Jacob Ioffe

✉ ji97@cornell.edu in jacob-ioffe jioffe502

Education

Cornell-Tech (Cornell University), New York, NY **Aug 2023 – May 2025**
Dual M.S. in Computer Science – Health Tech Concentration *GPA: 4.0/4.0*

Coursework: Machine Learning Engineering, NLP, ML Hardware & Systems, Deep Learning Efficiency, Language Modeling, Applied ML, Drug Development, ML for Health, Algorithms, Computational Biology, AI for Healthcare

Vanderbilt University, Nashville, TN **Aug 2019 – May 2023**
B.S. in Computer Science, Minors in Neuroscience and Mathematics *GPA: 3.7/4.0*

Technical Skills

Languages & Frameworks: Python, PyTorch, C++, SQL, R, TensorFlow, Scikit-learn, Hugging Face, Linux
AI & Deep Learning: CUDA, Numba, NLP, RAG, Quantization, Pruning, NAS, Multi-agent Systems
Data Processing & ML Ops: Pandas, NumPy, PySpark, Hadoop, AWS, Azure, Docker, SLURM, WandB

Experience

Johnson & Johnson Innovative Medicine | Computational R&D Intern **May 2024 – Aug 2024**

- Engineered a benchmarking framework leveraging NVIDIA MPS for GROMACS, optimizing GPU sharing in multi-node environments; improved job throughput by up to 30% for diverse molecular simulations (6k-12M atoms) and provided insights to reduce costs and enhance resource allocation for drug discovery teams
- Architected an LLM evaluation pipeline for medical table reasoning across diverse healthcare datasets, implementing multi-model comparisons, advanced prompt engineering, and clinically-relevant evaluation metrics

Weill-Cornell Medical School | Graduate Research Assistant **Jan 2024 – May 2024**

- Applied clustering techniques (UMAP, t-SNE, PaCMAP) and dimensionality reduction (CCA, cross-decomposition) to multi-omics and phenotypic data from the UK Biobank (N>500,000) to identify subtypes of depressive disorders
- Reimplemented Improved Deep Embedded Clustering (IDEC) in PyTorch, extending analysis to neural network-based clustering and collaborating with scientists to align computational findings with clinical insights

Humana | Software Engineering Intern **May 2021 – Aug 2023**

- Contributed across 5 internships in software, cloud, and data engineering, leveraging Python, PySpark, AWS, and Azure
- Engineered real-time log analysis system using Python and PySpark to diagnose performance issues in NLP-driven fax-to-PDF conversion tool for care providers; implemented SQL query optimizations, reducing load times by 35%
- Used Azure Functions and Stream Analytics to process medical claims data, resulting in PowerBI visualizations for IT

Projects

TinyConv **Python, PyTorch, TinyML, C++**

- Implemented and compared Quantization-Aware Training with Post-Quantization on Arduino's TinyML, analyzing accuracy and model size trade-offs between 2 and 8 bits, and with minifloat quantization
- Fine-tuned structured and unstructured pruning in PyTorch to optimize inference time within Arduino's resource limits
- Reduced model parameters and FLOPs by 75% through structured channel pruning, achieving an 18% runtime gain during on-device inference and 85% test accuracy on the Speech Commands dataset after 15 epochs of fine-tuning

Optimizing DNN Primitives Using TVM **TVM, CUDA, Python**

- Achieved over 20× speedups and near-peak hardware utilization on CPU and GPU for fundamental DNN operations (1D Convolution on CPU/GPU, GEMM, Depthwise Separable Convolution) using the TVM compiler framework
- Applied hardware-specific optimizations—including cache-aware tiling, SIMD vectorization, shared memory optimization, and efficient thread block mapping—to enhance computational efficiency while maintaining numerical accuracy

Heart Murmur Classification **Python, Scikit-learn, Librosa, Streamlit**

- Developed a machine learning pipeline for heart murmur detection using Mel-frequency cepstral coefficients (MFCCs) and traditional ML models, achieving up to 86.67% recall and 96.03% precision across different heart valve locations
- Implemented precise audio segmentation and noise reduction techniques, improving model performance by an average of 9.15% in F1-score and reducing false negatives by 7 instances across all models
- Built a Streamlit app for real-time murmur detection, integrating SHAP values giving providers interpretable results

Publications

1. Liu YH, Luo C, Golding SG, **Ioffe JB**, Zhou XM. *Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data*. Nature Communications. 2024;15(1):2447.
2. Li S, Guo Z, **Ioffe JB**, et al. *Text mining of gene–phenotype associations reveals new phenotypic profiles of autism-associated genes*. Scientific Reports. 2021;11(1):15269.