Stephen Bottos

778-966-8141 | bottos.steve@gmail.com | linkedin.com/in/stephen-bottos | github.com/stevebottos

Experience

Lead Machine Learning Engineer Aug. 2023 – Present
Kibeam, Inc. *Oakland, CA (Remote)*

- Built an LLM-powered content generation platform integrating OCR, RAG pipelines with embedding similarity search, and audio synthesis; enabled automated narration, transcription, and interactive game generation from book scans.
- Advanced LLM research, including agentic systems for iterative content development, multimodal LLM exploration, and prompt engineering strategies that reduced hallucinations and improved output reliability.
- Designed and implemented the company's core ML training system in PyTorch, including custom model training, int8 quantization (for ESP32 inference), evaluation pipelines, diagnostics, and alerting; this system powers all production models and experimentation across the organization.
- Architected company-wide ML infrastructure using Docker, Airflow (MWAA), AWS ECS/ECR, and MLflow, with Redshift integration for data warehousing and event tracking; enabled scalable iteration and deployment with 100+ new models/week and 1,000+ unique production models deployed to embedded devices.
- Developed lightweight CNNs optimized for low-latency inference on ESP32; models perform accurate recognition of hundreds of book-specific objects under strict memory constraints and without external connectivity.
- Contributed to Unity-based simulation pipelines generating 100% of CV training data; tuned augmentations and scene diversity to achieve production-grade generalization from synthetic data alone, eliminating manual data collection.
- Led cross-functional ML initiatives by defining project scopes, research direction, and development priorities across product, content, and engineering teams—ensuring alignment of modeling efforts with product vision and user needs.

Senior Machine Learning Engineer Aug. 2021 – Aug. 2023
Plainsight, Inc. *San Diego, CA*

- Primary developer of internal software which enabled the quick deduplication, clustering, and exploration of millions of images and videos, using Milvus, BigQuery, ClickHouse, and ANN algorithms, accelerating data exploration by up to 80%.
- Wrote Kubeflow and Apache Beam ETL pipelines handling large amounts of data from client data lakes into our data warehouses.
- Responsible for designing, implementing, deploying, and training custom model architectures in Python/Pytorch driving business logic for customer accounts worth up to $10M. While mostly Computer Vision models, many other modalities were leveraged on a per-use-case basis including video, text, and LiDAR pointcloud data.
- Built a vector database on top of open-source image data (scaling up to millions of records) to enable dataset supplementation through either text or image queries, allowing the internal team and customers to acquire large amounts of relevant data for their project in seconds rather than weeks.
- Built and scaled cloud workflows in GCP, using custom docker containers along with VertexAI to train and deploy models behind RestAPI endpoints for inference.
- Obtained the certificate: Google Cloud Certified Professional Machine Learning Engineer[1].

Machine Learning Engineer June 2020 – Aug. 2021
alwaysAI, Inc. *San Diego, CA*

- Designed, trained, and deployed custom AI Computer Vision models using Tensorflow and Pytorch.
- Responsible for back-end production and maintenance of the company's platform model retraining/inference software. Training of models and all ETL pipelines were handled using AWS ECS and custom docker containers, with models requiring quantization and pruning prior to being deployed offline on edge devices for real-time inference.
- Sole inventor of patented company system for object tracking and re-identification (link to Google Patents[2]). Consists of many deployed models across edge devices in a network which identify and track objects (usually people) across many cameras in a store in order to identify interactions in regions of interest to produce digital-like metrics in a physical environment. During employment, two contracts yielding upwards of $1M used this technology, and still do.

[1] https://www.credential.net/51be5dae-2e20-4700-b3da-478819ba76a9

[2] https://patents.google.com/patent/US20220335626A1/en?oq=+SYSTEMS+AND+METHODS+FOR+OBJECT+RE-IDENTIFICATION+17%2f332%2c522

Machine Learning Engineer Sep. 2019 – June 2020
Qimia, Inc. *San Diego, CA*

- Gained hands-on exposure to a broad range of ML applications, solely responsible for data engineering, data science, and machine learning demands of a large advertisement agency whose mission was to optimize campaigns regionally, with billions of data points relating to individuals across the United States. Also worked with other sub-contractors to develop early-stage spill detection and inventory analysis software commonly running on robots deployed to grocery/big-box stores.
- Typically utilized Apache Spark to facilitate a data science workflow including, data analytics/BI, data manipulation, feature engineering, and predictive modeling at scale, deploying applications using Docker and presenting results to stakeholders using custom dashboards.
- Often worked with large, complex datasets (billions of records) in AWS RedShift.
- Mainly used PySpark and Scala for working with big data.

Machine Learning Research Engineer Jan. 2018 – Aug. 2019
University of Windsor *Windsor, ON*

- First-authored two conference (*A novel slip-Kalman filter to track the progression of reading through eye-gaze measurements*, *Tracking the progression of reading using eye-gaze point measurements and hidden Markov models*) papers and one IEEE journal publication (*Tracking the progression of reading through eye-gaze measurements*). All of these can be viewed on my Researchgate profile[3].
- My thesis, *Statistical Methods to Measure Reading Progression Using Eye-Gaze Fixation Points*, can be viewed here[4].
- Primary area of research was, generally, developing and optimizing application-specific algorithms which utilize various machine learning/statistical modeling and predictive analysis techniques to extract information about an individual, given eye-gaze fixation data.
- Regularly developed models and techniques which include but are not limited to Principal Component Analysis, Neural Networks, Hidden Markov Models, Kalman Filters, K-Means Clustering, Gaussian Mixture Models, Support Vector Machines (SVM), and Decision Trees.
- Commonly used technologies included Matlab, Python, C, and C++. Regularly utilize many current Machine Learning/Data Science Python packages including Numpy, Pandas, Scipy, Scikit-Learn, Matplotlib, Seaborn, Pomegranate, Keras/TensorFlow, LightGBM , and OpenCV for computer vision tasks.

Education

University of Windsor Windsor, ON
M.Sc, Electrical and Computer Engineering *Jan. 2018 – Aug. 2019*

University of Windsor Windsor, ON
B.Eng, Mechanical Engineering *Sep. 2012 – Aug. 2016*

Technical Skills

General: Deep Learning, Data Science, Data Engineering, Computer Vision, NLP, Generative AI, MultiModal Models, Embedded ML, Model Quantization, Model Optimization, Edge AI
Languages: Python, C/C++, SQL, MATLAB
Developer Tools: Git, GitHub Workflows (CI/CD), Docker, Google Cloud Platform, Amazon Web Services (ECS, MWAA, ECR, Redshift), Microsoft Azure, Apache Airflow
Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, PyTorch, PySpark, Keras, TensorFlow, MLflow, TensorBoard, TFLite, OpenCV, ONNX
LLM Tooling: OCR Pipelines, Retrieval-Augmented Generation (RAG), Prompt Engineering, LangChain, Hugging Face Transformers

[3] https://www.researchgate.net/scientific-contributions/Stephen-Bottos-2153466785
[4] https://scholar.uwindsor.ca/etd/7776/