Staff Machine Learning Performance Engineer

Description

Minimum qualifications:

- Bachelor’s degree or equivalent practical experience.
- 8 years of experience in software development, and with data structures/algorithms.
- 5 years of experience testing, and launching software products, and 3 years of experience with software design and architecture.
- 5 years of experience with machine learning algorithms and tools (e.g. TensorFlow), artificial intelligence, deep learning, or natural language processing.

Preferred qualifications:

- Experience in a technical leadership role leading project teams and setting technical direction.
- Experience working in a complex, matrixed organization involving cross-functional, and/or cross-business projects.
- Experience in performance analysis and optimization, including system architecture, performance modeling, or other similar experience.
- Experience in compiler optimizations or related fields.
- Distributed development and large-scale data processing experience.

About the job

Google’s software engineers develop the next-generation technologies that change how billions of users connect, explore, and interact with information and one another. Our products need to handle information at massive scale, and extend well beyond web search. We’re looking for engineers who bring fresh ideas from all areas, including information retrieval, distributed computing, large-scale system design, networking and data storage, security, artificial intelligence, natural language processing, UI design and mobile; the list goes on and is growing every day. As a software engineer, you will work on a specific project critical to Google’s needs with opportunities to switch teams and projects as you and our fast-paced business grow and evolve. We need our engineers to be versatile, display leadership qualities and be enthusiastic to take on new problems across the full-stack as we continue to push technology forward.

The TPU Performance team is responsible for bleeding edge performance and extracting maximum efficiency for machine learning/AI training workloads. We drive Google ML performance using deep fleet-scale, benchmark analysis, and out-of-the-box auto-optimizations.

We focus on performance analysis to identify performance opportunities in Google production, research ML workloads, and land optimizations to the entire fleet. Our work demonstrates cutting edge ML performance on the largest scale and latest accelerators at MLPerf competition. We push state-of-the-art efficiency on multipod ML models.

Hiring organization
Candidate-1st

Employment Type
Full-time

Beginning of employment
asap

Job Location
London, UK

Working Hours
40

Base Salary
euro USD 60K - 112K *

Date posted
May 14, 2024

Google Cloud accelerates every organization’s ability to digitally transform its business and industry. We deliver enterprise-grade solutions that leverage Google’s cutting-edge technology, and tools that help developers build more sustainably. Customers in more than 200 countries and territories turn to Google Cloud as their trusted partner to enable growth and solve their most critical business problems.

Responsibilities

- Focus on large language models (Google Deepmind Gemini, Bard, Search Magi, Cloud LLM APIs, etc.), performance analysis, and optimizations.
- Identify and maintain LLM training and serving benchmarks that are representative to Google production, industry and ML community, use them to identify performance opportunities and drive TensorFlow/JAX TPU out-of-the-box performance toward state-of-the-art, and to gate TF/JAX releases.
- Engage with Google product teams to solve their LLM performance problems for example, onboarding new LLM models and products on Google new TPU hardware, enabling LLMs to train efficiently on very large-scale (i.e. thousands of TPUs), etc.
- Explore model/data efficiency techniques for example, new ML model arch/optimizer/training technique to solve a ML task more efficiently, new techniques to reduce the label/unlabeled ML data needed to train a model to target accuracy.

How the process will look like

Your teammates will gather all requirements within our organization. Then, once priority has been discussed, you will decide as a team on the best solutions and architecture to meet these needs. In continuous increments and continuous communication between the team and stakeholders, you’re part of making data play an even more important (and understood) part withing Brand New Day.

Job Benefits

USD 60K – 112K *