DICKSON NEOH

AI Engineer with 10+ years in computer vision and time series modeling. Expert in scalable AI deployment, model optimization, and multimodal retrieval. Named a Top 2% Scientist by Stanford (2023). Skilled in developer relations, open source, and technical content.

dickson.neoh@gmail.com dicksonneoh.com

WORK EXPERIENCE

Visual Layer Inc.

Machine Learning Engineer & Lead Developer Relations — Mar 2023 - Feb 2025

- Optimized object detection, OCR, and vision-language models, achieving a 8x inference speedup using ONNX, TensorRT, and OpenVINO.
- Architected a hybrid multimodal image retrieval system enabling advanced content-based searches.
- Reduced customer annotation time by 90% through a human-in-the-loop active learning framework.
- Automated dataset quality improvement via a data flywheel framework to detect missing labels/errors.
- Scaled open-source project fastdup, increasing GitHub stars from 731 to 1.7k in 1.5 years

ZenML GmbH

Developer Advocate — July 2022-December 2022

- Designed social media strategies generating over 400k impressions in six weeks without paid ads.
- Represented ZenML through blogs, podcasts, and public talks.

KEY PROJECTS

Supercharge Your PyTorch Image Models: Bag of Tricks to 8x Faster Inference 2024

- Boost TIMM PyTorch model inference by converting to ONNX, optimizing with ONNX Runtime & TensorRT, and embedding preprocessing—achieve up to 123x speedup over CPU-based PyTorch.

Bringing High-Quality Image Models to Mobile 2023

- Deploy image classification models on mobile with Hugging Face Spaces and Flutter, using cloud-based inference to bypass hardware limits. It includes setup, code, and cross-platform demos.

Supercharging YOLOv5: How I Got 182.4 FPS Inference Without a GPU 2022

- Run YOLOv5 at 180+ FPS on consumer CPUs using Neural Magic's SparseML & DeepSparse, leveraging pruning, quantization, and sparsification.

EDUCATION

National Energy University, Malaysia

Ph.D. in Electrical & Electronics Engineering - 4.00 GPA — March 2018-June 2022

- Developed self-supervised Transformer models for battery state-of-charge estimation (Published in Nature).

SKILLS

- Programming Languages: Python, C, Dart
- Frameworks: TensorFlow, PyTorch, Keras
- Optimization Tools: ONNX, TensorRT, OpenVINO
- UI Development: Flutter, Streamlit, Gradio
- Database: PostgreSQL, MySQL.