Bill Cai

Applied Scientist, Machine Learning

Work Experience

2022 | Present

Applied Scientist

Generative AI Innovation Center, Amazon Web Services — Singapore

- AWS Generative AI Innovation Center works with strategic customers across the globe to successfully build and deploy successful generative AI solutions. I lead science efforts in ASEAN, India and Korea, with previous coverage of Japan and Greater China Region customers.
- Optimized internal LLMs for Amazon Bedrock and customer engineering teams with optimal LLM quantization methods such as GPTQ on custom architectures to achieve up to reduction of 70% in memory footprint with similar throughput and perplexity as 16bit models. This enabled cost reduction for internal service teams and customer deployments of up to 50-80%.
- Tech lead for customer engagements on novel applications of Generative AI. Main tech lead for collaborations on socially impactful genAI projects, including for education and literacy education.
- Led a 2 month customer engagement with a globally recognisable asset manager to build a retrieval augmented generation pipeline to analyze and gather insights from SEC filings, earnings call transcripts. Led a 2 month engagement with a regional consumer bank to build a LLM agent that queries internal informational APIs and analyses consumer data.
- Main technical lead for a 3 month customer engagement with a large national entity to build a MLOps platform with Python/R notebooks, MLOPs pipelines including model packaging, model versioning with evaluation metrics, deployment to GPU/CPU endpoints.
- Leading research projects (2-3 scientists per team) on benchmarking LLM agent performance on machine learning problems, LLM performance for educational tasks (accepted in NAACL 2024)
- Contributor to open-source projects such as Alpaca Eval and AutoGPTQ

2019 | 2022

MLOps Lead

Data Science and AI Division, Government Technology Agency — Singapore

- Govtech Singapore is the technology arm of the Singapore government. The Video Analytics team works on developing and deploying computer vision and video understanding models for social good. Hired as a member of the founding technical team.
- Lead for AI modelling efforts in few-shot object detection, video activity recognition, and captioning models. Deployed and implemented using REST APIs in Python/Typescript, with K8s backends for infrastructure abstraction.
- Main architect and DevSecOps lead for petabyte-scale cloud-native computer vision platform and ML pipelines for image and video analytics. Scaled ML infra engineering squad from 1-3 people, supporting a larger data science and analytics team of 20, with petabyte-level of data-intensive products that save >1mil man-hours annually with cost savings of >$10m annually.
- Tech lead for crowd analytics project for 200+ cameras deployed on AWS. Designed and implemented full Terraform infrastructure-as-code, serverless architecture using AWS Lambda, API Gateway, and cloud-native solutions including AWS Rekognition, S3, ECR, API Gateway.
- Practical experience with deploying, updating, and maintaining a secure and compliance-ready cloud-native system. Designed a fully security-compliant system, and main representative for successfuly completed 3rd party security assessments for IM8 and security risk assessment.
- First and only DS/ML engineer emplaced onto Digital Technologist Scheme, which is a talent initiative with selection by external industry experts, and re-designed pay scheme and development pathways to retain and develop technical talent. Received COVID-19 Recognition Award from Minister for contributions to Singapore's pandemic response.

2018 | 2019

Data Scientist, Computer Vision and Deep Learning

One Concern — Menlo Park, California

- One Concern is a benevolent AI company that provides trusted insights that positively impact our communities. Our mission is to drive deep social impact through benevolent intelligence to save lives and livelihoods.
- Lead in-house inference of key features from unstructured data, such as satellite images and street-level iamgery. Extensive use of Keras, Tensorflow, PyTorch to build deep learning tools. Wrote and built Docker images, with deployment in Kubernetes.
- Backend engineering for a city-scale and real-time platform for infrastructure resilience. In charge of resilience and infrastructure recovery estimations, using combinatorial and graph optimisation techniques. Launched a new power and water grid estimation modeling effort that grew from a 2 person team into a 10+ person new product team, while leading algorithmic and ML engineering technical functionalities. In charge of key technical challenges including vectorising bottleneck computations to 100x in Python to increase computational output.
- Customization of open-source Javascript/HTML/CSS image annotation libraries with integration to Amazon Mechanical Turk.

