

David Chang

Seattle, WA 98109
(408) - 560 - 6268
changdavidb@gmail.com
davidbchang.github.io
github.com/davidbchang
linkedin.com/in/changdavidb

Education

University of Washington, Seattle

B.S. in Computer Science, Minor in Mathematics, GPA: 3.8/4.0

Sept 2018 – Jun 2022

Experience

Docugami, Remote - Seattle, WA

Machine Learning Engineer

Sept 2022 – Present

- Owned end-to-end development of customer-facing multimodal models; improved F1 by +10 points (absolute) on production test set via error analysis and targeted improvements across data/label quality, model architecture/training, and hyperparameter optimization.
- Implemented adaptive request-level batching for variable-length documents (1 to 3-5 docs/request) with model micro-batching, increasing document throughput 3.5x and raising average GPU utilization from 40% to ~90%.
- Redeployed models as TensorRT engines on NVIDIA Triton, boosting inference throughput 4x on lower-cost GPUs.
- Redesigned the distributed ingestion/inference pipeline from PySpark to Redis-backed queues and workers, cutting data ingestion latency 10x and improving ML microservice scalability under peak load.
- Cut model iteration cycle time 50% by building an MLOps pipeline (MLflow, DVC, GCP) to automate data processing, training, evaluation, experiment tracking, and reproducibility.
- Led the data collection/quality effort and managed an offshore annotation team with weekly check-ins and continuous QA, saving 10+ hrs/week of manual labeling.

Docugami, Remote - Seattle, WA

Machine Learning Engineer Intern

Jun 2022 – Sept 2022

- Built an extractive QA transformer model for business document contextual semantic understanding, improving extraction quality on internal evaluation sets and customer-facing documents.
- Designed and annotated a novel SQuAD-style dataset for contextual language understanding.
- Fine-tuned a distilled RoBERTa model on this dataset and achieved a higher F1 score of +8 points (absolute) and 3x faster inference speed on production test data.
- Performed large-scale data processing through distributed computing for model inference using Apache PySpark.

UW NLP, xlab, Seattle, WA

Research Assistant

Mar 2021– Jun 2022

- Created a vision and language task and dataset for social commonsense reasoning on movie scenes and captions.
- Fine-tuned OpenAI CLIP with an image-masking technique; improved RefCOCO testA accuracy by +1 point and RefCOCO+ testA by +2 points vs UNITER's strong baseline.

Paul G. Allen School of Computer Science & Engineering, Seattle, WA

Data Programming Teaching Assistant

Jan 2021– Mar 2022

- Taught Python and data programming; led a weekly section of ~30 students and mentored 1:1 during office hours.

Projects

Improving VLM Spatial Reasoning via Attention Interventions

Oct 2025– Feb 2026

github.com/davidbchang/visual-attention-intervention

- Implemented attention intervention techniques from 3 research papers that mitigate object hallucinations and attention sinks in VLMs and investigated their impact on spatial reasoning.
- Fine-tuned and evaluated Qwen3-VL model on the Visual Spatial Reasoning dataset using PyTorch, boosting accuracy by +0.5 points on random split and +1.4 points on zeroshot split using attention intervention.
- Generated interpretable visualizations such as visual attention heatmaps and attention weight distribution plots to guide hyperparameter tuning and experimentation.

Skills

Languages: Python, SQL, C#

ML/Data: PyTorch, Hugging Face Transformers, NumPy, Pandas, Redis

MLOps/Infra: MLflow, DVC, TensorRT, NVIDIA Triton, Docker, Kubernetes, Git, GCP, AWS S3, Azure