Chen Gaosong(Cedar)

+8613226435125 a1249812431@gmail.com
https://www.linkedin.com/in/cedarchen https://tallsong.github.io
Software Development Engineer

SUMMARY

Software engineer with 5 years of experience specializing in optimizing Large Language Model (LLM) performance and developing robust software solutions (Python, C++). Proven ability in designing and implementing LLM inference API services, leveraging tools like vLLM and low-bit quantization to enhance throughput and reduce memory usage. Proficient in GPU architecture, CUDA programming, PyTorch, and experienced with distributed systems and big data platforms.

SKILLS LIST

- Programming Languages: Python, C, C++, JavaScript, Shell
- Frameworks/Tools: PySpark, PyTorch, CUDA, Django, Qt, Flask, Celery, MySQL, Redis, Git, Docker, Kubernetes, Linux, Scrapy, CEPH, RESTful API
- Testing Platform: Ansible, Pytest

PROFESSIONAL EXPERIENCE

Runjian 2023.02 - 2025.04
Software enginner Beijing, China

- Led the development of EBConverter, a drawing conversion tool built with C++, ensuring seamless XML data conversion while maintaining system scalability. Improved data accuracy during conversion by implementing strict testing protocols
- Led the development of word processing software that automatically identifies Word documents, extracts content, and renders it in a GUI for user interaction, increasing UI response efficiency by 95% (Qt, C++)
- Developed the server-side user management module for nuclear power monitoring (FastAPI, Python)
- Designed and independently developed a Python (Flask)-based LLM inference API service aimed at optimizing large model inference performance
  - Tested the debugging capabilities of major open-source and closed-source LLMs on the market using Hugging Face benchmarks, and selected the one with the best performance
  - Integrated and applied the industry-leading VLLM inference engine, replacing the standard Hugging Face implementation, significantly enhancing service throughput and response speed
  - Implemented and evaluated 4-bit low-bit quantization technology (using bitsandbytes), effectively reducing model memory usage while maintaining acceptable inference accuracy
  - Built asynchronous non-blocking API interfaces using Flask to improve service concurrency handling. Used Docker for service containerization, simplifying deployment and ensuring environmental consistency
  - Technology Stack: Python, Flask, vLLM, PyTorch, CUDA, Docker, Git

JD.com 2021.10 - 2022.08
Software development engineer - Data Intellgence Beijing, China

Participated in the development of Nine Security Computing Platform, mainly responsible:

- SQL Parser: Generate AST by parsing the SQL input by the user through Lexer and parser, and then perform code generation to generate the corresponding secure computing API code (ANTLR, Pyspark, SQL, Python), and the parsing ability is improved by 50%
- Jupyter development: developed custom user login, code detection, custom kernel, and other functions (Jupyter, Python, Kubernetes, Docker)
- File service: developed user data storage, query, and other related functions (Flask, Pandas), supports access to file list of folders from the file system, HDFS, and S3 interfaces, creating folders, obtaining CSV and ORC format headers, obtaining the number of rows and columns, file size, uploading files, and reading the first N lines of files. The results are returned in restful style JSON, an average increase of 30% in final response speed

Institute of Information Engineering, Chinese Academy of Sciences 2020.07 - 2021.10
Software development engineer Beijing, China

- Optimized the user interface and implemented the backend functions (C++, QT, socket) of the security and security authorization management system of the virtualization platform, reducing the time to manage access by 40% through automation
- The design and implement a distributed storage system, mainly implementing user management, uploading and downloading files, file collection and sharing, user history viewing, and other functions (CEPH, Linux, Django, Boto3, HTML, CSS, JavaScript, python)
- Storage cluster monitoring module development and document online browsing function implementation (Flask, JavaScript, Python)

Yuexin Technology 2019.06 - 2019.12
Data development engineer Intern - Big Data Engineering Beijing, China

- Data warehouse extraction, scheduling, processing, cleaning, and analysis of hundreds of millions of data levels
- Use ETL tools to write scripts and SQL to process data (Kettle, JavaScript)
- Develop data crawlers for enterprise industrial and commercial data (Python, Scrapy, XPath)

EDUCATION

Guangdong Baiyun University 2016.09 - 2020.07
Information Management and Information System Bachelor Guangzhou, China