DIPIKA KHULLAR

Santa Clara, CA · (661) 301-6451 · dkhullar98@berkeley.edu · Linkedin.com/in/dipikakhullar/

EDUCATION

University of California, Berkeley B.A Computer Science, Major December 2020
B.A. Data Science, Minor

Awards: Berkeley Leadership Award

PROFESSIONAL EXPERIENCE

MATS Scholar ***Berkeley, CA***

- Researched lie detection in LLMs in the Anthropic megastream, exploring methodology to measure and correct self-sycophancy (inflated self-ratings) in harmfulness judgments.

Amazon AGI ***Santa Clara, CA***
Applied Scientist II *Mar. 2022 — May 2025*

- Identified, evaluated, and benchmarked high quality data sources for enhancing Bedrock Foundational Models (FMs), involving exploration and refinement of diverse datasets such as Eureka, Twitter, Wikimedia, arXiv, and Quora, as well as conducting comprehensive data value assessments and ablation studies to quantitatively evaluate data quality and variety. Released PILEv2 and automated future PILE data source refreshes.
- Developed a diffusion-based synthetic data generation pipeline to create diverse, domain-compatible images of scarce road scene objects, improving object detection model training for real-world applications
- Researched strategies to harness latent representations from LLM-decoder of Visual Question Answering (VQA) models like BLIP2, combined with visual embeddings, to achieve substantial performance improvements in few-shot image classification with limited data (1-5 examples per class)
- Proposed and developed a method to generate images by inserting synthetic infrequent objects into real backgrounds using a text-conditioned diffusion model. Utilized a mask generator to create areas for object insertion and ensured domain compatibility by blending synthetic objects seamlessly into real scenes
- Led the proposal, OCR research, modeling, and engineering development for a PDF structured data extraction pipeline to extract data from pdf documents AmazonBot fetches from the web, used in Nova model pretraining
- Led structured content extraction efforts for foundational model pretraining by training scalable models to identify, extract, and clean high-quality code, math, and tabular data from large-scale web and document corpora, significantly improving data quality and diversity for foundation model training.

Qualcomm Corporate Research and Development ***San Diego, CA***
Machine Learning Engineer *Dec. 2020 — Mar. 2022*

- Worked with the framework integration team to open source examples and benchmarks for deep learning model optimization toolkits and federated learning library
- Developed 8 bit quantization and compression methods for CNNs, transformers, and other state of the art models. Experiment with heuristics to identify output channels that will cause a high error term post quantization. Work open sourced as part of AIMET.
- Created a visualization tool, exposing statistics and data used for making model quantization and compression decisions. Analyzed compression performance, visualized evaluation scores and compression ratios to understand the candidates selected by greedy selection algorithms

Square ***San Francisco, CA***
Machine Learning Engineering Intern *Sept. 2020 — Dec. 2020*

- Utilized vector embeddings for categorical data to increase the predictive power of XGBoost loan model for Square Capital. Explored unsupervised learning techniques for misclassified business types
- Integrated new computational flow and modeling framework with the Square Capital loan platform software

Apple CoreML ***Cupertino, CA***
Machine Learning Engineering Intern *May 2020 — Aug. 2020*

- Prototyped and experimented with neural architecture search (NAS) methods for CreateML sound classification
- Exposed a parameterizable MLP head for the image classifier in CreateML, providing a higher capacity alternative to logistic regression. Determined optimal parameters for this new image classification model

- Profiled sound and image classifier performance improvements using a multilayer perceptron (MLP) classifier head over logistic regression with larger dataset sizes
- Determined an automatic selection of classifier heads for both sound and image classification tasks based upon the number of training examples, classes, and balance of a dataset

ACADEMIC EXPERIENCE

Research Fellow: SPAR *January 2025 — present*

- *Evaluated steering vectors to modify and analyze reasoning behaviors in large language models, enabling controlled interventions and insights into model alignment and decision pathways.*

Community Researcher: Eleuther AI *December 2024 — present*

- *Estimating volume basins in parameter space, quantifying how much perturbation weights could tolerate before significantly degrading performance.*

Undergraduate Researcher: Assistant AI Professor Canny's Group *August 2020 - March 2022*

- Working in Professor John Canny's lab on new techniques for OCR, learning methods in order to support a video-to-text translation pipeline
- Investigate and design representation learning algorithms that extract general and meaningful latent features and profile them against current video captioning systems

Research Assistant: Purdue University *August 2020 - May 2021*

- Understanding comparison networks and their structure, with the goal of improving their performance.

RESEARCH PUBLICATIONS

Anomaly Detection for Spatiotemporal Data in Action *Tutorial, KDD 2022*

Synthetic Data Generation for Scarce Road Scene Detection Scenarios *Poster, Workshop on Synthetic Data Generation with Generative AI, Neurips 2023*

Improved Few-Shot Image Classification Through Multiple-Choice Questions *WACV 2024, Poster, Amazon Computer Vision Conference 2023*

P84597: Fully automated and controlled Virtual Product Placement in Images and Videos US Patent

Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation

BLOG POSTS

Create Amazon SageMaker models using the PyTorch Model Zoo AWS Machine Learning Blog
***Dipika Khullar**, Marcelo Aberle, Ninad Kulkarni, and Yash Shah*

Build Streamlit apps in Amazon SageMaker Studio AWS Machine Learning Blog
***Dipika Khullar**, Marcelo Aberle, and Yash Shah*

INVITED TALKS

Domain-Compatible Synthetic Data Generation for Infrequent Object Detection *Amazon Science*

SERVICE

Amazon Computer Vision Conference 2023 (ACVC) *Peer Reviewer*
Winter Conference on Applications of Computer Vision 2023 (WACV) *Peer Reviewer*

Amazon Research Award* *Grant Reviewer Committee*

* *Reviewed research proposals sent in by PhD students and faculty to determine funding for the Amazon Research Awards*