Xuefei Ning

foxdoraame@gmail.com *(main mail)* · ningxuefei@mail.tsinghua.edu.cn *(org. mail)*
https://scholar.google.com/citations?user=oVslpJsAAAAJ
NICSEFC-EffAlg team I lead in Tsinghua University: https://nics-effalg.com/
Imagination-Research team with friends: https://github.com/imagination-research

Summary of Research Interests

- **Efficient deep learning (particularly efficient AIGC)**: I've been working on efficient DL since 2019. Since 2023, most of my work are centered around efficient AIGC. This is because from a young age, I have been fascinated by a future where we can be enriched with experiences beyond the constraints we face – whether physical, artificial, or otherwise. I think AIGC can play a vital role in shaping this future, thus I decided to shift my research focus toward AIGC applications, leveraging my expertise in efficient DL.

  Currently, I am mentoring 10+ graduate and undergraduate students with my Ph.D. advisor Prof. Wang to conduct research and engineering projects on efficient AIGC. The information about the team I lead in Tsinghua University can be found at https://nics-effalg.com/.

- **Towards better reasoning**: Recently, I have been thinking about the gaps between current techniques and the efficient, intelligent AI I imagine. While my thoughts on "what are the most key gaps and promising pathways to mitigate them" are evolving as our exploration carries on, I'm currently drawing insights from our own cognition and learning to help AI infer and learn in a more efficient and reliable way.

  Currently, I'm exploring these vague but exciting pathways with a small group of collaborators. The information about our unofficial team can be found at https://github.com/imagination-research.

Work Experience

Tsinghua University, Research-Track Assistant Professor 2024/03 – now
Huawei-Tsinghua Joint Postdoctoral Program, Postdoctoral Researcher 2021/12 – 2023/12
Advisor: Prof. Pinyan Lu, Prof. Yu Wang

Education

Tsinghua University, Bachelor of Electronic Engineering 2012/09 – 2016/07
Score: 92/100 *Ranking: 12/231* *Graudate with honor*

Tsinghua University, Doctor of Philosophy in Electronic Science and Technology 2016/9 – 2021/10
Advisor: Prof. Yu Wang, Prof. Huazhong Yang
Thesis: Neural Architecture Search for Efficient and Robust Convolutional Neural Networks

Mentorship Experiences

I built and lead the Efficient Algorithm (EffAlg) team in the NICSEFC group, starting in 2020 by instructing three undergraduates. Until 2024, I have mentored and continue to mentor 9 graduate students. I'm proud that four of them have earned their master's degrees, with one graduating with honors. For more information, please visit the NICSEFC-EffAlg website.

Teaching Experiences

C/UNIX Programming 2020 Autumn, 2022 Autumn, 2024 Autumn
Lecturer Course Instructor: Prof. Huazhong Yang, Prof. Yu Wang

Computer-Aided Design of Digital Circuits and Systems 2020 Spring, 2022 Spring, 2024 Spring

Teaching Assistant Course Instructor: Prof. Yu Wang

In the 2020 Spring course, I led the course collaboration project that produced the survey paper "Machine Learning for Electronic Design Automation: A Survey", published in TODAES 2021.

Industry Internship Experiences

Douban, Beijing, China 2015-7 – 2015-9

Software Engineer Intern of the Platform Group Mentor: Guillaume Bouriez

- Participate in the development of the RPC system on the private application cloud of Douban. This system acts as a vital component of the Micro-Service Architecture in Douban.
- Specifically, my task is to: (1) optimize "circuit breaker" of the RPC system; (2) develop service supervisor that reports service status; (3) support explicit interface declaration to increase the robustness of the system.

DeePhi Tech (now part of Xilinx), Beijing, China 2016-4 – 2016-8

Software Engineer Intern, IT Manager Mentor: Hong Luo

- Build up and maintain all the networking and servers in the startup.
- Lead the development of the compression toolchain for deploying LSTM (based on Kaldi) and CNN (based on Caffe) onto FPGA, including pruning and quantization functionalities.