Contact Info

Website and Projects:

billcai.com

LinkedIn Profile:

linkedin.com/in/billcai77

Email Address:

billcai@alum.mit.edu

Skills

Building large-scale and operational ML systems
Deep learning and ML frameworks: PyTorch, Tensorflow, sklearn
Python, Typescript, Julia, MATLAB, R, Stata, SQL
AWS Solutions Architect Pro, AWS DevOps Eng Pro
Docker, Kubernetes, Terraform, AWS CDK

Interests

Coding, Programming
ML Research (295+ citations, h-index of 7) Google Scholar
Economic Theory (PhD classes in Market/Auction Design, Computational Macro)

Conferences, Talks and Seminars

ICLR 2020, NeurIPS 2020, NeurIPS 2021, ICML 2021 Climate Change Workshop Program Committee
Reviwer for CVPR, NeurIPS, IEEE Internet of Things Journal
NCS Impact 2023 Invited Speaker
IPOS Intl 2022 Course for AI for Public Policy Speaker
Singapore Tech Forum 2019 Panelist

2017 | 2018

Graduate Researcher, Computer Vision

MIT Senseable City Lab — Cambridge, Massachusetts

- Trained computer vision models (segmentation, object detection models), and large-scale deployment to quantify urban canopy cover and parking utilization on large city-wide scales
- Sensor-fusion of lidar and camera data for obstacle detection in autonomous marine vehicle applications in Amsterdam and Boston/Cambridge
- Implemented state-of-the-art CNN architectures for classification, semantic segmentation and instance segmentation, including residual network, Mask-RCNN, PSPNet. Utilized gradient class activation (Grad-CAM) maps to understand learned features
- Extensive use of ROS, including Google Cartographer for SLAM, Velodyne lidar, IMU, USB cameras, for sponsored project by SNCF in Paris

2017 | 2017

Summer Associate, Product Analytics

Thumbtack — San Francisco, California

- Built live dashboards with Python, R, SQL, Javascript/HTML/CSS to track key metrics
- Modeled two-sided matching and dynamic marketplaces in Python. Our engineering blog post that explains more!
- Analyzed A/B test results, including using quasi-experimental methods, to understand impact of product feature changes on customer behavior
- Worked closely with product managers, engineers and designers to shape product decisions

Education

M.S. in Computational Science and Engineering

Center for Computational Science and Engineering, Massachusetts Institute of Technology
Cambridge, Massachusetts

- GPA: 5.00/5.00, thesis on applying computer vision and deep learning for large-scale quantification of urban and city dynamics (advised by Carlo Ratti)
- Selected Coursework: Advances in Computer Vision, Statistical Learning Theory and Applications, Numerical Methods in Partial Differential Equations, Optimization Methods

B.A. in Economics

University of Chicago — Chicago, Illinois

- GPA: 3.87/4.00, Graduated with Phi Beta Kappa (highest honors) and Dean's List fot all years

Research, Journal and Conference Publications

Mar 2024

Low-Cost Generation and Evaluation of Dictionary Example Sentences

Accepted and to appear in NAACL 2024
Bill Yang Cai, Ng Boon Liang Clarence, Daniel Tan Wee Liang, Shelvia Hotama

Dec 2020

DAMSL: Domain Agnostic Meta Score-based Learning

CVPR 2021 Workshop on Learning from Limited and Imperfect Data
John Cai, **Bill Yang Cai**, Shengmei Shen

Oct 2019

Quantifying Urban Canopy Cover with Deep Convolutional Neural Networks

Published in NeurIPS Workshop on Climate Change AI
Bill Yang Cai, Xiaojiang Li, Carlo Ratti

Dec 2018

Quantifying Legibility in Indoor Spaces Using Deep Convolutional Neural Networks: A Case Study in Train Stations

Published in Building and Environment
Wang Zhoutong, Liang Qianhui, **Bill Yang Cai**, Louis Charron, Fabio Duarte, Carlo Ratti

Dec 2018

Deep Learning Architect: Classification for Architectural Design through the Eye of Artificial Intelligence

Published in Computational Urban Planning and Management for Smart Cities
Yuji Yoshimura, **Bill Yang Cai**, Wang Zhoutong, Carlo Ratti

Aug 2018

Deep Learning Based Video System for Accurate and Real-Time Parking Measurement

Published in IEEE Internet of Things Journal
Special Issue on Enabling a Smart City: Internet of Things Meets AI
Bill Yang Cai, Ricardo Alvarez, Michelle Sit, Fabio Duarte, Carlo Ratti

Feb 2018

Using Street-level Images and Deep Learning for Urban Landscape Analysis

Published in Landscape Architecture Frontiers
Xiaojiang Li, **Bill Yang Cai**, Carlo Ratti