Pengcheng Ding

(667) 803 0823 | pding@fastmail.com | www.linkedin.com/in/pchding | github.com/pchding

SKILLS

Languages: Python, R, SQL (BigQuery), and Bash
Tools: Tensorflow, PyTorch, Scikit-learn, MLflow, Airflow, Kedro, PySpark, Vertex AI suite (GCP), Git
Applications: Machine learning (statistical/deep learning, weakly supervised learning, computer vision, NLP), Statistics / optimization (Stochastic approximation/programming, A/B testing, (non-)linear optimization, and robust optimization)

EXPERIENCE

Machine Learning Scientist II, *Wayfair,* Boston — Sep 2021–Dec 2023

R&D for Model Advancement to Improve Catalog Data Quality and Product Findability

- Developed novel training methods to overcome challenges with scarce true labels by integrating noisy label pre-training, programmatically generated labels, and human annotations. Achieved 20% average precision (AP) increase over existing models (based on CLIP) and 40% F1 increase against training on noisy labels alone.
- Created custom training loops to obtain optimal task grouping in multi-task models by identifying how tasks are impacted by updates to shared model layers. Led to 10 percentage point AP increase.
- Improved style models to include KNN-based tabular features. Increased accuracy by 2-5 percentage points and reduced the need for frequent retraining due to improved model robustness. Designed and verified positive revenue impact through A/B testing.
- Optimized image shot angle prediction models by designing label encodings to stakeholder needs, enhancing model reliability for downstream tasks and reducing the severity of prediction errors by 20%.

Production Pipeline and Training Framework

- Crafted end-to-end batch prediction pipeline in Airflow for predicting product tags. Improved resource utilization on shared Spark clusters and implemented automated failure recovery, reducing recovery time from hours to minutes.
- Designed and implemented a scalable end-to-end training framework for multi-objective image attribute extraction and classification models using either raw images or embeddings, enabling rapid training on tens of millions of images, resulting in reduction of total model building time from days to hours. Key contributions include: Developed a PySpark-based preprocessing pipeline for image and label acquisition, streamlining data preparation and converting data into efficient tf-records format. Built a modular training package in TensorFlow utilizing CNN orTransformer backbones. This package features flexible training loops, adaptable model classes, and customized loss functions for optimal performance.

Data Science Fellow, *Insight,* New York — May 2020–Aug 2020

- Assisted Alphabet Health to increase client accessibility to Multiple Sclerosis literature by building their first keyphrases extraction models for literature abstracts.
- Implemented LSTM and Bi-LSTM-CRF sequence-to-sequence named entity recognition models for extraction. Utilized multiple embeddings to improve performance. Built Elasticsearch backend for storage and clio-lite based Streamlit front-end to conduct contextual search.
- Enabled training of extraction models on any search term on PubMed, augmenting abstracts with extracted key phrases through end-to-end pipeline implemented in MLflow.

Research Assistant, *Johns Hopkins University,* Baltimore — Jul 2015–Aug 2021

- Designed pricing incentives to affect supply-demand equilibrium of power markets to recover infrastructure costs. Simulated electricity transportation network and its effects on market equilibrium. Determined optimal network expansion decisions to correct distortions from these prices in game theoretic bi-level optimization models. Built solution heuristics and solved models using HPC clusters.
- Aided World Bank to finance its multi-nation infrastructure investment by allocating benefits of international trading and planning using cooperative game theory in optimization models and calculating Shapley values.

EDUCATION

Johns Hopkins University, Baltimore, MD — Sep 2021
Ph.D. Power System Operations Research and Economics

École Polytechnique, Palaiseau, France — Jan 2014
M.S Energy and Environment

Nanjing University, Nanjing, China — Jun 2012
B.S. Physics