2026 ISNN Special Session: Compression and Acceleration of Large Models

Hanzhong, Shaanxi, China

Topic: The unprecedented capabilities of Large Language Models (LLMs) and other foundational AI models have come at a steep computational cost. Their massive scale, often involving hundreds of billions of parameters, poses critical challenges for practical deployment, including prohibitive memory footprints, high inference latency, and enormous energy consumption. These barriers restrict the accessibility, scalability, and real-world applicability of state-of-the-art AI, confining powerful models primarily to well-resourced data centers.

To democratize access and enable a new generation of efficient, responsive, and sustainable AI applications—from real-time assistants on mobile devices to cost-effective large-scale cloud services—research into model **compression** and **acceleration** has become a field of paramount importance. This special session focuses on innovative techniques to dramatically reduce the computational and storage demands of large models while striving to preserve their original performance and capabilities.

This endeavor requires a multi-faceted approach, tackling the problem at algorithmic, systems, and hardware levels:

Algorithmic Compression: Developing novel methods for pruning redundant parameters, quantizing high-precision weights into lower-bit representations, and distilling knowledge from cumbersome "teacher" models into compact "student" models.

Architectural Innovation: Designing inherently more efficient model architectures, attention mechanisms, and modules that achieve strong performance with fewer parameters and operations.

System & Hardware-Aware Acceleration: Creating specialized compilers, runtime engines, and kernels that optimize execution graphs for specific hardware (GPUs, NPUs, CPUs). This also involves co-designing algorithms and hardware to unlock maximal efficiency.

Efficient Training & Adaptation: Pioneering strategies to efficiently fine-tune or adapt compressed models for downstream tasks, ensuring compressed models remain versatile and effective.

This special session aims to gather leading researchers and engineers to present cutting-edge advancements, share insights, and discuss the future directions of making large models leaner, faster, and more deployable everywhere. We seek contributions that push the boundaries of efficiency without compromising the intelligence that makes these models revolutionary.

Organizers: Feng-Lei Fan, Juntong Fan, Haoran Wang

List of Confirmed Speakers (Alphabetical Order):

Fenglei Fan	City University of Hong Kong
Xiaocheng Feng	Harbin Institute of Technology
Ting Gao	Huazhong University of Science and Technology
Chen Li	Northestern University
Chuang Niu	Harbin Institute of Technology
Liu Shi	Nanchang University
Maolin Wang	City University of Hong Kong
Xu Yang	Xidian University
Yulun Zhang	Shanghai Jiaotong University

Fenglei Fan

Fenglei Fan is currently an Assistant Professor with Department of Data Science, City University of Hong Kong. His primary research interests lie in NeuroAI and its applications in model compression and medical imaging. He was the recipients of the IBM AI Horizon Scholarship, the 2021 International Neural Network Society Doctoral Dissertation Award, and He won OlympusMons Pioneering Award, a prestigious award in the field of storage. He has one paper selected as one of few 2024 CVPR Best Paper Award Candidates, one won the IEEE Nuclear and Plasma Society IEEE TRPMS Best Paper Award, one ESI highly cited paper, and one North American Highly Cited Paper in Biomedical Science by IOP. He organized special issues in journals like IEEE TRPMS, presented three tutorials in AAAI2023, IJCNN25, and WWW2025, and served as (senior) program committee members in AAAI and IJCAI.

