Marcus Neo Jing Quan

(+65) 98538012 | marcus.neo418@gmail.com | linkedin.com/in/marcusneo | github.com/marcus-neo

Founding ML Engineer with 4+ years at an AI startup, owning the full machine learning lifecycle — from training pipeline architecture and model export to inference serving and data annotation tooling. Built and scaled production systems across classical computer vision, 3D medical imaging, and vision-language model fine-tuning on GCP.

EDUCATION

Imperial College London Oct 2018 – Jun 2022
MEng Electronic & Information Engineering, First Class Honours

- Final Year Project: *Data Stream Evolution using Multiresolution Wavelet Transform* — Python-based ML research applying wavelet transforms as preprocessing for anomaly detection in data streams

Hwa Chong Institution, Singapore Jan 2014 – Dec 2015
GCE A-Levels — AAAA

EXPERIENCE

Datature Jul 2022 – Present
Founding Machine Learning Engineer

Model Training Service — Sole author, deployed on Google Compute Engine

- Enabled customers to train on any major framework without migration overhead by building a framework-agnostic pipeline supporting JAX, TensorFlow, and PyTorch with multi-GPU support (Data/Model Parallelism)
- Onboarded 16+ model architectures (YOLOv8/11/26, MaskRCNN, EfficientDet, DeepLabV3, PaliGemma, DFINE, RF-DETR, EfficientAD), giving users broad coverage across classification, detection, segmentation, pose estimation, anomaly detection, and 3D medical imaging
- Improved model performance on high-resolution imagery via sliding window (SAHI) training with NMS and configurable overlap ratios
- Reduced manual data preparation through class auto-balancing, multilabel classification (Brier score, co-occurrence confusion matrices), and custom training metadata integration
- Enabled faster iteration via continual learning, allowing users to resume from prior checkpoints without retraining from scratch
- Lowered the barrier for non-ML users by building an OpenAI-powered hyperparameter recommender for automated training configuration

3D Medical Imaging Pipeline — Built end-to-end using MONAI

- Unlocked a new product vertical in medical imaging by developing end-to-end pipelines for volumetric models (SwinUNETR, SegResNet, nnUNet, SegMamba) with NIfTI loading, 3D augmentations, and affine-aware evaluation
- Enabled seamless data exchange with existing medical workflows via NIfTI annotation import/export with class-indexed and one-hot mask exports, 3D TFRecord export, and aggregation statistics
- Enabled clinical radiotherapy workflow compatibility by adding DICOM input and RTSTRUCT output support
- Defined 3D AEMD schemas and bitmask chunk encoding in the shared core type library used across all services

Vision-Language Model Fine-Tuning Pipeline — Built from scratch, deployed on GCP

- Opened up the VLM product line by establishing a separate fine-tuning service with full CI/CD, supporting QWEN2.5-VL, QWEN3-VL, NVILA, CosmosReason1/2, and KimiVL
- Reduced GPU memory requirements via LoRA fine-tuning with quantization; enabled multi-GPU scaling through tensor parallelism with OOM-resilient training loops
- Supported diverse customer use cases — VQA, Chain-of-Thought reasoning with structured evaluation, freeform generation, and video training via PyAV ingestion
- Automated annotation labelling at scale by building the Intelliscribe caption generation microservice with JSON schema validation for structured annotation outputs

Model Export Service — Authored, deployed on Google Cloud Run

- Enabled deployment to edge devices, mobile, and real-time applications via cross-framework export (TFLite, CoreML, TensorRT, OpenVINO, ONNX) with Float16/Int8 quantization (up to 75% model size reduction) and model pruning up to 90% of parameters
- Ensured service reliability under heavy load by wrapping export conversion in multiprocessing with async apply and timeout to prevent Cloud Run request hangs

Model Hosting & Inference Service — Co-authored, deployed on Google Kubernetes Engine

- Powered production inference for all onboarded model architectures across image, video, and 3D volumetric inputs (NIfTI and DICOM) by integrating Triton Inference Server with Numba JIT optimisation
- Reduced annotation costs for customers by building an active learning pipeline with entropy-based metrics and bitmask annotation re-upload to the platform

Annotation Data Services

- Eliminated data migration friction via annotation import/export for 10+ formats (COCO, LabelMe, SuperAnnotate, ScaleAI, OpenAI, VoTT, Supervisely, IBM) with multipolygon, sliding window, and 3D NIfTI support
- Built VQA/Visual Intelligence annotation pipeline: JSONL and ViFull import/export, phrase grounding with caption/bbox support, and Neural Insights
- Improved training data quality assurance by implementing Mosaic and CutMix augmentation preview for training data visualisation

Video Annotation Services — Deployed on GKE with KEDA scale-from-zero

- Enabled frame-by-frame video annotation at scale by building video interpolation and SAM2-based tracking services with PubSub queue and multiprocessing to prevent GPU memory leaks
- Ensured launch readiness by scaling and hardening deployments: KEDA authentication, production configs, and memory-optimised caching

CI/CD & DevOps

- Accelerated release cycles by designing GitHub Actions CI/CD pipelines across all ML services — automated linting, testing, Docker builds, and deployment to GCP (Compute Engine, Cloud Run, GKE)
- Reduced deployment image sizes through multi-stage Docker builds, while managing CUDA runtime upgrades and Python version migrations across training, export, and inference services
- Maintained PyPI and npm packages for the shared Alchemy training library and nexus-core-schema type library, ensuring version consistency across services

Datature Apr 2021 – Jan 2022
Machine Learning Intern

- Built custom on-the-fly augmentations for the TensorFlow training pipeline and microservices for augmentation preview, dataset statistics aggregation, and hyperparameter recommendations
- Developed the full backend for Portal (neural net visualisation webapp): REST API, multi-framework model compatibility (TensorFlow, PyTorch, DarkNet), video/image predictions, and bulk inference
- Developed hosted ML model deployment service on GCP using Docker

CORE COMPETENCIES

- **Programming:** Python, C++
- **ML Frameworks:** PyTorch, TensorFlow, JAX, ONNX, MONAI, Ultralytics
- **ML Serving & Export:** Triton Inference Server, TensorRT, CoreML, TFLite, OpenVINO
- **VLM Fine-Tuning:** LoRA, Quantization, Tensor Parallelism, DeepSpeed
- **Cloud & Infrastructure:** GCP (Compute Engine, Cloud Run, GKE, PubSub), Docker, Kubernetes, KEDA
- **Languages:** English (native), Mandarin (native)

ACTIVITIES & INTERESTS

- **Scuba Diving** — PADI Advanced Open Water + Nitrox certified, 50+ logged dives
- **IC Hack 2020** — Semi-finalist (Smart Recipe Cookbook)
- **Imperial College Singapore Society Musical 2020** — Composed and conducted the orchestra