John Doe

San Francisco, CA | john.doe@email.com | johndoe.ai | johndoe | johndoe

Summary

AI researcher and entrepreneur with a track record of publishing at top venues (NeurIPS, ICML, ICLR) and translating research into products used by millions.

Currently building Nexus AI, a VC-backed infrastructure company for efficient large model deployment.

Education

Sept 2018 – May 2023 **Princeton University**, PhD in Computer Science – Princeton, NJ

Sept 2014 – June 2018 **Boğaziçi University**, BS in Computer Engineering – Istanbul, Türkiye

Experience

June 2023 – present **Co-Founder & CTO**, Nexus AI – San Francisco, CA

May 2022 – Aug 2022 **Research Intern**, NVIDIA Research – Santa Clara, CA

May 2021 – Aug 2021 **Research Intern**, Google DeepMind – London, UK

May 2020 – Aug 2020 **Research Intern**, Apple ML Research – Cupertino, CA

May 2019 – Aug 2019 **Research Intern**, Microsoft Research – Redmond, WA

Projects

Jan 2023 – present **FlashInfer**

Jan 2021 **NeuralPrune**

Publications

July 2023 **Sparse Mixture-of-Experts at Scale: Efficient Routing for Trillion-Parameter Models**

Dec 2022 **Neural Architecture Search via Differentiable Pruning**

July 2022 **Multi-Agent Reinforcement Learning with Emergent Communication**

May 2021 **On-Device Model Compression via Learned Quantization**

Selected Honors

- MIT Technology Review 35 Under 35 Innovators (2024)
- Forbes 30 Under 30 in Enterprise Technology (2024)
- ACM Doctoral Dissertation Award Honorable Mention (2023)
- Google PhD Fellowship in Machine Learning (2020 – 2023)
- Fulbright Scholarship for Graduate Studies (2018)

Skills

Languages: Python, C++, CUDA, Rust, Julia

ML Frameworks: PyTorch, JAX, TensorFlow, Triton, ONNX

Infrastructure: Kubernetes, Ray, distributed training, AWS, GCP

Research Areas: Neural architecture search, model compression, efficient inference, multi-agent RL

Patents

1. Adaptive Quantization for Neural Network Inference on Edge Devices (US Patent 11,234,567)

2. Dynamic Sparsity Patterns for Efficient Transformer Attention (US Patent 11,345,678)

3. Hardware-Aware Neural Architecture Search Method (US Patent 11,456,789)

Invited Talks

4. Scaling Laws for Efficient Inference — Stanford HAI Symposium (2024)

3. Building AI Infrastructure for the Next Decade — TechCrunch Disrupt (2024)

2. From Research to Production: Lessons in ML Systems — NeurIPS Workshop (2023)

1. Efficient Deep Learning: A Practitioner's Perspective — Google Tech Talk (2022)