Igor Fedorov

fedorov.uofi@gmail.com http://ifed-ucsd.github.io/

Summary

- At Meta, I am a tech lead for large language model compression and inference optimization for edge use cases. I have also been responsible for algorithm research and software development for neural architecture search on ads recommendation models.
- At ARM, I was responsible for algorithm research and software development of ARM's proprietary neural architecture search framework, enabling the automatic design of performant neural networks that can be deployed in extremely resource constrained hardware and that showcase ARM hardware capabilities

Education

- **University of California, San Diego** — San Diego, CA
  Ph.D in Electrical Engineering; GPA:4.0 — *Sept 2018*
  Advisor: Bhaskar D. Rao, co-advisor: Truong Q. Nguyen
- **University of Illinois at Urbana-Champaign** — Urbana-Champaign, IL
  M.S. in Electrical Engineering; GPA:3.78 — *May 2014*
  Advisor: Pierre Moulin
- **University of Illinois at Urbana-Champaign** — Urbana-Champaign, IL
  B.S. in Electrical Engineering; GPA:3.90 — *May 2012*
  James Scholar, Highest Honors

Experience

- **Meta** — Remote
  AI Research Scientist — *June 2022 – Present*
  - Tech lead for LLM compression / optimization
  - Model optimization research for family of apps and mixed reality devices
  - AutoML and neural architecture search for ads recommendation system
- **ARM Research** — Boston, MA
  Staff Research Engineer — *April 2021 – June 2022*
  - Machine learning research group
  - Neural architecture search for hardware-aware, efficient neural networks
  - Differentiable neural architecture search, Bayesian optimization, unstructured pruning, channel pruning, quantization, compression
- **ARM Research** — Boston, MA
  Senior Research Engineer — *Sept 2018 – April 2021*
  - Machine learning research group
- **Samsung Research** — San Diego, CA
  Intern — *June 2017 – Sept 2017*
  - Deep learning research group
- **Qualcomm** — San Diego, CA
  Intern — *May 2015 – Aug 2015*
  - Developed continuous multi-modal authentication system for verifying mobile user's identity
- **Qualcomm** — San Diego, CA
  Intern — *May 2013 – Sept 2014*
  - Developed real-time, fixed point C implementation of Fast Stereo Independent Vector Analysis

- **Qualcomm** San Diego, CA
  Intern *Jun 2012 – Aug 2012*
  - Developed novel voice activity detector using non-negative matrix factorization
- **Cisco** San Jose, CA
  Intern *Jun 2011 – Aug 2011*
  - Implemented testing framework for NX-OS
- **ComEd** Libertyville, IL
  Intern *Jun 2010 – Aug 2010*
  - Worked with Transmission and Substation Department in the Testing Group

Patents

- M. Haddon, **I. Fedorov**, R. Jeyapaul, P. N. Whatmough, Z. Liu, "Error detection," *US patent application.*
- H. Tann., R. Navarro, **I. Fedorov**, C. Zhou, P. Whatmough, M. Mattina, "System, Devices and/or Processes for Defining a Search Space for Neural Network Processing Device Architectures," *US patent application.*
- **I. Fedorov**, P. Whatmough, "Neural network system and training method," *US patent application, 17576101*, 2023
- **I. Fedorov**, R. Matas, C. Zhou, H. Tann, P. Whatmough, M. Mattina, "System, devices and/or processes for designing neural network processing devices," *US patent application 17394048*, 2023
- M. El-Khamy **I. Fedorov**, J. Lee, "Image denoising neural network architecture and method of training the same," *US patent*, 2020.

Publications (by topic)

Large Language Models

- Zechun Liu, Changsheng Zhao, **Igor Fedorov**, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, Tijmen Blankevoort, "SpinQuant: LLM Quantization with Learned Rotations," *ICLR*, 2025.
- **Igor Fedorov**, Kate Plawiak, Lemeng Wu, Tarek Elgamal, Naveen Suda, Eric Smith, Hongyuan Zhan, Jianfeng Chi, Yuriy Hulovatyy, Kimish Patel, Zechun Liu, Changsheng Zhao, Yangyang Shi, Tijmen Blankevoort, Mahesh Pasupuleti, Bilge Soran, Zacharie Delpierre Coudert, Rachad Alao, Raghuraman Krishnamoorthi, Vikas Chandra, "Llama Guard 3-1B-INT4: Compact and Efficient Safeguard for Human-AI Conversations," *Arxiv*, 2024.
- Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, **I. Fedorov**, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, L. Lai, V. Chandra, "MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases," *ICML*, 2024.

Neural Architecture Search

- W. Wen, K. Liu, **I. Fedorov**, X. Zhang, H. Yin, W. Chu, K. Hassani, M. Sun, J. Liu, X. Wang, L. Jiang, Y. Chen, B. Zhang, X. Liu, D. Cheng, Z. Chen, G. Zhao, F. Han, J. Yang, Y. Hao, L. Xiong, W. Chen, "Rankitect: Ranking Architecture Search Battling World-class Engineers at Meta Scale," *ArXiv*, 2023.
- H. Zheng, K. Liu, **I. Fedorov**, X. Zhang, W. Chen, W. Wen, "SiGeo: Sub-One-Shot NAS via Information Theory and Geometry of Loss Landscape," *ArXiv*, 2023.
- T. Zhang, W. Wen, **I. Fedorov**, X. Liu, B. Zhang, F. Han, W. Chen, Y. Han, F. Yan, H. Li, Y. Chen, "DistDNAS: Search Efficient Feature Interactions within 2 Hours," *ArXiv*, 2023.

