Jian (Ken) Jian
Mobile: +1 (315) 802 9717
Email: ukgang.ukgang@gmail.com
Location: Cupertino, CA
LinkedIn: https://www.linkedin.com/in/jian-jian
Website: https://jianjian-portfolio.github.io

SUMMARY

Expertise in analyzing, building, and optimizing systems utilizing Big Data, Deep Learning, Computer Vision, NLP, and Regression & Forecasting models. Proven track record of guiding clients to invent AI algorithms that turn their ideas into applications. Successfully orchestrated machine learning and deep learning methods to software engineering to productionize AI technologies for various industries such as life science, intellectual property management, medical research, and social information science.

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 — 10/2021 – Present
- Developed deep learning models to detect confounding factors in images, improving model generalization ability
- Improved model performance by creating a benchmark matrix and fine-tuning deep learning model hyperparameters
- Implemented an evaluation pipeline to prevent non-deterministic results and improve product quality
- Collaborated with data scientists to accelerate inference speed of deep learning models using TensorRT/ONNX format
- Built a model deployment pipeline using Airflow and Beam, bringing research values to the user-end
- Created an infrastructure to consolidate machine learning models to edge computing units

AirDNA — Denver, CO
Data Scientist (Part-time / Remote) — 09/2021 – 09/2022
- Analyzed Airbnb data using SEM, PCA, and factor analysis to understand customer behavior in the short-term rental market
- Identified key factors that influence the level of informativeness, helpfulness, and persuasiveness in customer interactions
- Utilized BERT pre-trained model to track changes in customer preferences over time
- Developed strategies to improve the trustworthiness of business entities by analyzing firm-level data
- Tuned machine learning algorithms to accurately identify patterns of behavior among user groups in the sharing economy market

IPwe, Inc — Dover, DE
Data Scientist Intern — 02/2021 – 04/2021
- Designed a search engine with funnel-based architecture that enhanced search precision and met system service level agreements
- Constructed a BERT-based knowledge graph that resolved selection bias in the ranking stage

SOS+CD Laboratory — Syracuse, NY
Assistant Researcher — 01/2020 – 05/2021
- Analyzed network resource health using Survival Regression, Random Forest, Elastic Net, and Shapley value analysis
- Enhanced the Survival Regression model by incorporating an elastic net regularization term as a prior probability
- Re-designed the image labeling system using Django, MongoDB, Minio, and FabricJS
- Constructed a research database by combining structured and unstructured data gathered using Selenium

Active Network, LLC — Dallas, TX
Senior Java Software Engineer — 01/2016 – 06/2019
- Devised a Token Bucket Algorithm to regulate net traffic and improve its efficiency through a backpressure mechanism
- Improved the throughputs of the REST API for e-commerce microservice to 333+ TPS (Transactions Per Second)
- Implemented MongoDB cache for popular data to reduce physical I/O and improve performance
- Streamlined the ETL pipeline of the pricing system by integrating Spring State Machine that simplified the business process
- Integrated Apache Kafka & MongoDB Change Stream for reactive architecture

NCS Pte. Ltd. — Singapore
Senior Java Software Engineer — 03/2011 – 12/2015
- Led a team to upgrade two hospital information systems with suitable ejb2/struts1 architecture for extension development
- Launched invoice e-sign system project integrating backend (PDFBox2.0/iText4) with frontend (Base64/HTML5) techniques
- Benchmarked project workload for browser compatibility refactoring

NEC Corporation — Beijing, China
Associate Lead Java Software Engineer — 11/2007 – 02/2011
- Led a QA intern team, resolving 1.79 bugs/kilobyte for on-site product testing
- Designed monitoring dashboard for cluster nodes, upgraded hospital database, provided ERP on-call support

EDUCATION

Syracuse University, School of Information Studies Syracuse, NY
M.S., Applied Data Science 08/2019 – 05/2021

Chongqing Technology and Business University, School of Finance China
Bachelor., Economics 09/2004 – 07/2008

PUBLICATIONS

Predicting the longevity of resources shared in scientific publications (Under Reviewing by HSSCOMMS) 07/2020 – Present

This study, funded by US Office of Research Integrity (ORIIIR190049), looks at factors that impact scientific publication success, such as venue, authors, references, and sharing methods. Using web archives, we analyze a large collection of publications using Logistic Regression, Random Forest, Lasso, Ridge, Elastic Net Regression, and advanced Tobit survival analysis to identify key factors. Our findings show resource sharing as the most important factor and underscore the importance of standards and technology in promoting reproducibility and recognition of scientific datasets.

PROJECT EXPERIENCES

AI Platform for Single Cell Taxonomy 10/2021 – Present

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

This AI platform, which I helped develop as a senior machine learning engineer, enables researchers to analyze and classify cellular tissue composition by sorting and analyzing cell morphology. Working with data scientists, I fine-tuned deep neural models on the Google Cloud Platform using various network architectures to sort, clean, and classify cell images. To optimize computational resources, I utilized preemptible TPUs and developed a hyperparameter tuning tool for scheduling and monitoring training performance. Additionally, I implemented TensorRT and ONNX model formats for model compression and translation, and built a model management service to deliver the value to end-users. Our solution integrates a proprietary instrument with cloud computing resources, aiding biologists and cross-functional teams in researching unknown cell clusters.

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

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

As the leader of a 3-member team at IPWe, Inc, a startup company in the intellectual property industry, I was responsible for developing algorithms to expand query capabilities for candidate documents and integrating them into the platform's patent searching engine. Utilizing logistic regression and recall as evaluation metrics, we improved searching accuracy and met system SLA requirements. In addition, we integrated an ETL pipeline with a public database and BERT-based NER tagging to enhance candidate documents and reduce selection bias. Our work resulted in improved accessibility of intellectual acquisitions and enhanced user experience for online patent searching.

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

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

I led a research team to develop a solution that improves the generalization ability of AI models by experimenting with the generative adversarial network (GAN) architecture for image augmenting and embedding. I set a baseline performance for the GAN by building the generator and discriminator with a stack of CNN and batch normalization layers, and then improved the network by adding an edge detecting layer. I also utilized TensorFlow's checkpoint to improve the training process and coped with GPU resource limitations. We evaluated the model by conducting a double-blinded test with labelers and found that the difference in accuracy between the human and GAN's discriminator was not significant. Finally, we enriched the original dataset with synthetic images to alleviate the model's variance issue.

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

- *Java, Spring Boot, 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*

As a Senior Java Engineer on an enterprise internet infrastructure service project, I mentored junior team members, reviewed pull requests, developed, and benchmarked system performance. I integrated and facilitated the use of CPaaS platforms Twilio and Sparkpost within the Spring Boot framework, devised an SMS sending service in batch mode to improve performance, refactored the application API for better collaboration with other microservice clusters, and reorganized and simplified the metadata of the distributed infrastructure system. I also redefined the JMS message queue to decentralize system cache, scheduled tasks to periodically purge NFS space, and used JMap and JStack toolkit to benchmark system throughput. As a result of these efforts, the system's throughput was 312.73% higher than the legacy system, allowing the company to spare more computational resources for other business.