count GEMM for 8-bit and 1-bit quantization parts of the model to accelerate the inference speed on Turing NVIDIA GPUs. Experiments on GLUE tasks show that *SENSIMIX* compresses the original BERT model to an equally effective but lightweight one, reducing the model size by a factor of $8\times$ 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 Assistant

- 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)