Jeet Kanjani

Website LinkedIn Github

jeetkanjani7@gmail.com
+1 412-626-1947

Education

- **Carnegie Mellon University - School of Computer Science** — Pittsburgh, PA
  Master of Science in Computer Vision (4.22/4.33) — *Jan 2021 - May 2022*
- **The LNM Institute of Information Technology** — Jaipur, India
  Bachelor of Engineering in Computer Science — *Jun 2014 – Apr 2018*

Experience

- **Meta AI** — Menlo Park, CA
  Machine Learning Engineer Lead — *Feb 2025 - Present*
  - Led training and inference optimization for early stage ranking on Instagram, directly unblocking two major releases.
  - Built critical components for hierarchical attention caching in foundation ranking models, achieving 2x faster inference.
  - Developed custom Triton and CUDA kernels in Pytorch to accelerate key ops across Meta's recommendation models.
- **Amazon Go, AWS** — Palo Alto, CA
  Machine Learning Engineer Lead — *Jun 2022 - Feb 2025*
  - **Optimizations:** Led optimization efforts of Video Foundation model training, enabling **3x** longer sequences with **2.5x** speedup by stochastic backpropagation, efficient attention kernels, model compilation, improving Tensor Core utilization, optimizing memory-bound operations and upgrading to the latest PyTorch versions while maintaining reproducibility.
  - Developed a QAT pipeline on our 1st gen NNP with a custom backend and shipped 10+ models with quantization-friendly architectural updates, maintaining accuracy parity at 8-bit precision and reducing deployment time from months to days.
  - **MultiModal Transformer:** Co-led the design and development of a multimodal transformer model that analyzed a timeseries of customer hand crops and product descriptions to determine their shopping receipt.
  - Achieved a significant 44% reduction in human-in-the-loop inquiries within six months. This was the first project in our organization to demonstrate the impact of hand-centric modeling and leveraging multimodality.
- **NVIDIA** — Santa Clara, CA (Remote)
  ML Software Intern — *May 2021 - Aug 2021*
  - Integrated modular skeletal-based action recognition pipeline into Metropolis eMDX for low-latency edge devices. [Link]
  - Achieved **2.5X** speedup on Triton Inference Server by writing custom kernel for tensorRT and GPU memory optimizations.
- **Oxehealth Ltd** — Oxford, UK
  Research Engineer — *May 2018 - Dec. 2020*
  - Contributed to the development of ML models for Person tracking, Fallen Person Detection, Person on Edge of Bed detection and Sleep Staging from video. These models resulted in reduction of patients falling from 33% to **48%**. [Link]
  - Achieved a 22% improvement in finetuned Yolov3 mAP by joint learning it with an optical flow based motion model.
  - Led efforts to build inference service with GRPC on Coral TPUs to serve deep learning models in production. Achieved **32%** speedup and reduced deployment cost by **10x**.
- **University of Oxford** — Oxford, UK
  Research Intern — *May 2017 - Sept. 2017*
  - **Research**: Worked with a DPhil student at **Torr Vision Group** on 3D Pose Estimation from Monocular images using structured learning approaches. Improved on previously built 2D Pose Estimator using CRF as RNN. [Link]
  - **Development (QuickHOG)**: CUDA implementation of HOG-SVM based Pedestrian Detection to the **OxSight glasses** used by the visually impaired. Achieved **80x** run time improvement over sequential implementation and **1.2X** over state of the art parallel implementation. Implemented a novel NMS by adopting a *map/reduce* parallelization pattern. [Link]
- **Tonbo Imaging** — Bengaluru, India
  Research Intern — *Jan. 2018 - April 2018*
  - **Research**: Addressed the issue of long term tracking of objects in thermal infrared videos by using fully convolutional siamese networks (SiameseFC) with LSTMs. Achieved a 2.9% AUC improvement on Tonbo's infrared dataset. [Link]

Publication

Far3Det Towards Far-Field 3D Detection
Shubham Gupta*, **Jeet Kanjani***, Shu Kong, Martin Li, James Hayes, Francesco Ferroni, Deva Ramanan
Published at WACV 2023 **Link**

Academic Project

- **Robust Factor Graph Attention Net** — Pittsburgh, PA
  (MMML Course Project) — *Sept 2021 - Dec 2021*
  - **Research**: Conducted experiments to mitigate uncertainty in Visual Question Answering models and make them robust to linguistic variations. Formulated contrastive loss and generated unimodal/multimodal relevance scores for training. Improved on the NDCG metric by **7.3%** over the FGA baseline (Awarded best poster presentation). [Poster][Paper]