Xiao-Cheng Feng

Xiaocheng Feng is Professor and Doctoral Supervisor at the Center for Social Computing and Information Retrieval, Faculty of Computing, Harbin Institute of Technology (HIT), where he also serves as Associate Dean of the School of Artificial Intelligence and Assistant Director of Heilongjiang Provincial Key Laboratory of Chinese Information Processing. His research focuses on natural language processing, large models, text generation, and machine translation, with over 40 publications in CCF Rank A/B venues including ACL, AAAI, IJCAI, TKDE, and Science China Information Sciences (Google Scholar citations: 5,800+; two Google Scholar 2020 Highly Cited Papers; Paper Digest EMNLP 2020 Top Ten Most Influential Paper). He serves as Senior/Program Committee member for NeurIPS, ICML, AAAI, IJCAI, ACL, etc., while holding concurrent roles as Deputy Researcher at Peng Cheng Laboratory, Deputy Secretary-General of the Chinese Information Processing Society's Natural Language Generation Committee, and Chair of CCF YOCSEF Harbin. His honors include China Association for Science and Technology's Young Talent Support Project (6th cohort), Chinese Information Processing Society's Outstanding Doctoral Dissertation Award, Heilongjiang Science & Technology Progress Award (Second Prize), 2023 CCF-NLP Young Rising Star Award, 2022 WAIC Yunfan Award, and 2023 Xiaomi Young Scholar. He leads multiple grants including sub-projects of China's National Key R&D Program "Next-Generation AI" and National Laboratory Key Projects, NSFC General/Youth Programs, Heilongjiang Key R&D Program, Outstanding Youth Foundation, and MSRA Collaborative Research Program, maintaining research partnerships with Huawei, Tencent, iFLYTEK, Microsoft, etc.

Ting Gao

Gao Ting earned her Ph.D. from the Illinois Institute of Technology (USA) in 2015. She has served as a Senior Data Scientist at MZ (a leading US mobile gaming company) and a Machine Learning Algorithm Engineer in Twitter's Big Data Product R&D department, where she developed multiple deep reinforcement learning/deep learning-based recommendation systems and real-time bidding models for big data stream online learning. Currently affiliated with the School of Mathematics and Statistics at Huazhong University of Science and Technology, her primary research focuses on: **identification of non-Gaussian stochastic dynamical systems, prediction of effective dynamics and optimal control, with applications in information communication and financial mathematics**. She has published multiple high-impact papers in journals including *Applied Mathematics and Computation (AMC), Communications in Nonlinear Science and Numerical Simulation (CNSNS), SIAM Journal* series, *International Journal of Bifurcation and Chaos (IJBC),* and *Inverse Problems and Imaging (IPI)*.

Chen Li

Chen Li earned his Ph.D. from Associate Professor with Long-Term Appointment at Northeastern University, and expert of the Sichuan Provincial "Thousand Talents Plan." He is serving as an editorial board member of IEEE TMI, associate editor of Frontiers in Microbiology, and associate editor of Journal of X-Ray Science and Technology. He holds the position of Deputy Director of the Review Committee for the "Xiaoping Innovation Laboratory" under the Central Committee of the Communist Youth League, and is a member of IEEE and CCF. He is primarily engaged in research in artificial intelligence and microscopic image analysis. Awards received include the Second Prize of the National Invention and Entrepreneurship Achievement Award and the Second Prize of the Sichuan Provincial Science and Technology Progress Award. Has published over 100 papers as first or corresponding author, including more than ten ESI hot papers and highly cited papers.

Chuang Niu

Chuang Niu is currently a full professor in Harbin Institute of Technology. He serves as an Associate Director in AXIS Lab led by Prof. Ge Wang. He received my B.S. in biomedical engineering in 2015, and my Ph.D. in pattern recognition and machine intelligence in 2020, from Xidian University . He was a visiting student from 2019 to 2020 and a Postdoc from 2020 to 2023 at RPI. His research interest is in Medical Multimodal Multitask Foundation Model (M3FM), self-supervised/unsupervised learning, weakly-supervised learning, representation learning, clustering, and biomedical imaging and analysis. He has been working on weakly-supervised especially self-supervised learning algorithm development since 2018, with the applications in image segmentation, image classification/clustering, representation learning for object recognition and detection, medical CT imaging, and large foundation models for medical multimodal AI.

