Tairen Piao

Email: tairenpiao@gmail.com ◇ **Website:** tairenpiao.github.io
LinkedIn: https://www.linkedin.com/in/tairenpiao/

ABOUT ME

I am an AI Research Engineer at Nota Inc., specializing in developing hardware-aware AI model optimization and compression methods, especially focusing on Large Language Model (LLM) quantization. Previously, I was a Machine Learning Engineer at Xiaomi, working on LLM acceleration and building large-scale recommender systems. I received a master's degree at Seoul National University, advised by Prof. U Kang, where I conducted research related to deep learning model compression.

EXPERIENCE

AI Research Engineer, *NetsPresso Core Research, Nota Inc.* *Jul. 2023 - Present*
1) Conducting research on Generative AI model quantization, developing key quantization methods and toolchains for the Nota NetsPresso platform. 2) Training various computer vision models and deploying the models to diverse chips and edge devices. 3) Handling various deep learning IR such as ONNX and TFLite including IR conversion, quantization, and graph optimization.

Machine Learning Engineer, *AI Lab, Xiaomi* *Oct. 2021 - Jun. 2023*
1) Applied advanced quantization methods (e.g.,INT8 CUTLASS kernel, SmoothQuant) to accelerate Xiaomi LLM. 2) Processed million-level data using Spark, developed the entire machine learning (ML) pipeline for the recommender system and optimized the baseline model using ML techniques to improve both accuracy and efficiency.

Research Assistant, *Data Mining Lab, SNU* *Aug. 2019 - Aug. 2021*
1) Conducted deep learning model compression research, including pruning, quantization, and decomposition, mainly targeting the pre-trainining language models. 2) Published an SCI-E paper (SensiMix) related to the mixed precision quantization method on BERT.

EDUCATION

Seoul National University *Aug. 2019 - Aug. 2021*
M.S. in Computer Science and Engineering
GPA: 4.12 / 4.3
Advisor: *Prof. U Kang*

Harbin Engineering University *Aug. 2015 - Jun. 2019*
B.Eng. in Computer Science and Technology
GPA: 3.7 / 4.0

PUBLICATIONS

[2] **EdgeFusion: On-Device Text-to-Image Generation**
Thibault Castells, Hyoung-Kyu Song, Tairen Piao, Shinkook Choi, Bo-Kyeong Kim, Hanyoung Yim, Changgwun Lee, Jae Gon Kim, Tae-Ho Kim
CVPR 2024 (EDGE Workshop)

[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 (SCI-E Journal, 2022)

SELECTED PROJECTS

Nota Inc.

1. **NetsPressor Quantizer.** Develop NetsPresso platform Quantizer.
   - **Advanced Quantization** Quantize eager mode models (e.g., Hugging Face, PyTorch), and apply advanced quantization algorithms, such as AWQ, AutoRound, and GPTQ.
   - **Graph Quantization** Apply quantization to all model graph including non-GEMM operators.
2. **EdgeFusion.** The goal was to apply quantization to the stable diffusion (SD) model to accelarate the inference speed and to reduce the memory usage, enabling its deployment on the Samsung Exynos NPU while maintaining accuracy.
   - **Mixed Precision Quantization (MPQ).** In the SD model, layers such as softmax, swish, and LayerNorm are extremely sensitive to W8A8 quantization. To tackle this, We apply W8A16 quantization to softmax and FP16 to swish and LayerNorm layers.
   - **Deploy to Samsung Exynos NPU.** By utilizing MPQ and hardware model-level tiling, the model is successfully deployed on the Samsung Exynos NPU. The EdgeFusion model can generate images in 0.7 seconds.
3. **ONNX2TFLite Converter Optimization.** The goal was to optimize the ONNX2TFLite tool to support more ONNX models.
   - **NCHW to CNHW.** The ONNX format uses NCHW (batch, channel, height, width) ordering to represent the model, while TFLite uses CNHW. During conversion, many operators struggled with this format difference. I fixed various corner cases to support more models such as PIDNet, Segformer.
   - **Operator Parameter Optimization.** During operator conversion, operator parameters need to be copied and mapped to the correct positions in TFLite. I updated the algorithm to support more cases.
4. **MPQ for TFLite Vision Models.** The goal was to recover the full INT8 quantized TFLite model accuracy using MPQ.
   - **Improve the accuracy of full INT8 quantized models.** The INT8 TFLite uses a default min-max quantization scheme. Models like YOLOv5 have severe quantization accuracy loss due to many concatenation layers in the detection head of the model.
   - **Mixed Precision Quantization.** Applied the FP32 precision to the detection head and full INT8 to the remaining part of the model to improve accuracy. For YOLOv5n, the mAP50 of the quantized model is incresed from 0.0 to 0.436.

Xiaomi

1. **Mall Products Recommendation.** The goal was to discover the high-potential customers who are interested in purchasing products at the Xiaomi online Mall and offering coupons to some of the top-scoring users to increase Gross Merchandise Volume (GMV). 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 million-level 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 Click-Through Rate prediction modules) and applied Random Search-based HPO to optimize the hyper-parameters, which gains 1M dollars income improvement.

2. **AI Advertising.** The goal was to improve the Ads' effective Cost Per Mile (eCPM) and model efficiency of the Xiaomi Ad system to improve the customer experience and gain business growth. I mainly focused on optimizing the baseline features and models in the system. The highlights are as follows:
   - **Layer-wise Knowledge Distillation (KD).** To overcome the Query Per Second (QPS) 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 inference time and even 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. **SensiMix (BERT 1&8-bit Quantizaion).** The goal was to compress the pre-trained BERT model to a lightweight one while maintaining its accuracy. We propose *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 propose three novel 1-bit training methods to minimize the accuracy drop and apply XNOR-Count GEMM to 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 a noticeable accuracy drop.
   - **SensiMix accepted by PLOS One 2022.** More specific methodology and experimental results can be found in the paper.
   - **Deploy to Android Devices.** To make the compressed model inference on real edge devices, I deployed the SemsiMix model to Android phones based on the PyTorch Mobile framework. The 1-bit XNOR GEMM kernel is also ported to the Android platform. The kernel implementation is open-sourced, which can be found in *https://github.com/tairenpiao/XNOR-popcount-GEMM-PyTorch-CPU-CUDA*
2. **BERT Model Compression.** I was a research assistant at SNU Data Mining Lab focused on BERT model compression. Besides quantization, I also 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, "A Cross-Task Knowledge Distillation Model for Multi-Task CTR Prediction, CN-Registration (2023)
2. Tairen Piao, "Layer-Wise Knowledge Distillation Method for Compressing CTR Prediction Models.", CN-Registration (2022)
3. Tairen Piao, "Auto Feature Selection Method for CTR Prediction Models based on Power Law Data Distribution.", CN-Registration (2022)

4. 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, Shell, SQL
Frameworks, tools & IR: PyTorch, TensorFlow, TFLite, ONNX, TensorRT
Spark, Pandas, Optuna, Matplotlib, AIMET
Development: Linux, Git, Docker
Language: Korean (Advanced), English (Advanced), Mandarin (Native)