Tencent AI Lab, Beijing, China 2019-3 – 2019-5

Machine Learning Research Intern Mentor: Yin Zheng

- Revise the paper "Nonparametric Topic Modeling with Neural Inference", published in Neurocomputing, which designs a nonparametric and hierarchical Dirichlet Process prior for VAE for flexible topic modeling.
- Develop a Gumbel-Softmax extension of differentiable neural architecture search. I am grateful that this is the start of the NAS research for my Ph.D. thesis, a direction suggested by Dr. Zheng during my internship.

Project Experiences

I have participated in more than 10 engineering projects with enterprises. I am the project PI for two engineering projects with Oppo and Baidu. I have published 10 patents as one of the first three authors.

Honors and Awards

Runner-up in NeurIPS 2024 Edge-Device LLM Competition (2nd/21) 2024/12
Solution: Decomposition and Finetuning for LLM Compression (As advisor)

Runner-up in NeurIPS 2018 Adversarial Vision Challenge Competition (2nd/339) 2018/12
Solution: Mutual Adversarial Training with Diverse Early Stop PGD (A joint work with Wenshuo Li)

Outstanding Graduate of Tsinghua University (10%) 2016/07

Future Scholar Scholarship of Tsinghua University (2/103) 2016/07

Grand Prize in the 5th Creativity Competition of Tsinghua University 2016/04
Solution: A Cross-Platform Mobile Application for Tsinghua Web Learning System (A joint work with Sihan Li)

National Scholarship for Encouragement, Academic Excellence Award 2015, 2014

First Prize in National High Schools Physics Competition 2011

Academic Service

Reviewer for ICML (2022, 2023, 2024), NeurIPS (2022, 2023, 2024), ICLR (2023, 2024, 2025), CVPR (2022, 2023), ICCV / ECCV (2022, 2023, 2024), AAAI (2023, 2024), AISTATS (2025), ECML-PKDD (2022). Reviewer for TPAMI, IJCV, CUSR, TCSVT, Pattern Recognit., TECS, TCAD, TODAES. Senior area chair for ACL 2025. Area chair for CVPR 2025. TPC member for DAC 2025 (AI track).

Selected Talks

- **Efficient Inference for Large Language Models – Algorithm, Model, and System**.
  Upcoming tutorial at *EMNLP 2025*.
- **Generative Model Compression and Acceleration**. Slide
  Invited talks at *Huawei; Apple-China; University of Chinese Academy of Sciences; University of Electronic Science and Technology of China; VIVO; AMD-China; and others.*
- **An Introduction to Quantization of Large Language Models**. Video and slide.
  Invited tutorial for *an Competition organized by AWS-China.*
- **Model Compression Towards Efficient Deep Learning Inference**. Slide
  Invited talks at *Huawei; Inceptio.ai; Beihang University; and others.*
- **Neural Architecture Search and Architecture Encoding**. Slide
  Invited talks at *DAMO Academy of Alibaba Group (U.S.); Renmin University of China; and others.*

Publications

* indicates equal contribution, + indicates corresponding author and project advisor

Towards Better Reasoning

- **Xuefei Ning***, Zifu Wang*, Shiyao Li*, Zinan Lin*, Peiran Yao*, and Others, Can LLMs Learn by Teaching for Better Reasoning? A Preliminary Study, In NeurIPS 2024.

Efficient NN Inference

- Zixuan Zhou*, **Xuefei Ning***+, Ke Hong*, Tianyu Fu, and Others, A Survey on Efficient Inference for Large Language Models, In arXiv 2024.

