Yang Bai

Email : ybai62868@gmail.com
Technical Blog
Github: ybai62868, Mobile : +86-176-0074-9506

Education

- **The Chinese University of Hong Kong** N.T. Hong Kong
  Ph.D. in Computer Science: Compiler and Deep Neural Network Design Automation *Aug. 2020 – Present.*
- **Chinese Academy of Sciences** Beijing, China
  Master of Science in Computer Science: Machine Learning Systems and Computer Vision *Aug. 2017 – July. 2020*
- **Xidian University** Shaanxi, China
  Bachelor of Engineering in Electronics and Communication *Aug. 2013 – July. 2017*

Research Interests

Machine Learning System Compiler Optimization Computer Architecture Design Automation

Publications

- **Conference Paper**:
  1. ACPNet: Anchor-Center based Person Network for Human Pose Estimation and Instance Segmentation. **Y Bai**, W Wang. *IEEE International Conference on Multimedia and Expo (ICME) 2019.*
  2. UHRSNet: A Semantic Segmentation Network Specifically for Ultra-High-Resolution Images. L Lei, M Li, **Y Bai**, and W Wang. *IEEE International Conference on Pattern Recognition (ICPR) 2021.*
  3. Global-Local Attention Network for Semantic Segmentation in Aerial Images. M Li, L Shan, **Y Bai**, W Wang, B Luo, S Chen, K Lv. *IEEE International Conference on Pattern Recognition (ICPR) 2021.*
  4. AutoGTCO: Graph and Tensor Co-Optimize for Image Recognition with Transformers on GPU. **Y Bai**, X Yao, Q Sun, B Yu. *IEEE International Conference on Computer-Aided Design (ICCAD) 2021.*
  5. A High-Performance Accelerator for Super-Resolution Processing on Embedded GPU. W Zhao, Q Sun, **Y Bai**, W Li, H Zheng, N Jiang, J Lu, B Yu, MDF Wong. *IEEE International Conference on Computer-Aided Design (ICCAD) 2021.*
  6. Fast and efficient DNN deployment via deep Gaussian transfer learning. Q Sun, C Bai, T Chen, H Geng, X Zhang, **Y Bai**, B Yu. *IEEE International Conference on Computer Vision (ICCV) 2021.*
  7. PCL: Proxy-based Contrastive Learning for Domain Generalization. X Yao, **Y Bai**, X Zhang, Y Zhang, Q Sun, R Chen, R Li, B Yu. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2022.*
  8. GTuner: Tuning DNN Computations on GPU via Graph Attention Network. Q Sun, X Zhang, H Geng, Y Zhao, **Y Bai**, H Zheng, B Yu. *IEEE Design Automation Conference (DAC) 2022.*
  9. AutoGraph: Optimizing DNN Computation Graph for Parallel GPU Kernel Execution. Y Zhao, Q Sun, Z He, **Y Bai**, B Yu. *AAAI Conference on Artificial Intelligence (AAAI) 2023.*
  10. DiffPattern: Layout Pattern Generation via Discrete Diffusion. Z Wang, Y Shen, W Zhao, **Y Bai**, G Chen, F Farnia, B Yu. *IEEE Design Automation Conference (DAC) 2023.*
  11. ALCOP: Automatic Load-Compute Pipelining in Deep Learning Compiler for AI-GPUs. G Huang, **Y Bai**, L Liu, Y Wang, B Yu, Y Ding, Y Xie. *Conference on Machine Learning and Systems (MLSys) 2023.*
- **Journal Paper**:
  1. A High-Performance Accelerator for Super-Resolution Processing on Embedded GPU. W Zhao, **Y Bai**, Q Sun, W Li, H Zheng, N Jiang, J Lu, B Yu, MDF Wong. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. (TCAD)*

Research and Development Experiences

- **SmartMore (Heterogeneous Computing Center)**
  Research Intern - Compilation and Acceleration for AI applications *Jul. 2020 - Mar. 2021.*
  - **OCR - GPU Acceleration**:
    Developed the codebase for OCR application on NVIDIA GPUs (TX2, AGX, 2080Ti).
    Including PyTorch → ONNX → TensorRT, Image pre-processing and post-processing (BGR → RGB, HWC → CHW, Normailization, FP16, FP32, Int8, Custom Plugin). Got **9.753x** speed-up in text detection and **10.411x** speed-up in text recognition. More details can be found in this **talk**.
  - **Ultra HD Video - AI Solutions**:
    Deployed the image super-resolution model (Lapar) on GPU platforms (TX2, NX, AGX) and Atlas. Including Int8, FP16, FP32 format ranging from 360P → 720P, 540P → 1080P, 720P → 2K, 1080P → 4K.

