COOPER PELLATON

+1(203) 291-9803 ◇ San Francisco, CA (open to remote)
c@cepp.ch ◇ linkedin.com/in/cooperpellaton ◇ github.com/cooperpellaton

OVERVIEW

Senior Software Engineer with 5 years of experience specializing in ML Systems and close-to-the-metal device work. Proven track record in optimizing Android systems, building AI platforms, and engineering NLP systems. Seeking an L6 role as Software Engineer, Systems ML - Frameworks / Compilers / Kernels at a Big Tech company.

EXPERIENCE

Senior Software Engineer Spring 2023 - Present
Humane *San Francisco, CA*

- Led effort to use idiomatic Android concepts. Personally deleted 1.3 million lines of code resulting in 50% reduction in OTA size and 4x improvement in battery life.
- Led design of an Android Voice Interaction Service, utilizing custom quantization aware training STT and TTS models, achieving sub-250ms inference time and .1 Watt power usage, enhancing user experience and runtime speed by 8x.
- Revamped large-scale multi-project Gradle build system by employing advanced caching and parallel execution techniques which sped up CI builds 4x and local builds 10x.
- Mentored junior engineer on implementation of neural voice TTS project, leading to her promotion to Senior in September 2023.

Software Engineer Spring 2022 - Spring 2023
Humane *San Francisco, CA*

- Independently converted 60M parameter model to TensorFlow Lite, implemented Transformer encoder/decoder logic in C++, reducing inference time to 50ms and enabling on-device intent determination.
- Implemented on-device inference library with Hexagon Delegate hardware support, overcoming device and Android version constraints, adopted by 3 teams enhancing their workflow efficiency.
- Established an AI Platform team from ground up, leading recruitment and on-boarding of 5 engineers, mentoring in ML systems development, resulting in successful delivery of key projects including LLM-based AI assistant functionality spanning on-device and server.

Software Engineer Fall 2019 - Spring 2022
Memora Health *San Francisco, CA*

- Developed high-performance NLP system handling 50K+ medical facts, serving whole customer base, improving answer quality by 20%.
- Architected SMS serialization/deserialization system using Node.js and Redis, reducing operational costs by 60% and ensuring compliance with telecom protocols for 10K+ daily messages.
- Designed and implemented FHIR data adapter, reducing user on-boarding time by 100% and enhancing data export functionalities, leading to 4x shorter customer on-ramping and marketplace distribution.

PROJECTS

Real-time Discord Bot for Meme Creation (2019 – 2022): Developed a Python-based bot using Discord.py that processes user requests in real time using Websockets and DynamoDB. Generated 1K+ memes per day.

fMRI Toolbox (2016-2019): Developed a suite of tools in C and Python for 4DFP and time-series realignment, ROI discovery and statistical analysis of BOLD responses. Improved data processing efficiency by 200%.

ML-powered YouTube Content Extractor (2016): Developed a Python-based client using TensorFlow to extract key contents from YouTube videos using ML and classical CV techniques. Won 2nd place at YHACK 16 and the A+E Challenge. Processed videos with 72% accuracy.

EDUCATION

Pursued Bachelor of Computer Science, Georgia Institute of Technology 2016 - 2019
Relevant Coursework: Comp. Architecture; Advanced Comp. Micro Architecture; Processor Design; Operating Systems; Algorithms; Compilers; Programming Languages; Big Data & Society

ADDITIONAL EXPERIENCE AND AWARDS

- Additional Experience: Interned at Alibaba (Summer 2018), implemented novel graph embedding research improving training time by 48x. Video++ (Summer 2017), contributed to the development of a CV based video-advertising platform with 600M monthly users. Cigna (Summer 2016) implemented a model to assist with incident debugging in CI/CD decreasing response time by 10%.
- Contributions: Contributed to open source projects such as PsychoPy, where I worked on support for Python 3 modules, and Intel Caffe where I fixed a bug that improved cache coherence 2x.
- Awards: Recipient of 2 US Patents for developing a novel machine learning inference scheme dubbed the "Intent Pyramid" and a system to extract data from an electronic medical record using OCR on PDFs.

LANAGUAGES AND TECHNOLOGIES

Programming Languages	Java, C++, Python, Rust
Machine Learning Frameworks	Tensorflow, PyTorch, CUDA
Build Systems	CMake, Bazel, Soong
Package Management	Poetry, Nix, Maven