Jian (Ken) J. jjian03@syr.edu | (315) 802 9717 Cupertino, CA

LinkedIn: https://www.linkedin.com/in/jian-jian Website: https://jianjian-portfolio.github.io

SUMMARY

A machine learning data scientist with extensive experience in artificial intelligence and software engineering. Adept at analyzing, building, and optimizing systems that consolidate Big Data, Deep Learning, Computer Vision, NLP, and Regression & Forecasting models. Guided clients to invent AI algorithms that turn their imaginations into applications. Orchestrated machine learning and deep learning methods to software engineering that productionize the AI technologies in life science, intellectual property management, medical research, social information science, and other industries.

TECHNICAL SKILLS

Languages:	Python, Java, SQL, Shell, JavaScript, R, C, C++, Octave	
Databases:	MongoDB, Federated Queries, SQL Server + T-SQL, Oracle + PL/SQL, SQLite, MySQL, Postgres	
Algorithms:	Deep Neural Network, Sequence Model, BERT, Generalized Linear Model, Support Vector Machine (SVM) Transfer Learning, Principal Component Analysis (PCA), Structure Equation Model (SEM), Factor Analysis	
Frameworks	**- Data Science:**	PyTorch, TensorFlow, Keras, Pandas, NumPy, scikit-learn, SciPy, SpaCy, PySpark, matplotlib
	- Backend:	Domain Driven Design (DDD), Aspect Oriented Programming (AOP), Reactive Programming Event Driven Architecture, CQRS, 3-Tier Architecture Design, Test Driven Design (TDD) Django, Flask, Spring Boot, EJB, SQLAlchemy, Mybatis, PyTest, JUnit, Mockito, Gradle, Maven
	- Frontend:	HTML5, Ajax, Twitter Bootstrap, Backbone, RequireJS, JQuery
Midware:	Spark, Kubernetes, Kafka, OpenMQ, Tomcat, Jetty	
CICD:	GoCD, CircleCI, Jenkins, Bitbucket Pipeline	
Cloud Services:	GCP, AI Platform, Cloud Composer (Airflow), Dataflow, BigQuery, GCR, GCS, GKE, PVM(TPU), App Engine	

WORK EXPERIENCES

Deepcell, Inc Menlo Park, CA
Senior Machine Learning Engineer 2021/10 – Present

- Productionizing deep learning models to detect confounding factors in images that improve the model generalization ability
- Advancing the model performance by creating a benchmark matrix and fine-tuning hyperparameters of deep learning models
- Introducing an evaluation pipeline, preventing models from generating non-deterministic results to improve product quality
- Collaborating with data scientists on accelerating the inference speed of deep learning models with TensorRT/ONNX format
- Integrating Airflow and Beam to build a model deployment pipeline that brings values from the research side to users' end
- Devising an infrastructure to manage machine learning models that reconcile ML/DevOps

AirDNA Denver, CO
Data Scientist (Part-time / Remote) 2021/09 – 2022/09

- Studied Airbnb data, applied SEM, PCA, and factor analysis to model customers' behaviors in the short-term rental market
- Identified critical factors to the level of informativeness, helpfulness, and persuasiveness
- Adapted BERT pre-trained model to differentiate customers' preferences over time
- Strategized and illuminated the market team on improving the trustworthiness of business entities by analyzing the firm-level data
- Fine-tuned machine learning algorithms to identify behaviors among user groups in sharing economy market

IPwe, Inc Dover, DE
Data Scientist Intern 2021/02 – 2021/04

- Designed and implemented a search engine with funnel architecture that improved the search accuracy and system SLA
- Developed a BERT-based knowledge graph that addressed selection bias issues in the ranking stage

Syracuse University (SOS+CD Laboratory) Syracuse, NY
Data Scientist 2020/01 – 2021/05

Active Network, LLC Dallas, TX / China
Senior Java Software Engineer 2016/01 – 2019/06

NCS Pte. Ltd. Singapore / China
Senior Java Software Engineer 2011/03 – 2015/12

EDUCATION

Syracuse University, School of Information Studies Syracuse, NY
M.S., Applied Data Science 2019/08 – 2021/05
Chongqing Technology and Business University, School of Finance China
Bachelor., Economics 2004/09 – 2008/07

PUBLICATIONS

Predicting the longevity of resources shared in scientific publications (Under Reviewing by HSSCOMMS) 2020/07 – Present
This project is funded by the US's Office of Research Integrity grant (ORIIR190049). In this paper, we study a range of explanatory features related to the publication venue, authors, references, and where the resource is shared. We analyze an extensive repository of publications and, through web archival services, reconstruct how they looked at different time points. To set a benchmark, we build a group of tree-based models - Logistic Regression and Random Forest Algorithm. We further inspect the dataset using Lasso, Ridge, and Elastic Net Regression. In response to the right-truncated dataset, we advance the Tobit survival analysis by attaching an elastic net to it. We discover that the most important factors are related to where and how the resource is shared. By examining the places where long-lasting resources are shared, we suggest that it is critical to disseminate and create standards with modern technologies. Finally, we discuss implications for reproducibility and recognizing scientific datasets as first-class citizens.

