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 Honors 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 *Jul. 2019 - Jan. 2020*

- Advisor: Prof. Song Han.

PUBLICATIONS

Privacy-Preserving Cloud Edge Inference
Zhanghao Wu*, Zhijian Liu*, Ligeng Zhu and Song Han ***CVPR 2020*** *(targeted)*

Efficient Transformer for Mobile Application
Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin and Song Han [Paper] ***ICLR 2020*** *(submitted)*

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. *Jul. 2019 - Present*

Project 1: Efficient Transformer for Mobile Application

- **Goal:** Enable Natural Language Processing tasks for computational limited condition by re-designing the transformer architecture.
- Defined the mobile settings for the natural language processing tasks, analyzed the computational intensive transformer architecture, and proposed a novel efficient primitive (LSRA) with specialized information extractor to reduce the MultAdds of model inference.
- 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 Neural Architecture Search based Evolved Transformer under mobile settings.
- Wrote and submitted a paper to the *International Conference on Learning Representations* **(ICLR 2020)**.

Project 2: Privacy-Preserving Cloud-Edge Inference

- **Goal:** Mediate between the resource-constrained edge devices and the privacy-invasive cloud servers, protecting sensitive user data.
- Analyzed the limitation of both the local and cloud inference, proposed a efficient encryption and decryption methods using the linearity of the neural network we observed and designed attacking methods to evaluate our methods and previous work.
- Experimented the proposed method on three modalities, including vision, language and speech, and achieved impressive results over all previous work on all the modalities, indicating the effectiveness, efficiency and generalizability of method.
- Wrote a paper target for the *IEEE Conference on Computer Vision and Pattern Recognition* **(CVPR 2020)**.

Project 3: On Device Image Classification

- **Goal:** Deploy image classification models on hardware devices with strict efficiency constraints.
- Participated in the CVPR'19 LPIRC and VWW competition for designing the low-latency but high-accuracy deployable neural networks and ranked first in both competition among the academic groups.
- Searched for a 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.
- Wrote and 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 improve state-of-the-arts results.
- Wrote and published a paper to Annual the 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.
- Released the first Deep Canonical Correlation Analysis (DeepCCA) implementation in Pytorch on GitHub.
- Assisted the training of DNN-based framework with the statistic based representations and achieved better performance than original speaker verification systems.

Course (CS087): Computer Science: Advanced Topics *Shanghai, China*
Course Project, Advised by Prof. John Hopcroft. *Apr. 2018 - Jun. 2018*

Project 1: Adversarial Robustness Exploration

- **Goal:** Investigate the attacking methods for classification networks and design a general defense against adversarial examples.
- Proposed a local Lipschitz regularization term to constraint the complexity of the boundary between different classes.
- Experimented on MNIST and CIFAR-10 and proved that the regularization can improve adversarial robustness of small neural networks.

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:

- **First place**, in Visual Wakeup Words (VWW) Challenge of CVPR'19. *2019*
- **Third place**, in Low Power Image Recognition Competition of CVPR'19 (first 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 **first place**). [Code] *May. 2019*
- **DeepCCA**, the first implementation of Deep Canonical Correlation Analysis in Pytorch. (Got **40 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 Experience

- **Compiler (MS208)**, Teaching Assistant *2019*
- **C++ Programming (CS152)**, Teaching Assistant *2018*

Extracurricular Activity

Participated in Deep Learning Books Translation, SJTU

- Richard S. Sutton and Andrew G. Barto, **Reinforcement Learning: an Introduction**, Chinese edition is published.