Andre Xian Ming Chang

Andrechang | linkedin | scholar | mysite | andrechang67@gmail.com | 7655431912

Experience

Principal Machine Learning Engineer Micron, Seattle WA 09/2019 - now

- Lead Machine Learning software team to deliver software stack (SDK and compiler) to support Micron Deep Learning Accelerator (MDLA)
- Drive ML software roadmap development for MDLA to achieve model coverage, performance, ease of use and backward compatibility
- Integrate ML frameworks into MDLA compiler to run models on multiple hardware accelerator versions in ASIC and FPGA
- Interact with hardware and ML engineers and customers to gather product requirements and architect and develop new features to MDLA software stack
- Implement SDKs, compiler support, examples and tutorials to develop ML software ecosystem and for customer facing projects

Co-Founder and Lead Compiler Engineer (Acquired by Micron) FWDNXT 09/2017 - 05/2019

- Architected and implemented SDKs and compiler to enable deployment of customer DNN models on custom hardware accelerator
- Developed compiler optimizations to run ML model on custom DNN accelerators
- Assisted design of DNN accelerators through DNN profiling and creation of tools

Research Assistant e-Lab Purdue University, IN 09/2014 – 08/2017

- Wrote library to accelerate deep learning on mobile phones using openGL
- Designed a low power hardware accelerator for recurrent neural networks using FPGA

Research Intern Sonoscan, IL 05/2015 – 08/2015

- Implemented firmware for the transceiver, Android App to communicate with mobile phones and program to communicate with PC

Education

Ph.D. - Purdue University West Lafayette, IN, US 09/2019

Electrical and Computer Engineering: DNN Hardware Accelerators and compilers (3.88/4.0)

Master of Science - Purdue University West Lafayette, IN, US 05/2016

Electrical and Computer Engineering: RNN Hardware Accelerators (3.8/4.0)

Bachelor - Universidade Tecnológica Federal do Paraná (UTFPR), Brazil 07/2014

Electronic Engineering (0.87/1)

Skills

AI: CNN, RNN, GAN, Transformer, ViT, RL, GNN, Private ML, Diffusion, NAS
Coding: C, C++, Python, Verilog, Pytorch, TVM, React, JScript, Android, ONNX, Torchdynamo, Solidity, Tensorflow, Streamlit, Linux, Git, JIRA, Confluence, Bitbucket, Kubernetes, Docker, Doxygen, Sphinx
Languages: English Fluent and Portuguese Fluent

Publications

- **A.X.M. Chang**, P. Khopkar, B. Romanous, A. Chaurasia, E. Culurciello, et.al. "*Reinforcement Learning Approach for Mapping Applications to Dataflow-Based Coarse-Grained Reconfigurable Array.*" Arxiv 2022.
- **A.X.M. Chang**, A. Zaidy, M. Vitez, L. Burzawa E. Culurciello. "*Deep neural networks compiler for a trace-based accelerator.*" JSA 2020.
- **A.X.M. Chang**, A. Zaidy, L. Burzawa E. Culurciello. "*WIP: Deep Neural Networks compiler for a trace-based accelerator.*" LCTES 2018.
- **A.X.M. Chang**, A. Zaidy, E. Culurciello. "*Efficient compiler code generation for Deep Learning Snowflake co-processor.*" EMC2 Workshop 2018 co-located with ASPLOS 2018.
- **A.X.M. Chang**, A. Zaidy, E. Culurciello. "*Compiling Convolutional Neural Networks for FPGA Accelerator.*" OpenSuCO Workshop 2017 co-located with Super Computing 2017.
- **A.X.M. Chang**, A. Zaidy, V. Gokhale, E. Culurciello. "*Compiling Deep Learning Models for Custom Hardware Accelerators.*" arXiv preprint https://arxiv.org/pdf/1708.00117.pdf (2017).
- **A.X.M. Chang**, A. Zaidy, V. Gokhale, E. Culurciello. "*Generating instructions for deep learning custom hardware.*" Poster presentation at Computational Science and Engineering Student Conference - CSESC 2017.
- **A.X.M. Chang**, E. Culurciello. "*Hardware Accelerators for Recurrent Neural Networks on FPGA.*" In press - ISCAS May 27 2017.
- V. Gokhale, A. Zaidy, **A.X.M. Chang**, E. Culurciello. "*Snowflake: an Efficient Hardware Accelerator for Convolutional Neural Networks.*" In press - ISCAS May 27 2017.
- A. Zaidy, **A.X.M. Chang**, V. Gokhale, E. Culurciello. "*Snowflake: A Hardware Accelerator for Convolutional Neural Networks.*" Poster presentation at Workshop: Hardware and Algorithms for Learning On-a-chip (HALO) - ICCAD November 10 2016.
- **A.X.M. Chang**, B. Martini, E. Culurciello. "*Recurrent Neural Networks Hardware Implementation on FPGA.*" arXiv preprint arXiv:1511.05552 (2015).

Patent applications

- *Non-Uniform Splitting of a Tensor for Shuffling* US-2021141203.
- *Reinforcement Learning workload mapping to Streaming Engine* US-2022141658.
- *Evolutionary Imitation Learning* US Patent App. 17/124,389.
- *Implement the computation of an artificial neural network using multiple deep learning accelerators* US Patent App. 17/092,038.
- *Compiler with an artificial neural network to optimize instructions generated for execution on a deep learning accelerator of artificial neural networks* US Patent App. 17/092,040.
- *Runtime optimization of computations of an artificial neural network compiled for execution on a deep learning accelerator* US Patent App. 17/092,044.
- *Discovery of hardware characteristics of deep learning accelerators for optimization via compiler* US Patent App. 17/092,033.
- *Compiler configurable to generate instructions executable by different deep learning accelerators from a description of an artificial neural network* US Patent App. 17/092,013.
- *Deep learning accelerators with configurable hardware options optimizable via compiler* US Patent App. 17/092,023.
- *Deep Neural Networks Compiler for a Trace-Based Accelerator* US Patent App. 17/003,476.
- *Inference Engine Circuit Architecture* US Patent App. 16/833,610.
- *Hardware Accelerator for Convolutional Neural Networks and Method of Operation Thereof* US Patent App. 15/990,365.