Liu Shi

Liu Shi is a specially appointed associate researcher at Nanchang University, primarily engaged in research on 3D X-ray imaging for packaging inspection and intelligent image processing. In 2019, he earned a Bachelor of Science degree in Information and Computing Science from the School of Mathematics and Statistics at Chongqing University. In 2024, he obtained a Doctor of Philosophy degree in Particle Physics and Nuclear Physics from the Institute of High Energy Physics, Chinese Academy of Sciences. He has published 10 academic papers in domestic and international journals and holds 3 authorized patents. He has participated in key projects such as the National Key R&D Program's Major Scientific Instrument and Equipment Development Project "X-ray 3D Layered Imaging System" and the Chinese Academy of Sciences Innovation Interdisciplinary Team Project "Advanced Packaging X-ray 3D Inspection Technology Innovation Interdisciplinary Team." He has also secured funding from the Jiangxi Provincial Early Career Young Talent Program, the 2025 Nanchang "One Enterprise, One Doctor" Science and Technology Talent Service Project (first batch), and the Nanchang University Youth Talent Cultivation Fund.

Xu Yang

Xu Yang is an Associate Professor (Huashan Scholar - Elite) and Master's Supervisor at Northwestern Polytechnical University, specializing in multimodal content understanding, machine learning, and computer vision. With 30+ publications in top-tier venues including IEEE Transactions journals (TPAMI, TNNLS, TIP), CCF Rank A conferences (CVPR, NeurIPS, AAAI), and one ESI Highly Cited Paper, he leads a National Natural Science Foundation (NSFC) project while contributing to NSFC Key Programs and National Key R&D Programs. His honors include Shaanxi Province's Young Talent Support Program, First Prize in Shaanxi Natural Science & Technology Award, ACM Xi'an Outstanding Doctoral Dissertation Award (2021), and Wu Wen Jun AI Science & Technology Award for Excellent Doctoral Dissertation (2021). He serves as reviewer and (Senior) Program Committee member for TPAMI/TIP/TCYB/TNNLS journals and CVPR/AAAI/NeurIPS/ICML conferences. English homepage: https://scholar.google.com/citations?user=Uc3piAIAAAAJ&hl=en.

Maolin Wang

Dr. Maolin Wang is a Research Assistant Professor at the Hong Kong Institute of AI for Science (HKAI-Sci), City University of Hong Kong. He received his Ph.D. from the Department of Data Science, City University of Hong Kong, under the supervision of ***Prof. Xiangyu Zhao*** and co-supervision of ***Dr. Ruocheng Guo and Prof. Junhui Wang***. His research focuses on graph learning, model compression, tensor/matrix decomposition, and LLMs. Prior to his doctoral studies, he received his master's and bachelor's degrees from the University of Electronic Science and Technology of China (UESTC) under the guidance of ***Prof. Zenglin Xu***. His ultimate goal is to contribute to the advancement of sustainable AI that is both powerful and environmentally responsible, making AI more accessible to everyone.

Yulun Zhang

Yulun Zhang is a tenured-track Associate Professor at Shanghai Jiao Tong University. He received his Bachelor's, Master's, and Ph.D. degrees from Xidian University, Tsinghua University, and Northeastern University (USA), respectively. After completing his Ph.D., he conducted postdoctoral research at ETH Zurich. His research interests lie in computer vision and machine learning, with a particular focus on image/video restoration, model compression, computational imaging, multimodal computing, and large language models. He has published over 100 academic papers in top-tier international journals and conferences. His publications have been cited more than 31,000 times on Google Scholar, with his most cited first-authored paper receiving over 6,800 citations. He received the Best Student Paper Award at IEEE VCIP 2015, the Best Paper Award at the IEEE ICCV RLQ Workshop 2019. He was recognized as one of the "Top 100 Chinese Rising Stars in AI" (2021), has been listed in Stanford's "World's Top 2% Scientists" for several consecutive years (2021–2025), and was named an Elsevier "Highly Cited Chinese Researcher" in 2024. In recent years, he has served as an Area Chair for top conferences such as CVPR, ICCV, ECCV, ICLR, NeurIPS, ICML, ACM MM, AAAI, and IJCAI.