Zhanghao Wu

UNDERGRADUATE IN COMPUTER SCIENCE

Mobile: +1 (857) 228-4863 | **Email:** zhanghao.wu@outlook.com | **Homepage:** zhanghaowu.me

RESEARCH INTERESTS

Deep learning, especially Natural Language Processing, Speech and Efficient Machine Learning.

EDUCATION

Shanghai Jiao Tong University (SJTU) *Shanghai, China*
Bachelor of Engineering in Computer Science at ACM Honors Class, Zhiyuan College *Sep. 2016 - Jun. 2020*
- **ACM Honors Class** is an elite CS program for top 5% talented students in Computer Science Department.
- GPA: **92.47/100 (4.01/4.3)** | **Ranking: 2nd/37** (in ACM Hornors Class).
- Advisors: Prof. Kai Yu, Prof. Yanmin Qian and Prof. Yong Yu.

Massachusetts Institute of Technology (MIT) *Cambridge, USA*
Research Assistant at HanLab, Microsystems Technology Laboratories *Jun. 2019 - Jan. 2020*
- Advisor: Prof. Song Han.

PUBLICATIONS

FaceMix: Privacy-Preserving Facial Attribute Classification on the Cloud
Zhanghao Wu*, Zhijian Liu*, Ligeng Zhu, Chuang Gan and Song Han ***CVPR 2020*** *(under review)*

Efficient Transformer for Mobile Application
Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin and Song Han [Paper][Slides] ***ICLR 2020*** *(under review,* ***6.7/8****)*

Data Augmentation using Variational Autoencoder for Embedding based Speaker Verification
Zhanghao Wu, Shuai Wang, Yanmin Qian and Kai Yu [Paper] [Slides] ***Interspeech 2019 (oral)***

On-Device Image Classification with Proxyless Neural Architecture Search and Quantization-Aware Fine-tuning
Han Cai, Tianzhe Wang, **Zhanghao Wu**, Kuan Wang, Ji Lin and Song Han [Paper] ***ICCV 2019 workshop***

RESEARCH EXPERIENCE

HanLab, Microsystems Technology Laboratories, MIT *Cambridge, USA*
Research Assistant, Advised by Prof. Song Han. *Jun. 2019 - Present*

Project 1: Efficient Transformer for Mobile Application
- **Goal:** Enhance machine translation on resource-constrained conditions by re-designing the transformer architecture.
- Defined mobile settings for the NLP tasks, analyzed the bottleneck of the computational intensive transformer architecture, and proposed a novel efficient primitive (LSRA) motivated by specialization to reduce the computation and model size of the model.
- Experimented on three machine translation datasets (IWSLT De-En, WMT En-De and WMT En-Fr) and achieved better performance than the original transformer architecture, as well as the AutoML based Evolved Transformer under mobile settings.
- Raised people's awareness about the importance of design insights and concerns over the massive cost of AutoML.
- Submitted a paper to the *International Conference on Learning Representations* **(ICLR 2020)**, and received average score of **6.7/8**.

Project 2: Privacy-Preserving Inference on the Cloud
- **Goal:** Mediate between the resource-constrained edge devices and the privacy-invasive cloud servers, protecting sensitive data.
- Analyzed limitations for both the edge and cloud inference, proposed efficient encryption and decryption methods by encouraging the linearity of neural networks and designed GAN-based attacking methods to evaluate our methods and previous work.
- Experimented our method on two popular facial attribute classification tasks, CelebA and LFWA, protecting both the input and output privacy. Our method outperforms all previous encryption techniques over effectiveness and efficiency.
- Further experimented our proposed method on other modalities, language, and speech, and achieved impressive results.
- Submitted a paper to the *IEEE Conference on Computer Vision and Pattern Recognition* **(CVPR 2020)**.

Project 3: Hardware-Aware Transformer
- **Goal:** Search for efficient transformer architectures specialized for various hardwares.
- Analyzed the correlation between model size, Mult-Adds and latency. Concluded that latency is highly related to specific hardware and is a more important metric related to the real-time experience for the user.
- Designed and implemented a HyperNet based Neural Architecture Search (NAS) for the hardware-aware transformer models, and experimented on two datasets and different hardwares, comparing to the model size based NAS algorithm (evolved transformer).
- Prepared a paper targetted for the Annual Conference of the Association for Computational Linguistics **(ACL 2020)**.