1. **Algorithm-level Techniques**:
   - **[Vision Generation]** Enshu Liu*, **Xuefei Ning***+, Zinan Lin*, Huazhong Yang, Yu Wang+, OMS-DPM: Deciding The Optimal Model Schedule for Diffusion Probabilistic Model, In ICML 2023.
   - **[Language Generation] Xuefei Ning***, Zinan Lin*, Zixuan Zhou*, Zifu Wang, and Others, Skeleton-of-Thought: Prompting Large Language Models for Efficient Parallel Generation, In ICLR 2024.
   - **[Vision Generation]** Enshu Liu, **Xuefei Ning+**, Huazhong Yang, Yu Wang+, A Unified Sampling Framework for Solver Searching of Diffusion Probabilistic Models, In ICLR 2024.
   - **[Vision Generation]** Enshu Liu, **Xuefei Ning+**, Yu Wang, Zinan Lin+, Distilling Auto-regressive Models into Few Steps 1: Image Generation, In arXiv 2024.
   - **[Vision Generation]** Yao Teng, Han Shi, Xian Liu, **Xuefei Ning**, and Others, Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding, In arXiv 2024.
2. **Model-level Design or Compression**:
   - **(Chinese Book)** Yu Wang, **Xuefei Ning**, Efficient Deep Learning: Model Compression and Design, Published by the Publishing House of Electronics Industry, 2024/07. (Introduce the model compression and efficient model design techniques systematically)
   - **(Book Chapter)** Yu Wang, **Xuefei Ning**, Shulin Zeng, Yi Cai, and Others, Hardware Design and Software Practices for Efficient Neural Network Inference, In the Low-Power Computer Vision Book, Published by the CRC Press, 2022. (Introduce our practical efficient deployment solutions)
   - **Xuefei Ning***, Tianchen Zhao*, Wenshuo Li, and Others, DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation, In ECCV 2020 (**Spotlight**).
   - Xiangsheng Shi*, **Xuefei Ning***+, Lidong Guo*, Tianchen Zhao, and Others, Memory-Oriented Structural Pruning for Efficient Image Restoration, In AAAI 2023.
   - Tianchen Zhao, **Xuefei Ning+**, Ke Hong, Zhongyuan Qiu, and Others, Ada3D: Exploiting the Spatial Redundancy with Adaptive Inference for Efficient 3D Object Detection, In ICCV 2023.
   - **[Language Generation]** Shiyao Li, **Xuefei Ning+**, Luning Wang, Tengxuan Liu, and Others, Evaluating Quantized Large Language Models, In ICML 2024.
   - **[Vision Generation]** Tianchen Zhao*, **Xuefei Ning***+, Tongcheng Fang*, and Others, MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization, In ECCV 2024.
   - **[Vision Generation]** Zhihang Yuan*, Hanling Zhang*, Pu Lu*, **Xuefei Ning+**, and Others, DiTFastAttn: Attention Compression for Diffusion Transformer Models, In NeurIPS 2024.

- **[Vision Generation]** Tianchen Zhao, Tongcheng Fang, ..., **Xuefei Ning+**, Yu Wang+, ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation, In arXiv 2024.
- **[Language Generation]** Tianyu Fu*, Haofeng Huang*, **Xuefei Ning***+, Genghan Zhang, and Others, MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression, In arXiv 2024.
- **[Language Generation]** Enshu Liu*, Junyi Zhu*, Zinan Lin+, **Xuefei Ning+**, and Others, Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models: Enhancing Performance and Reducing Inference Costs, In arXiv 2024.

Efficient NN Training

- Kai Zhong, **Xuefei Ning**, Guohao Dai, Zhenhua Zhu, and Others, Exploring the Potential of Low-bit Training of Convolutional Neural Networks, In TCAD 2022.
- Minxue Tang, **Xuefei Ning**, Yitu Wang, Jingwei Sun, and Others, FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning, In CVPR 2022.
- **[Vision Generation]** Enshu Liu*, Junyi Zhu*, Zinan Lin+, **Xuefei Ning+**, and Others, "Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models Better", In arXiv 2024.

