Vyom Pathak

San Francisco, CA, USA
+1(925)6605503 | angerstick3@gmail.com | linkedin.com/in/01-vyom | 01-vyom.github.io | github.com/01-vyom

Experience

Amazon, *Applied Scientist* November 2024 - Present
Core Search

- Formulating scaling laws using model size and data composition in LLM Semantic Matching models (1B–10B scale)
- Spearheading **automated evaluation** and query benchmarking to **iteratively refine LLM Semantic Matching**

Chronograph, *Applied Scientist* September 2023 - Present
Applied NLP research in financial domain

- Developed an **instruction tuned Llama 3.1 8B Quantized model** for **information extraction** in **JSON format** with an exact score of 80% on out of domain dataset
- Built an end-to-end **DeepSpeed** framework for distributed multi-GPU training across all stages, **tokenization** to **downstream tasks**, achieving a 40% improvement in model finetuning efficiency for **8B scale models**
- Explored the application of **in-context learning** for **information retrieval from long financial text** using **Claude 3.5 Sonnet** showing 3% improvement compared to previous production models
- Deployed a **longformer** based model for **long-context multi-class classification** improving upon the previous production model's performance by 9%

Data Science Research Lab, *Research Assistant* June 2023 - Present
ECOLE a DARPA funded Continuous Learning System (Mentor: Daisy Wang)

- Explored the application of **reinforcement learning from human feedback** for improving Knowledge Graphs and Scene Graphs
- Researched improvement, and faithfulness of multi-modal reasoning using **in-context learning for large language models**

Learned Bloom Filter for High-Dimension Similarity Join (Mentor: Daisy Wang) December 2022 - May 2023

- Orchestrated several experimental designs for training **SelNet, Mixture of Experts (MoE), XGB, LightGBM, and Support Vector Regressor** on 3 text based and 3 image based embedding datasets using **distributed training framework** on Pytorch
- Composed the introduction, related work, and experimental details sections of the research paper for baseline experiments using Latex

Amazon, *Applied Science Intern* August 2022 - December 2022
Alexa Smart Home Team (Mentor: Sven Eberhardt)

- Removed **redundant features** by performing feature importance analysis using **visualization techniques, and online model regression analysis**
- Adopted a novel **self-attention** based architecture for **sequence classification task** to model user behaviors
- **Interpreted attention scores** to find insights on important segments from the history

Apple, *Machine Learning Research Intern* May 2022 - August 2022
Siri Text to Speech Team (Mentor: Kishore Prahallad)

- Implemented a novel **end-to-end acoustic model for text-to-speech synthesis**
- Built a data synthesis and training pipeline to train deep learning models on **large scale speech corpus (500 hours)** using multi-GPU training framework on MLflow

Data Science Research Lab at UF, *Research Assistant*
Multi-Answer Question Answering Benchmarking Dataset (Mentor: Daisy Wang) March 2022 - May 2022

- Annotated 1,000 QA pairs for the **multi-answer QA benchmark dataset**, to **track misinformation, and disinformation**
- Performed **Tweet stance annotation** for 1,000 samples QA pairs from Twitter API Querier using **Dense Distinct Tweet Retriever (DDTR) model**

AIDA a DARPA funded Hypothesis Generation System (Mentor: Daisy Wang) September 2021 - March 2022

- Built pipeline for **cross-lingual Natural Language Inferencing using mT5 model** trained on multiple A100 GPUs, improving the recall score by 38%
- Extracted **sentence embeddings** from **100,000 sentences** using **XLM Roberta for similarity clustering**

Publications

- **Pathak, V.**, Bhatt, B., Sahay, A. and Raman, M., 2021, December. **Neural Network Based Retrieval of Inherent Optical Properties (IOPs) Of Coastal Waters of Oceans**. In 2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS) (pp. 285-288). IEEE. doi.org/10.1109/InGARSS51564.2021.9792013
- Raval, D., **Pathak, V.**, Patel, M. and Bhatt, B., 2021. **Improving deep learning based automatic speech recognition for Gujarati.** Transactions on Asian and Low-Resource Language Information Processing, 21(3), pp.1-18. dl.acm.org/doi/full/10.1145/3483446
- Raval, D., **Pathak, V.**, Patel, M. and Bhatt, B., 2020, December. **End-to-End automatic speech recognition for Gujarati**. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 409-419). aclanthology.org/2020.icon-main.56

