Ethan Kim

(617) 758-9553 | ethan_kim@college.harvard.edu | New York, NY, Github: https://github.com/ethankim00/ | Site: https://ethankim00.github.io./

EDUCATION

HARVARD UNIVERSITY

Bachelor's in Statistics, Secondary in Computer Science

Graduated December 2021, SAT: 2370/2400 (V800, M800, W770)

EXPERIENCE

CYNDX | Data Scientist — Nov '21 – Present

- Train transformer based Dense Information Retrieval models to replace existing search algorithm
- Employ novel scoring methods to improve search precision on long documents by over 300% over previous algorithm
- Train and deploy models on GCP including text classification, keyword recognition, Named Entity Recognition, summarization and marine translation. Upgrade internal infrastructure for distributed training of pytorch models
- Optimize models for inference efficiency employing quantization, teacher-student distillation and ONNX runtime
- Architect full stack workflow for machine learning using Kubeflow Pipelines, DBT, Cloud Run and Triton Inference Server
- Develop pipelines to process and extract features from millions of documents using Spark
- Standardize internal workflows with common python packages, pipeline components and CI/CD
- Spearhead implementation of SHAP based explainability features into customer facing models

ORACLE | Data Science Intern — May '21 – Aug '21

- Design and Implement architecture to Incorporate SHAP value-based model interpretability into Oracle's Marketing ML platform
- Design standard interface to deploy explainability into all internal and client models including including XGBoost, Random Forest, and NLP models
- Research and Implement methods to explain Recommender Systems

HARVARD AI4LIFE LAB | Research Assistant

- Research extensions and limitations of SHAP and LIME methods with professor Hima Lakkaraju.

THINKCERCA | Data Science Consultant — May '21 - Aug'21

- Fine-tune BERT models for Automatic Essay Scoring and Feedback

BGI LLC | Data Science Intern — Sep '20 – Jan '21

- Develop transfer learning and data augmentation techniques for Speech to Text Transcription using Mozilla DeepSpeech to achieve over 90% Character Accuracy Rate

HARVARD COLLEGE DATA ANALYTICS GROUP | Director of Consulting — Jan '20 - Dec '21

- Manage 24+ Data Science consulting teams working with nonprofits, startups, and Fortune 500 companies
- Source clients to help student club grow 5x to 100+ members with $130,000 in revenue
- Provide technical guidance to case teams on projects ranging from marketing analytics to using NLP to scrape financial databases to billing anomaly detection
- Lead team of six on Data Science research project to develop live win probability prediction models for a sports analytics company. Deploy efficient model for real time inference.

SKILLS

Python, Scikitlearn, TensorFlow, Pytorch, XGBoost, SQL, R, git, PySpark, Docker, AWS, Google Cloud Platform, Kubeflow

PUBLICATIONS

Google Big Bench

- Develop task to probe reasoning capabilities of Large Language Models

Differentiable Entailment For Parameter Efficient Few Shot Learning

- First Author paper achieve high performance for text classification the few shot learning regime

Prompt Tranfer For Domain Adaptation in Neural Information Retrieval

- Explore using soft prompts for domain transfer in information retrieval
- Novel method achieves strong zero shot out of domain performance on information retrieval benchmarks

Bloom: A 176 Billion Parameter Open Access Multilingual Model

- Principled Data filtering to create corpus to train large multilingual language model

PROJECTS

Involution

- Pytorch implementation of Convolutional Operator alternative

Parallelizing OpenCV for Real Time Object Tracking

- Use OpenMP and C++ to Parallelize multiobject tracking across multiple cores to run in real time