Efficient and Automatic Optimization Process

1. **Efficient Neural Architecture Search:**
   Summary Website: https://sites.google.com/view/nas-nicsefc
   Unified NAS Framework (containing following researches): https://github.com/walkerning/aw_nas
   - **Xuefei Ning**, Yin Zheng, Tianchen Zhao, Yu Wang, Huazhong Yang, A Generic Graph-based Neural Architecture Encoding Scheme for Predictor-based NAS, In ECCV 2020.
   - Wenshuo Li*, **Xuefei Ning***, Guangjun Ge, Xiaoming Chen, Yu Wang, Huazhong Yang, FTT-NAS: Discovering Fault-Tolerant Neural Architecture, In ASP-DAC 2020.
   - Shulin Zeng*, Hanbo Sun*, Yu Xing, **Xuefei Ning**, Yi Shan, Xiaoming Chen, Yu Wang, Huazhong Yang, Black Box Search Space Profiling for Accelerator-Aware Neural Architecture Search, In ASP-DAC 2020.
   - **Xuefei Ning**, Changcheng Tang, Wenshuo Li, Zixuan Zhou, and Others, Evaluating Efficient Performance Estimators of Neural Architectures, In NeurIPS 2021.
   - **Xuefei Ning**, Guangjun Ge, Wenshuo Li, Zhenhua Zhu, and Others, FTT-NAS: Discovering Fault-Tolerant Convolutional Neural Architecture, In TODAES 2021.
   - Zixuan Zhou*, **Xuefei Ning***+, Yi Cai, Jiashu Han, and Others, CLOSE: Curriculum Learning On the Sharing Extent Towards Better One-shot NAS, In ECCV 2022.
   - **Xuefei Ning***, Zixuan Zhou*, Junbo Zhao, Tianchen Zhao, and Others, TA-GATES: An Encoding Scheme for Neural Network Architectures, In NeurIPS 2022 (**Spotlight**).
   - Junbo Zhao*, **Xuefei Ning***+, Enshu Liu, Binxin Ru, and Others, Dynamic Ensemble of Low-fidelity Experts: Mitigating NAS Cold-Start, In AAAI 2023 (**Oral**).
   - **Xuefei Ning**, Yin Zheng, Zixuan Zhou, Tianchen Zhao, Huazhong Yang, Yu Wang, A Generic Graph-based Neural Architecture Encoding Scheme with Multifaceted Information, In TPAMI 2023.
   - Hanbo Sun, Zhenhua Zhu, Chenyu Wang, **Xuefei Ning+**, and Others, Gibbon: Efficient Co-Exploration of NN Model and Processing-In-Memory Architecture, In DATE 2022 & TCAD 2023.
2. **Efficient Performance Evaluation:**
   - **[Vision Generation]** Lin Zhao*, Tianchen Zhao*, Zinan Lin, **Xuefei Ning+**, and Others, FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Generative Models, In CVPR 2024.

Other Researches

- **Xuefei Ning**, Yin Zheng, Zhuxi Jiang, Yu Wang, and Others, Nonparametric Topic Modeling with Neural Inference, In Neurocomputing 2020.
- Tong Wu, **Xuefei Ning**, Wenshuo Li, Ranan Huang, Huazhong Yang, Yu Wang, Physical Adversarial Attack on Vehicle Detector in the Carla Simulator, Technical report in arXiv 2020.
- Tianchen Zhao, Niansong Zhang, **Xuefei Ning**, He Wang, Li Yi, Yu Wang, CodedVTR: Codebook-based Sparse Voxel Transformer with Geometric Guidance, In CVPR 2022.
- Ye Mu*, Weilin Liu*, Chao Yu, **Xuefei Ning+**, and Others, Multi-Agent Vulnerability Discovery for Autonomous Driving with Hazard Arbitration Reward, In CASE 2024.

- Lidong Guo*, **Xuefei Ning***+, Yonggan Fu, Tianchen Zhao, and Others, Rad-NeRF: Ray-decoupled Training of Neural Radiance Field, In NeurIPS 2024.
- Tao Yuan, **Xuefei Ning**+, Dong Zhou, Zhijie Yang, and Others, LV-Eval: A Balanced Long-Context Benchmark with 5 Length Levels Up to 256K, In arXiv 2024.
- Kaiyi Huang, Yukun Huang, **Xuefei Ning**, Zinan Lin, Yu Wang, Xihui Liu, GenMAC: Compositional Text-to-Video Generation with Multi-Agent Collaboration, In arXiv 2024.

Latest update: 2024/12