Pengcheng Ding

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

SKILLS

Languages: Python, R, SQL, and Bash
Tools: Tensorflow, PyTorch, MLflow, Airflow, Kedro, PySpark, Vertex AI suite
Applications: Deep learning (CNN, transformer, pre-training, weak supervision, computer vision, NLP, multi-task learning), Optimization (stochastic approximation/programming, (non-)linear optimization, and robust optimization)

EXPERIENCE

Machine Learning Scientist II, *Wayfair,* Boston — Sep 2021- Present

- Pipeline and Training Framework
  - Crafted end-to-end batch prediction pipeline with Airflow for predicting product style from image embeddings. Improved resource utilization on shared Spark cluster and implemented automated failure recovery, reducing recovery time from hours to minutes.
  - Instituted training framework for multi-task image label prediction using either raw images or embeddings, resulting in reduction of data processing time from days to hours:
    - Introduced PySpark-based preprocessing package for image and label acquisition, converting them into tf-records datasets.
    - Provided a modular training package in TensorFlow, complete with custom training loops, adaptable model classes, and tailored loss functions.
- R&D for Model Advancement
  - Innovated training approach combining noisy label pre-training, programmatically generated labels, and human annotations, realizing a 20% boost in average precision (AP) over existing models and a 40% F1 increase against noisy labels.
  - Devised custom training loops for optimal task grouping in a multi-task framework by identifying how tasks are affected when shared model layers are updated. Led to 10 percentage point AP increase for specific tasks.
  - Enhanced style prediction accuracy by 2-5 percentage points using features from both image embeddings and k-nearest human-annotated labels, improved model robustness through multi-task losses. Minimized product onboarding time by negating the need for frequent retraining.
  - Engineered loss function to counteract the challenges of missing labels during multi-label training, consistently gaining +2 percentage point increase in AP.

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 tagging 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

- Design add-on charges to affect market equilibrium of power auction market to recover infrastructure costs. Determine pricing structures and make proactive planning decisions to correct distortions from such charges in game theoretic bi-level optimization models. Build solution heuristics and solve model using HPC cluster.
- Aided World Bank to finance its multi-nation infrastructure investment by allocating benefits of coordinated trading and system planning using cooperative game theory and calculating Shapley values.

EDUCATION

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

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

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