PROJECT EXPERIENCES

AI Platform for Single Cell Taxonomy 2021/10 – Present

- *Python3, K-means, Deep Learning, Confounder Identification, GCP, Tensorflow2, PyTorch, Sci-Kit Learn, TPU, Apache Airflow, Apache Beam*

By sorting and analyzing cell morphology, this AI platform allows researchers to characterize the composition of tissues at the cellular level. As a senior machine learning engineer, I collaborate with data scientists on fine-tuning deep neural models on the Google Cloud Platform. We experiment with various network architects to sort, clean, and classify cell images. To leverage the computational resources on the cloud, I employ preemptible TPU (a cheaper computational unit on GCP) to train the model. Also, the hyperparameter tuning tool I devised offers data scientists a convenient way to schedule & monitor training jobs and their performance. We apply TensorRT and ONNX model formats to compress and translate the original network into user-ready features. In addition, the model management service I built ships the value from the data science team to the user end. And our product solution integrates the homegrown instrument with cloud computing resources, which assists biologists and other cross-functional teams in researching and exploring unknown cell clusters.

Search Engine – Information Retrieval for BioTech Patents 2021/02 – 2021/04

- *Python3, Scispacy, NLTK, Pandas, JavaScript, HTML*

IPWe, Inc is a startup company that offers a system to promote trades in the intellectual property industry. Its patent searching engine is the key feature of this platform. As the leader of 3 members, my team was responsible for inventing algorithms to filter for the candidate documents and feeding them to the funnel architectural system. To improve the searching accuracy and meet the system SLA, we applied logistic regression and chose recall as its evaluation matrices to get as many relevant documents as possible. In addition, we integrated the ETL pipeline with a public database and BERT-based NER taggings to enrich the candidate documents to reduce the model's selection bias at the next stage. With the extracted species information that connects different patents, this project expanded the accessibility of intellectual acquisitions and improved the user's online patent searching experience.

Data Augmentation for X-Ray Screening Images 2020/08 – 2020/12

- *Python3, GAN, TensorFlow, Sobel Filter, Data Augmentation, Cloud Computing, NHST*

This project was to offer a solution to improve the AI model's generalization ability. Previously, our classification models were performing badly on the test dataset. After a few tries and errors, my research team decided to experiment with the generative adversarial network (GAN) architecture to augment our dataset. To set a baseline performance of the GAN, I used a stack of CNN and batch normalization layers to build the generator and discriminator in GAN. Then I added another edge detecting layer to that original network and conducted the NHST to compare the performance between the 2 models. To cope with the GPU resource limitation during the training, I improved the training program by introducing TensorFlow’s checkpoint as a snapshot of the training progress. Once trained, we asked a few labelers to participate in a double-blinded test on distinguishing a sample where we mix the generated images with the authentic ones. And the difference in the accuracy rate between the human and the GAN's discriminator was not significant. Finally, we enriched the original dataset with synthetic images as the negative observations and alleviated the model's variance issue.

Foundation Email & File Storage Service 2016/11 – 2019/06

- *Java, Spring, Reactor, WebFlux, JQuery, Twitter Bootstrap2, Jersey2, Apache-httpclient, Jackson, Google.Guava, Mockito, PowerMock, Hibernate4, MyBatis, Gradle, MemoryAnalyzer, Token Bucket Algorithm, Python3, openMQ, NIO, Jersey2, Jetty*

This application was a part of the enterprise's internet infrastructure service. As a Senior Java engineer, I mentored juniors' on-job training, PR reviewing, developing, and benchmarking the system performance. In this project, I integrated and facilitated the CPaaS Twilio and Sparkpost platforms into the Spring Boot framework. I devised an SMS sending service in a batch mode to advance the system performance. I refactored the application API to better collaborate with other microservice clusters. I reorganized and simplified the metadata of the distributed infrastructure system, redefined the JMS message queue to decentralize the system cache, and scheduled tasks to periodically purge the NFS space to maximize the usage of hardware assets in the company. Based on the performance report generated by JMap and JStack toolkit, I benchmarked the system's throughput, which offered us the clue and direction to the system optimization. After the refactoring, the throughput of the infrastructure system was 312.73% times higher than the legacy system. Thus, the company could spare more computational resources for other businesses.

OTHER ENGINEERING PROJECTS

Market Research	**Customer Review Analysis for Airbnb**	2021/09 – 2022/09
Open-Source Project	**ImageAnnotatorJS**	2020/05 – 2021/05
Scientific Research Project	**Image Labeling System**	2020/04 – 2021/05
eCommerce Project	**eCommerce Service Platform**	2016/01 – 2019/06
Hospital Information System	**Mount Alvernia HIS**	2014/05 – 2015/12
Hospital Information System	**SATA Community Medical Record System**	2012/07 – 2015/12
Company Registration System	**British Virgin Island (BVI) Business Entity Management System**	2011/03 – 2012/07
ERP	**NEC Software Project Management Platform**	2010/08 – 2011/02
Library System	**NEC Library System at Japan**	2010/07 – 2011/09
Insurance Claiming System	**Integrated Social Security & HIS**	2009/12 – 2010/02