Shaan Pakala

San Jose, CA | Riverside, CA | shaan.pakala@gmail.com | [website, github, linkedin]

EDUCATION

University of California, Riverside. Sept. 2021 - June 2025
Data Science & Engineering, B.S – **3.91** Upper Division GPA

Relevant Courses: Machine Learning & Data Mining, Artificial Intelligence, Database Management Systems, Algorithms for Bioinformatics, Statistics for Data Science I & II, Design of Experiments

RELEVANT EXPERIENCE

Data Science Research Intern – UC Riverside [github] June 2024 - Present

- Submitted paper as first author to IEEE Big Data in December 2024 (under review)
  - Paper title: "Automating Data Science Pipelines with Tensor Completion"
- Funded by National Science Foundation (NSF) Research Experience for Undergrads (REU)
  - Under the guidance of Professors Vassilis Tsotras, Vagelis Papalexakis, and Jia Chen
- Leveraging tensor completion methods to accelerate hyperparameter tuning
  - Includes non-deep learning hyperparameter tuning & deep learning architecture tuning
  - Requires only a small percentage (1-5%) of evaluated models to infer the results of all
- Also applying this to inferring database query cardinalities, using a fraction of computed results
  - Efficiently estimating query results, for the purpose of query optimization

Bioinformatics Research Assistant – UC Riverside Spring 2024

- Bioinformatics Research Assistant under Professor Stefano Lonardi
- Recreated, verified, and extending the work of other bioinformatics research papers
- Fine-tuning large-language models (LLM) for protein cleavage site prediction for drug discovery research
  - Includes gathering & preparing protein sequence data for LLM inputs

Computer Science Grader – UC Riverside Spring 2024

- Grader for Upper Division Data Analysis Methods (CS 105) with Dr. Elena Strzheletska
- Facilitated students' lab and project demonstrations for evaluation
- Graded quizzes and project reports

Data Science Fellow – Lawrence Livermore National Laboratory (LLNL) [github] Summer 2023

- Designed Neural Network for heart disease diagnosis using electrocardiogram (ECG) data (with PyTorch)
  - Improved minority classes' accuracy around 10% with custom loss function
- Reconstructed spatio-temporal map of hearts using a Convolutional Neural Network (using TensorFlow)
  - Mapped ECG data from 12 different leads over 500ms into heart activity of 75 locations over 500ms
- Composed EDA graphs and experimented with normalization techniques to highlight patterns in data
- Presented results at poster session with an audience comprising mainly of LLNL scientists and interns

PERSONAL PROJECTS

Image Segmentation Projects [github]

- Developed custom U-Net implementing skip connections for image segmentation (using PyTorch)
- Segmented images of people, cells, and for multiclass segmentation of bears & deers
- Experimented with different image preprocessing techniques for better results (using OpenCV)

Visualization & Classification of Wikipedia Articles [github]

- Collected & cleaned data from over 400 Wikipedia articles
- Leveraged Google's pretrained BERT model to generate hidden states
- Mapped 2D hidden states to just 2 scalars (features) through linear combinations & t-SNE
  - Scatter Plot of the 2 scalars displayed the dissimilarity between articles of different topics
- Used a Support Vector Machine on the transformed hidden states for topic classification (92%+ accuracy)

SKILLS

Coding Languages: Python, SQL, R, C++, Java, Tableau
Libraries: PyTorch, TensorFlow, tensorly, scikit-learn, OpenCV, Pandas, Numpy, matplotlib, seaborn, tidyverse