Projects

Preliminary Survey on Foundation Language Models, *Research Project* January 2023 - May 2023

- Performed a **thorough analysis on large language models**, and wrote a 9 page research report
- Trained **10** large language models on **NVIDIA A100 GPU** using Pytorch in a **distributed manner**

mRNA COVID-19 vaccine degradation prediction, *Kaggle Genomics Project* January 2022 - May 2022

- Established a **hybrid Bi-LSTM Bi-GRU model** to achieve good MCRMSE score of 0.3577 over 5 column values
- Demonstrated that **graph-based architectures better capture sequential patterns in mRNA**, with a **10% improvement in MCRMSE score**

Image based melanoma detection, *Medical Research Project* January 2022 - May 2022

- Attained a **sensitivity score of 92.4% by ensembling ResNet and EfficientNet based models** by **finding threshold margin maximized over G-Means value**
- Evaluated the application of **self-supervised pre-training based on Bootstrap Your Own Latent (BYOL) model**, achieving an **increase of 3% for both ResNet and EfficientNet based models**

Schema based dialogue system, *Spoken Dialogue Systems Research Project* September 2021 - December 2021

- Invented a **zero-shot dialogue system** using a **schema-guided attention model** for **wire-framing dialogue systems**
- Designed **robust intent extraction module** based on **Dialog-GPT2 with decent priming** for final round robin evaluation achieving an **average User Satisfaction of 7.3**
- Systematized the user study protocol for evaluation on 20 users

CommonLit Readability Prize, *Kaggle NLP competition* May 2021 - August 2021
Silver medal · 106th/3566 (Top 3%)

- Formulated **2D attention algorithm for Roberta large** increasing the performance by 15%
- Experimented with fine-tuning techniques by implementing **differential learning rate, gradient accumulation, and custom attention heads** to attain a competitive RMSE of 0.460
- Utilized **Forward Selection OOF to ensemble various models** and boost the overall RMSE to 0.4588 by generalizing the target value

Open-Source Contributions, *Python Libraries* April 2021 - July 2021

- Added ground-up implementation of **Spec-Augment**, and performed **bug fixes**, and **documentation fixes** (**Hugging Face**, PR #11614, PR #11752)
- Updated **loss metric**, and **fixed bugs** (**Pytorch Ignite**, PR #2027, PR #2116)
- Added new **visualization algorithm**, **unit tests**, **bug fixes** and **documentation fixes** (Optuna, PR #2834, PR #2806, PR #2712, PR #2711, PR #2710)

End-to-End Speech Recognition for low-resource language, *NLP Research Project* December 2019 - March 2021

- Tailored **beam search decoding** by introducing **multi-level language modeling**, reducing the word error rate by 2.1%
- Innovated **spell correction post-processing using BERT language model**, outpacing the previous performance by 3%
- Performed **word-level analysis** on model output for **quality assurance**, and **iterative improvement** of the ASR system by 5.1%

- Scrapped and forged **316M word corpus** for developing **language models using KenLM package with ablation study**

Education

University of Florida, Gainesville, FL, United States August 2021 - May 2023
Master of Science in Computer Science (*Machine Learning specialization*) (*GPA: 3.85/4*)

Dharmsinh Desai University, Nadiad, Gujarat, India July 2017 - May 2021
Bachelor of Technology in Computer Engineering (*GPA: 9.01/10*)

Honors & Clubs

- Paper reviewer for 29th International Conference on Neural Information Processing (ACM ICONIP 2022)
- Mentored a group of 5 juniors as the ML Team Head at Developers Student Clubs (DSC) by Google Developers
- Arranged and taught 6+ seminars and workshops on various machine learning concepts at DSC
- Achieved 1^{st} place in university for ACM-ICPC, Gwalior-Pune regionals online qualifier

Skills

Programming Languages & Databases: Python, C, C++, Matlab, Latex

Framework & Libraries: Pytorch, Tensorflow, Keras, Librosa, Hugging Face, MLflow, Airflow

Tools & Services: Git, Github, GCP, AWS, CUDA, Docker, CUDA, HPC, Kubernetes