BAI, HAOLI

HOMEPAGE: haolibai.github.io
(+852) 5531 8737 / (+86) 191 1325 4820 ⋄ hlbai@cse.cuhk.edu.hk
Rm 101A, SHB, CUHK, Hong Kong.

RESEARCH INTEREST

My research interest majorly lies in **efficient deep learning**, including **network quantization**, **distillation**, **architecture search** and their applications in computer vision and natural languages.

EDUCATIION

The Chinese University of Hong Kong Aug. 2017 - Present
PhD in Computer Science and Engineering
Supervisors: Michael Lyu and Irwin King

University of Electronic Science and Technology of China Sep. 2013 - Jun. 2017
BEng in Computer Science, Yingcai Honor's College GPA: 3.93/4.00
Supervisor: Zenglin Xu Ranking: 2/87

EXPERIENCES

Huawei Noah's Ark Lab, Speech and Semantic Group Jul. 2020 - Present
Topic: Network Compression in NLP Tasks

Tencent AI Lab, Machine Learning Group, Jun. 2018 - Jun. 2020
Topic: Network Compression and Neural Architecture Search

PROJECTS

1. **PocketFlow: an Automated Network Compression Framework**. Tencent AI Lab
The project (https://github.com/Tencent/PocketFlow/) has received **2200+** stars and **480+** forks. I design the network quantization modules together with its automatic searching engine. Our 8-bit quantized MobileNet-V2 achieves around 3.0× speed-up deployed by TF-Lite, with no performance drop (Top-1 Acc. 72.26%) on ImageNet.

2. **Low-bit Transformer Quantization** Huawei Noah's Ark Lab
The project explores low-bit network quantization for Transformer on NLP tasks. On the GLUE benchmark, our recent work BinaryBERT https://arxiv.org/abs/2012.15701 reduces the model size by 24× and computation overhead by 15× with negligible performance drop. On machine translation, the binarized Transformer-base has only 2.0 ↓ of BLEU score on IWSLT-14 (en-de).

SELECTED PUBLICATIONS

1. **Haoli Bai***, Jiaxing Wang*, Jiaxiang Wu, Xupeng Shi, Junzhou Huang, Irwin King, Michael Lyu, and Jian Cheng. Revisiting Parameter Sharing for Automatic Channel Number Search, NeurlPS, 2020. (* equal contribution in the random order)

2. Kuo Zhong, Yin Wei, Chun Yuan, **Haoli Bai**, and Junzhou Huang. TranSlider: Transfer Ensemble Learning from Exploitation to Exploration, KDD, 2020.

3. Jiaxing Wang, **Haoli Bai**, Jiaxiang Wu, Jian Cheng. Bayesian Automatic Model Compression, IEEE Journal of Selected Topics in Signal Processing, 2020.

4. **Haoli Bai**, Jiaxiang Wu, Irwin King, Michale Lyu. Few Shot Network Compression via Cross Distillation, AAAI, 2020.

5. Jiaxing Wang, Jiaxiang Wu, **Haoli Bai**, Jian Cheng. MetaNAS: Meta Neural Architecture Search, AAAI, 2020.
6. Yuhang Li, Xin Dong, Saiqian Zhang, **Haoli Bai**, Yuanpeng Chen, Wei Wang. RTN: Reparameterized Ternary Network, AAAI, 2020.
7. Liangjian Wen, Xuanyang Zhang, **Haoli Bai**, Zenglin Xu. Structured Pruning of Recurrent Neural Networks through Neuron Selection, Neural Networks, 2020.
8. **Haoli Bai**, Zhuangbin Chen, Michael Lyu, Irwin King and Zenglin Xu. Neural Relational Topic Models for Scientific Articles, CIKM, 2018.
9. Hao Liu, Lirong He, **Haoli Bai**, and Zenglin Xu. Structured Inference for Recurrent Hidden Semi-Markov Model, IJCAI, 2018.
10. **Haoli Bai**, Zenglin Xu, Bin Liu and Yingming Li. Hierarchical Probabilistic Matrix Factorization with Network Topology for Multi-relational Social Network, ACML, 2016. **Best Student Paper Runner-up**.

Preprints

1. **Haoli Bai**, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, Irwin King. BinaryBERT: Pushing the Limit of BERT Quantization, Preprint arXiv: 2012.15701, 2020.
2. **Haoli Bai***, Xianghong Fang*, Jian Li, Zenglin Xu, Michael Lyu and Irwin King. Discrete Autoregressive Variational Attention Models for Language Generation, Preprint arXiv: 2004.09764, 2020. (* equal contribution in the random order)
3. **Haoli Bai**, Jiaxiang Wu, Irwin King and Michael Lyu. Cross Distillation: A Unified Approach for Few Shot Network Compression, submitted to Neural Networks.
4. Yuhang Li, Wei Wang, **Haoli Bai**, Ruihao Gong, Xin Dong, Fengwei Yu. Efficient Bitwidth Search for Practical Mixed Precision Neural Network. arxiv:2003.07577, 2020.

SERVICES

Senior PC Member: IJCAI-21

PC Member: ICML-21, NeurIPS-20, AAAI 19-21, IJCAI 20

Journal Reviewer: Cognitive Computation, Neural Networks, Neurocomputing

SELECTED AWARDS

AAAI Student Travel Grant of AAAI 2020.

ACM Student Travel Grant of CIKM 2018.

Postgraduate Studentship of the Chinese University of Hong Kong, 2017-2021.

Best Student Paper Runner-up of Asian Conference on Machine Learning, 2016.

National Scholarship (Top 2%), 2015

First Provencial Prize of the National Mathematical Contest in Modeling, 2015.

Meritorious Winner of the American Mathematical Contest in Modeling, 2016.

TECHNICAL SKILLS

Programming	PyTorch, Tensorflow, Python, MATLAB
Developping Tools	Git, Vim, Linux
TOEFL	100 (R:26, L:25, S:23, W:26)