Cory Pruce

corypruce [at] gmail [dot] com • cpruce.github.io • Cpruce

Experience

Robinhood Markets — **Menlo Park, CA**
Senior Machine Learning Engineer — *April 2021–October 2022*

- Low-latency, highly-available, fault-tolerant Golang server handling gRPC requests to serve real-time models and store relevant data. Deployed with Kubernetes, stores in Postgres and S3, and continually tested and deployed with CI on Apollo and Jenkins and CD via internal service.
- Created and deployed production binary classifier that uses the TensorFlow Universal Sentence Encoder to produce embeddings for feature extraction and dimensionality reduction, which are subsequently fed into a Logistic Regression model in sklearn. Model helps significantly reduce Account Takeovers by leveraging direct email texts to predict requests that need 2fa when deterministic rules are not possible. Best version achieves 0.89 Recall and 0.91 Precision on a validation set of over a million samples. Offline, batch model scheduled in an Airflow DAG on a Hadoop cluster.
- Low-cost Python data annotation service that interfaces with AWS SageMaker GroundTruth and our private Identity Provider. Implemented frontend GUI's with jQuery to dynamically show inputs and outputs, storing results in S3 and PrestoSQL.
- Wrote LDA and embeddings topic modelling and visualization data analysis of customer ticket texts to support recommendations on user issues.
- Administrator of team's AWS subaccount, minimizing costs and upholding least-privileged model while facilitating development using Terraform.

Amazon Web Services — **East Palo Alto, CA**
Software Development Engineer II — *May 2018–April 2021*

- Advanced the state of the art in image synthesis from 4.4 to 3.6 FID50k score by observing that the Adaptive Instance Normalization (AdaIN) operations in the StyleGAN architecture were hindering convergence. StyleGAN authors later addressed the AdaIN issue in the StyleGANv2.
- Led team of 3 in developing the scalable and reliable cluster management product and CI/CD for SageMaker distributed data parallel on EC2.
- Reduced shortest time-to-train of StyleGAN from 6 days 14 hours to 1 day 23 hours via data-parallelism over NVIDIA V100's using Horovod.
- Produced the 2nd best internal throughput (~45k sequences per second) and time-to-train for BERT by leveraging BytePS. Throughput/time-to-train could be improved further once the Elastic Fabric Adapter is supported by BytePS.
- Developed, tested, and deployed key data structures, methods, and automated tests for AWS TensorFlow Elastic Inference service in both Python and C++, helping reduce inference costs by up to 80%. Highlighted cost reduction in a cost analysis blog.

Parallel Machines — **Sunnyvale, CA**
Machine Learning Software Engineer — *June 2017–April 2018*

Juniper Networks — **Sunnyvale, CA**
Software Engineer II — *February 2016–June 2017*

- Identified a >$3million/yr reduction in maintenance costs through automating the triaging process. Designed ETL, feature engineering, and prediction model pipeline around LinearSVC+TF-IDF, which outperformed Logistic Regression, Naive Bayes, and Random Forests. Top-1 accuracy achieved over 50% and Top-5 accuracy surpassed 80% for thousands of classes. Using the Wagner-Fischer dynamic programming algorithm, created a personalized variation of the Damerau-Levenshtein distance for hierarchical-category degrees of separation.

Projects

Mobile Mask RCNN — **Online**
Contributor and Author — *December 2017–Present*

- Worked in a team to convert the Keras/Tensorflow Mask RCNN's Detection Layer to Tensorflow operations for cross-environment serializability.
- Debugged aforementioned Mask RCNN model to be compatible with Tensorflow Mobile, extracted Tensorflow Protobuf model, and successfully exported more operations to the built libtensorflow_inference.so library.
- Replaced FPN's ResNet50 backbone with MobileNet for lower latency on embedded devices, reducing inference time to 62% of the original speed.

Kaggle — **Online**
Competitions Expert — *January 2017–Present*

- Human Protein Atlas Image Classification: Weighted average ensemble of Xception, InceptionResNet, and DenseNet scored 0.542 macro-average F1-score on test set. Top model found was DenseNet121 on image resolution 512x512. Attained 37th of 2172 teams as individual competitor.
- 2018 Data Science Bowl: Fine-tuned Mask R-CNN with ResNet101 backbone. Placed 321st of 3634 teams as an individual.
- Planet Aerial Image Classification: Ensembled pretrained Resnet[18, 34, 50], Densenet[121, 169, 201], VGG[16, 19], and Inception[V3, V4] Pytorch models which ended with a 17-class sigmoid layer to predict whether a label exists in the aerial image. Tried cyclic learning rates and test-time augmentations such as random crops, flips, and rotations. Team achieved 9th of 938.
- Carvana Image Masking Challenge: Trained Unet (256, 512, and 1024 resolutions) using dice coefficient loss function for semantic segmentation of car images, reaching 79th of 735 teams. Used run-length encoding lossless compression for submissions.
- Cervical Cancer Screening: Wrote feature selection notebook using OpenCV k-means clustering, thresholds with contour visualizations, and hue ranges to create cropping masks for a given cervix image. Found non-aggressive, generalizable solution for black borders and some instruments by shaping the max perimeter bounding box formed from the contours' bounding boxes. Notebook inspired several others in the competition.
- NIPS Adversarial Attack. Demonstrated targeted and non-targeted Generative Adversarial Networks in order to fool the best discriminative models. Iterative-Fast Gradient Sign methods achieved over 60% incorrect classifications against the PyTorch ImageNet-pretrained InceptionV3 model.

Education

Carnegie Mellon University — **Pittsburgh, PA & Mountain View, CA**
Master of Science in Information Technology, Boeing Scholar & INI Scholar — *May 2014–December 2015*

Pitzer College, Pomona College track — **Claremont, CA**
Bachelor of Arts in Computer Science, Pitzer Grant Scholar — *August 2010–May 2014*

Publications & Patents

Herring: Rethinking the Parameter Server at Scale for the Cloud — **Amazon Web Services**

Framework integration for instance-attachable accelerator — **Amazon Web Services**