Project 4: On Device Image Classification

- **Goal:** Deploy image classification models on hardware devices with strict efficiency constraints.
- Participated in the CVPR'19 Low-Power Image Recognition Challenge and Visual Wake Words Challenge on designing the low-latency but high-accuracy deployable neural networks and ranked 1st in both competition among the academic groups.
- Searched for an efficient neural architecture directly with latency constraints using ProxylessNAS, quantized the model to 8-bit and achieved comparable latency but much higher accuracy than MobileNet V2 on the Google Pixel 2 mobile phone.
- Published a paper to the workshop of *IEEE International Conference on Computer Vision* **(ICCV 2019 workshop)**.

SpeechLab, Computer Science and Technology Department, SJTU — *Shanghai, China*
Undergraduate Researcher, Advised by Prof. Kai Yu and Prof. Yanmin Qian. — *Jul. 2018 - Jun. 2019*

Project 1: Data Augmentation for Robust Speaker Verification

- **Goal:** Improve the data efficiency and the robustness of speaker verification systems with generative models.
- Established a variational auto-encoder based data augmentation method improving the robustness of speaker verification systems against noises and reverberations in the real-world scenarios.
- Experimented on *x*-vector (DNN based) and *i*-vector (statistic based) speaker representations and achieved state-of-the-arts.
- Published a paper to the Annual Conference of the International Speech Communication **(Interspeech 2019 oral)**.

Project 2: Speaker Representations Enhancement

- **Goal:** Combining the orthogonal information captured in neural network based and statistic based speaker representation.
- Assisted the training process of DNN-based framework *x*-vector with the statistic based representations *i*-vector for speaker verification and achieved better performance than original separated systems.
- Released the first Deep Canonical Correlation Analysis (DeepCCA) implementation in Pytorch on GitHub.

Honors & Awards

Scholarships

- **Chinese National Scholarship**, highest honor for undergraduates, **top 0.2%** nation wide. *2018, 2019*
- **Fan Hsu-Chi Chancellor's Scholarship**, **top 0.1%** of 17,000 students in SJTU. *2017*
- **Zhiyuan Honorary Scholarship**, **top 5%** of 17,000 students in SJTU. *2017, 2018*

Competitions

- **1st place**, in Visual Wake Words (VWW) Challenge of CVPR'19. *2019*
- **3rd place**, in Low Power Image Recognition Challenge of CVPR'19 (**1st place** for academic participants). *2019*
- **Outstanding Winner**, in Mathematical Contest in Modeling (**top 0.5%** out of 8,800 international participants). *2017*

Selected Projects

- **Visual Wake Words**, the code for the CVPR'19 Visual Wake Words challenge. (Won **1st place**). [Code] *May. 2019*
- **DeepCCA**, the first implementation of Deep Canonical Correlation Analysis in PyTorch. (Got **41 stars**). [Code] *Dec. 2018*
- **Quantum Shor Algorithm**, a simulation of Qshor algorithm in Q# and Python. (Course project, **99/100**). [Code] *Jul. 2018*
- **Mx Compiler**, a compiler for a C-alike language Mx to NASM. Faster than GCC-O1. (Course project, **98/100**). [Code] *Jun. 2018*
- **RISC-V CPU**, implementation of 5-stage pipeline CPU in Verilog. FPGA supported. (Course project, **100/100**). [Code] *Jan. 2018*

Teaching & Activities

- **Teaching Assistant of CS152: Programming**, gave an introduction to programming and designed homework. [Link] *2018*
- **Teaching Assistant of MS208: Compiler**, designed projects and assisted in grading and answering questions. [Link] *2019*
- **Contributor of Popular Repositories**, contributed to the PyTorch/fairseq and wookayin/gpustat. [fairseq|gpustat] *2019*
- **Conference Reviewer**, served as the secondary reviewer of the AAAI Conference on Artificial Intelligence **(AAAI 2020)**. *2019*
- **Presentation at the Efficient Deep Learning Workshop**, presented on efficient transformer at the workshop, MIT. *2019*
- **Deep Learning Book Translation**, participated in the translation of *"Reinforcement Learning: an Introduction", by Richard S. Sutton and Andrew G. Barto*. The translation group was led by Prof. Kai Yu. (Chinese edition is now published). [Link] *2019*

Technical Skills

Programming languages: C/C++, Python, Java, MATLAB, Verilog-HDL, C#, Q#, Liquid.
Deep Learning Packages: PyTorch, Keras, TensorFlow, scikit-learn.
Scientific Softwares: Unity, Mathematica, Origin.