- Y. Chai, D. Tripathy, C. Zhou, D. Gope, **I. Fedorov**, R. Matas, D. Brooks, G. Wei, P. Whatmough, "PerfSAGE: Generalized Inference Performance Predictor for Arbitrary Deep Learning Models on Edge Devices," *ArXiv*, 2023.
- A. Kag, **I. Fedorov**, A. Gangrade, P.N. Whatmough, V. Saligrama, "Efficient Edge Inference by Selective Query," *ICLR*, 2023.
- **I. Fedorov**, R. Matas, H. Tann, C. Zhou, M. Mattina, P.N. Whatmough, "UDC: Unified DNAS for Compressible TinyML Models for Neural Processing Units," *NeurIPS*, 2022.
- K. Bhardwaj, J. Ward, C. Tung, D. Gope, L. Meng, **I. Fedorov**, A. Chalfin, P. Whatmough, D. Loh, "Restructurable Activation Networks," *ArXiv*, 2022.
- A. Kag, **I. Fedorov**, A. Gangrade, P.N. Whatmough, V. Saligrama, "Achieving High TinyML Accuracy through Selective Cloud Interactions," *ICML DyNN workshop*, 2022.
- C. Banbury*, C. Zhou*, **I. Fedorov***, R.M. Navarro, U. Thakker, D. Gope, V.J. Reddi, M. Mattina, P.N. Whatmough, "MicroNets: Neural Network Architectures for Deploying TinyML Applications on Commodity Microcontrollers," *MLSys*, 2021.
- **I. Fedorov**, M. Stamenovic, C. Jensen, L. Yang, A. Mandell, Y. Gan, M. Mattina, P.N. Whatmough, "TinyLSTMs: Efficient Neural Speech Enhancement for Hearing Aids," *INTERSPEECH*, 2020.
- S. Sandha, M. Aggarwal, **I. Fedorov**, M. Srivastava, "Mango: A Python Library for Parallel Hyperparameter Tuning," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.
- **I. Fedorov**, R.P. Adams, M. Mattina, P.N. Whatmough, "SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers," *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

RNN Compression

- U. Thakker, **I. Fedorov**, J. Beu, D. Gope, C. Zhou, G. Dasika M. Mattina, "Pushing the limits of RNN Compression," *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

Multimodal Dictionary Learning

- K. Kreutz-Delgado, B.D. Rao, **I. Fedorov**, S. Das, "Dictionaries in machine learning," *Signal Processing and Machine Learning Theory, 2023.*
- **I. Fedorov**, B.D. Rao, "Multimodal Sparse Bayesian Dictionary Learning," *arXiv preprint*, 2018.
- **I. Fedorov**, B.D. Rao, T.Q. Nguyen, "Multimodal Sparse Bayesian Dictionary Learning Applied to Multimodal Data Classification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Sparsifying Deep Neural Networks

- C. Lee, **I. Fedorov**, B.D. Rao, H. Garudadri, "SSGD: Sparsity-promoting Stochastic Gradient Descent Algorithm for Unbiased DNN Pruning," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.
- **I. Fedorov**, B.D. Rao, "Sparsifying Deep Neural Networks," arXiv preprint, 2018.

Non-negative Matrix Factorization

- **I. Fedorov**, A. Nalci, R. Giri, B.D. Rao, T.Q. Nguyen, H. Garudadri, "A Unified Framework for Sparse Non-Negative Least Squares using Multiplicative Updates and the Non-Negative Matrix Factorization Problem," *Signal Processing*, Volume 146, May 2018, Pages 79-91, ISSN 0167-1648.

- A. Nalci, **I. Fedorov**, M. Al-Shoukairi, T. T. Liu, B.D. Rao. "Rectified Gaussian Scale Mixtures and the Sparse Non-Negative Least Squares Problem," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3124-3139, June 2018.

Robust Sparse Signal Recovery & Face Recognition

- **I. Fedorov**, R. Giri, B.D. Rao, T.Q. Nguyen, "Relevance Vector Machine: A Novel Person Re-Identification Framework," *arXiv preprint arXiv:1703.10645*, 2017.
- **I. Fedorov**, R. Giri, B.D. Rao, T.Q. Nguyen, "Robust Bayesian Method for Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition," *IEEE International Conference on Image Processing (ICIP)*, 2016.

Single Photon Emission Computed Tomography

- **I. Fedorov**, S. Obrzut, B. Song, B.D. Rao, "SPECT Image Reconstruction under Imaging Time Constraints," *51st Asilomar Conference on Signals, Systems, and Computers*, 2017.
- **I. Fedorov**, B. Song, B.D. Rao, I. Levitan, S. Obrzut, "Total Variation Regularization in I-123 Ioflupane SPECT Reconstruction," *Journal of Nuclear Medicine*, 2017.

Action Recognition

- **I. Fedorov**, "Kinect depth video compression for action recognition," *Master's thesis*, 2014.
- A. Khosrowpour, **I. Fedorov**, A. Holynski, J.C. Niebles, and M. Golparvar-Fard, "Automated Worker Activity Analysis in Indoor Environments for Direct-Work Rate Improvement from long sequences of RGB-D Images," *Construction Research Congress: Construction in a Global Network*, 2014.

Miscellaneous

- P.S. Shenoy, **I. Fedorov**, T. Neyens, P.T. Krein, "Power delivery for series connected voltage domains in digital circuits," *International Conference on Energy Aware Computing*, 2011.

Skills

Python, Tensorflow, Pytorch, Matlab, C/C++, LaTex, Fluent in Russian

Teaching

WES 267: Intro to Digital Signal Processing, UCSD
ECE 161B: Digital Signal Processing, UCSD
ECE 445: Senior Design, UIUC

Honors and Activites

ARCS Fellowship, 2015-2018
ECE Departmental Fellowship, UCSD, 2014
Jules D. Falzer Scholarship for outstanding scholastic record, UIUC, 2012