GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings

Bang Fu, Di Feng
Zilliz Inc.

Abstract

The rise of ChatGPT [1] has led to the development of artificial intelligence (AI) applications, particularly those that rely on large language models (LLMs). However, recalling LLM APIs can be expensive, and the response speed may slow down during LLMs' peak times, causing frustration among developers. Potential solutions to this problem include using better LLM models or investing in more computing resources. However, these options may increase product development costs and decrease development speed. GPTCache [2] is an open-source semantic cache that stores LLM responses to address this issue. When integrating an AI application with GPTCache, user queries are first sent to GPTCache for a response before being sent to LLMs like ChatGPT. If GPTCache has the answer to a question, it quickly returns the answer to the user without having to query the LLM. This approach saves costs on API recalls and makes response times much faster. For instance, integrating GPTCache with the GPT service offered by OpenAI can increase response speed 2-10 times when the cache is hit. Moreover, network fluctuations will not affect GPTCache's response time, making it highly stable. This paper presents GPTCache and its architecture, how it functions and performs, and the use cases for which it is most advantageous.

1 Introduction

Since OpenAI released ChatGPT, large language models have impressed many people and have been frequently integrated into our daily work and lives. At the same time, more open-source enthusiasts and tech companies have invested time and effort into developing open-source LLMs, such as Meta's LLama (Touvron et al., 2023a,b), Google's PaLM (Chowdhery et al., 2022), Stanford's Alpaca (Wang et al., 2023; Taori et al., 2023), and Databrick's Dolly (Conover et al., 2023).

There are two ways to use large language models: online services provided by companies like OpenAI, Claude, and Cohere or downloading open-source models and deploying them on your servers. Both methods require payment. Online services charge you based on tokens, while deploying models on your own server requires purchasing specific computing resources. The choice depends on individual needs.

While online services are more expensive, they are more convenient and effective and provide a better user experience than deploying models yourself. Costs and user experience are two critical considerations for building LLM applications. As your LLM application gains popularity and experiences a surge in traffic, the cost of LLM API calls will increase significantly. High response latency will also be frustrating, particularly during peak times for LLMs, directly affecting the user experience.

GPTCache is an open-source semantic cache designed to improve the efficiency and speed of GPT-based applications by storing and retrieving the responses generated by language models. Unlike traditional cache systems such as Redis, GPTCache employs semantic caching, which stores and retrieves data through embeddings. It utilizes embedding algorithms to transform the queries and LLMs' responses into embeddings and conducts similarity searches on these embeddings using a vector store such as Milvus. GPTCache allows users to customize the cache to their specific requirements, offering a range of choices for embedding, similarity assessment, storage location, and eviction policies. Furthermore, GPTCache supports both the OpenAI ChatGPT interface and the Langchain interface, with plans to support more interfaces in the coming months.

Through experiments using the `paraphrase-albert-small-v2` model (Reimers

[1] `https://openai.com/chatgpt`
[2] `https://github.com/zilliztech/GPTCache`

Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), pages 212–218
December 6, 2023 ©2023 Association for Computational Linguistics