Zhanghao Wu

Mobile: +86 137-369-39077 | **Email:** wz.wzh@sjtu.edu.cn | **Homepage:** zhwu.me

Research Interests

Efficient deep learning, especially for Natural Language Processing and Speech, and computer architecture.

Education

Shanghai Jiao Tong University (SJTU) *Shanghai, China*
Bachelor of Engineering in CS 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: 92.5/100 (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 *Jun. 2019 - Jan. 2020*

- Served as a visiting undergraduate student for research purpose without taking courses.
- Advisor: Prof. Song Han.

Publications

HAT: Hardware-Aware Transformers for Efficient Neural Machine Translation
Zhanghao Wu*, Hanrui Wang*, Zhijian Liu*, Han Cai, Ligeng Zhu, Chuang Gan and Song Han ***ACL 2020*** *(under review)*

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 Applications
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. *Apr. 2019 - Present*

Project 1: Efficient Transformer for Mobile Applications

- **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 benchmarks and achieved better tradeoff between computation and BLEU score than the traditional 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)**, with a 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.
- 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 good 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 hardware.
- Analyzed the correlation between model size, Mult-Adds and latency, and provided design insights for Transformer architectures.
- Designed and implemented a supernet based Neural Architecture Search (NAS) for the hardware-aware transformers, and experimented on three machine translation datasets and different hardware, achieving up to 3× speedup and 25× smaller size.
- Submitted a paper to 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 neural networks and ranked 1st in both competition among the academic participants.
- 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 two traditional speaker representations and achieved state-of-the-arts with better data efficiency.
- Published a paper to the *Annual Conference of the International Speech Communication* **(Interspeech 2019 oral)**.

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

Project 1: Adversarial Robustness Exploration

- **Goal:** Investigate attacking and defending methods for adversarial examples on deep classification neural networks.
- 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 could improve adversarial robustness of the networks.

Honors & Awards

Scholarships

- **Chinese National Scholarship**, the highest honor for undergraduates, **top 0.2%** nationwide. *2018, 2019*
- **Fan Hsu-Chi Chancellor's Scholarship**, **top 0.1%** of 17,000 students in Shanghai Jiao Tong University. *2017*
- **Zhiyuan Honorary Scholarship**, **top 5%** of 17,000 students in Shanghai Jiao Tong University. *2016 - 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%** of 8,800 international participants). *2017*

Selected Projects

- **Visual Wake Words**, the code for the CVPR'19 VWW Challenge. (**1st place**, highlighted by Google). [Code | Link] *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 & Activities

- **Teaching Assistant of CS152: Programming**, gave an introduction to programming and designed homework. [Link] *2017*
- **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 Textbook 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). *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.