- **ByteDance AI Lab (Machine Learning Systems Group)** Advisor: Yunfeng Shi
  Software Development Intern - Deep Learning Compiler *Jan. 2020 - Jun. 2020*
  - **Deep Learning Compiler Survey**:
    Dived into Deep Learning Compiler including TVM, Glow, TC and Halide. Made an in-depth comparison between them. More details can be found in this talk.
  - **ResNet-50 verification for MVP (ASIC)**:
    Modified a deep learning compiler (Glow) to extract the input feature map, output feature map and weights of each layer to verify the arithmetic logic units of MVP.
  - **Quantization Tool Chain for MVP** :
    Developed profile-guided quantization codebase which can observe execution during inference to estimate the possible numeric range for each stage of the neural network. Devised visualization function for each layer by graphviz. Testing code is open source in this repo.
- **Cornell University (Computer Systems Lab)** Advisor: Zhiru Zhang
  Research Intern - Software Defined Heterogeneous Computing *Jun. 2019 - Oct. 2019*
  - **OpenCL Backend Development for HeteroCL**:
    Developed the Xilinx & Intel OpenCL backend for TVM-inspired HeteroCL. Implemented critical compiler backend optimization, e.g., loop unrolling, loop pipelining and partition for Xilinx OpenCL backend and implemented arbitrary precision integers for Intel OpenCL backend. Implemented the whole pipeline from Python-based domain-specific language to FPGA-targeted compilation flow (kernel code generation and runtime system). Tested samples *(KNN-DigitRec, GEMM, K-means, LeNet, Smith-Waterman Sequencing)* for HeteroCL OpenCL backend on software and hardware simulation. All of the code is open source in this repo.
  - **AWS-F1 Tutorial for HeteroCL**:
    Wrote tutorials for running *KNN-DigitRec* example on AWS-F1 using HeteroCL. Designed a new target backend for AWS development, combined HeteroCL Vivado HLS C++ code and host file based on Rosetta automatically generate host and wrapper files for design automation. More details can be found in this repo.
  - **Object Detection System Design for Drones (GPU Platform)**:
    Used CornerNet-Lite (Anchor Free Object Detection) and DJI dataset as baseline. Changed the original structures/layers to increase the receptive fields, included feature map fusion, and balanced the positive and negative examples during training to capture small object. Finally got **0.87** mIoU (big-model) on validation dataset. Designed a flexible and multi-module framework for small object detection based on PyTorch for UAV.
- **Horizon Robotics (Algorithm R&D Group)** Advisor: Lichao Huang
  Software Development Intern - Computer Vision *Apr. 2018 - Apr. 2019*
  - **Multi-Task Vision System Development for Smart IoT**:
    Optimized multiple metrics of the previous multi-task vision system that can do object detection, instance segmentation, and keypoint detection simultaneously by designing different backbones and head branches. Compared with design from Mask R-CNN, our system achieved **0.455/0.465** mAP on COCO instance segmentation task and **0.631/0.647** mAP on COCO keypoint task.
  - **PoseTrack System Development**:
    Implemented Realtime Multi-Person 2D Pose Estimation System (Bottom-up) in PyTorch. Implemented Dev-Head Pose Estimation system (top-down) in MXNet/Gluon. Designed and Implemented an efficient and effective human pose estimation & tracking system based on optical flow in MXNet/Gluon. Implemented the whole pipeline from person detection to single person pose estimation to multi-object tracking in successional video frames (3-stage). Documented a Gluon tutorial including basic CNN and RNN models.

Honors & Awards

- **2019**: Pacemaker to Merit Student in Chinese Academy of Sciences
- **2013-2017**: Four times Outstanding School-level scholarship (first prize & second prize)
- **2013-2017**: Top Ten Outstanding Undergraduate in Xidian University
- **2015 & 2016**: Second prize in the MCM/ICM Contest
- **2016**: Silver medal in the 2016 ACM-ICPC Shaanxi Province Contest
- **2015**: Silver medal in the 2015 ACM-ICPC Asia Dalian Region-Invitational Contest
- **2015**: Bronze medal in the 2015 ACM-ICPC Asia Shenyang Regional Contest
- **2015**: Aerospace CASC Outstanding Undergraduate Scholarship