Malcolm **Greaves**

Machine Learning Software Engineer

U.S. Citizen, New York City

(414) 704-9696 | greaves.malcolm@gmail.com | www.malcolmgreaves.io | malcolmgreaves | malcolm-greaves-49959919

Experience

Scale AI, Inc.

San Francisco & New York City

Senior ML Infrastructure Engineer — *March 2020 - Present*

- Created Scale's first document processing models and supporting infra., securing a $1M contract that led to the creation of Scale Document.
- Re-wrote model serving architecture across 4 different ML teams, each saving between 40% and 66% in inference costs: saving $200k annually.
- Creator, GPU utilization working group: up leveled other ML eng to optimize their team's service from 20% to 86% GPU utilization.
- Advised and mentored ML engineer who optimized Kubernetes service start time from 2 miuntes to under 30 seconds.
- Created ML model serving framework used by all ML engineers and hundreds of internal and external services.
- Optimized internal LLM services, achieving 10x speedup in end-to-end inference.
- Re-architected realtime image embedding generation & k-NN search pipeline with 10M+ requests/day with efficient vector database, reducing operating costs by 50%.
- Optimized batch embedding generation service used by multiple internal teams, decreasing costs by 5x while keeping overall running time constant.

Change Healthcare

Emeryville, CA

Senior Data Scientist — *November 2018 - Februray, 2020*

- Developed deep NN model for predicting missing charges & deployed to US's largest medical insurance payment network.
- Increased scientist productivity by 25x with internal model training & serving framework.

Volley Labs, Inc.

San Francisco, CA

Senior Research Engineer, ML Tech Lead — *December 2016 - October 2018*

- End-to-end ownership & development of multiple-choice question generation system from unsupervised to supervised neural network approach. Product-differentating feature that helped Volley stand-out and make its first sale to JPMC.
- Technical lead of engineering and ML team. Data pipelines in Airflow; deployed ML models using Keras, Spacy, Tensorflow; Python 3.

Nitro Software, Inc.

San Francisco, CA

Research Engineer — *March 2015 - October 2016*

- Created novel machine learning based solution for automatic form field detection (FFD) and semantic classification. End-to-end service development & deployment to Nitro Cloud.

Alpine Data Labs

San Francisco, CA

Software Engineer, Machine Learning — *Jun. 2014 - Mar. 2015*

- Algorithm optimization for large-scale data processing & modeling in Spark: L-BFGS, random forest models, feature encoding, etc.

Education

Carnegie Mellon University

Pittsburgh, PA USA

B.Sc. and M.Sc. in Computer Science — *Aug. 2009 - May 2014*

- Graduated with School of Computer Science Honors. GPA 3.5 (undergrad and graduate).
- Master's Thesis in semantic relation extraction from unstructured text with probabalistic logic & SVMs: `http://goo.gl/DzMr6c`

Work Portfolio

Languages	*Proficient*: Python, Scala, Java; *Moderate*: BASH, SQL, Go, C; *Familiar*: Typescript, C++11, Rust, LaTeX
Machine Learning & Data	Pandas, PyTorch, HuggingFace, NVidia Triton, Deepspeed, torchvision, torchaudio, TensorFlow, Numpy, Scipy, Scikit Learn, Onnx Matplotlib, Redash, Spark, OpenCV, AWS SageMaker, Airflow, Dagster
Infrastructure & Backend	Kubernetes, Terraform, AWS, CircleCI, Datadog, Kubecost, Snowflake, Postgres, MongoDB, qdrant, Flask, FastAPI, OpenAPI, Protobuf, DynamoDB, Temporal, Celery, Redis, Elasticache, Kafka, Akka