УДК 004.65:004.8:004.738.5

П.В. Травничева
Витебский государственный университет имени П.М. Машерова, Витебск, Беларусь

ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ ВЫСОКОПРОИЗВОДИТЕЛЬНОЙ СИСТЕМЫ ХРАНЕНИЯ И ИНДЕКСИРОВАНИЯ НЕЙРОДАННЫХ С СЕМАНТИЧЕСКИМ ПОИСКОМ

Ключевые слова: нейроданные, векторная база данных, семантический поиск, HNSW, ChromaDB, FastAPI, Kubernetes.

Введение. В современную эпоху больших данных и искусственного интеллекта нейросетевые модели становятся неотъемлемой частью множества сфер деятельности. С увеличением сложности и масштабов моделей экспоненциально растут объемы нейроданных — весов, эмбеддингов и метаданных. Это создает потребность в эффективных методах их хранения, индексации и семантического поиска, что особенно актуально для задач дообучения моделей, управления версиями и MLOps. Современные системы хранения данных, такие как реляционные и NoSQL базы данных [1], не всегда способны эффективно работать с высокоразмерными векторными представлениями и выполнять семантический поиск по сходству. В связи с этим возникает необходимость в разработке специализированного программного обеспечения, способного обеспечить высокопроизводительное хранение и интеллектуальный поиск по нейроданным.

Основной текст. Цель работы заключалась в проектировании и реализации высокопроизводительной системы хранения и индексации нейроданных, обеспечивающей масштабируемость, низкую задержку при операциях чтения/записи, эффективный семантический поиск по векторным представлениям и универсальность применения в различных конвейерах машинного обучения.

Методика проектирования включала анализ существующих технологий и практическую реализацию системы. В результате изучения были выбраны следующие технологии:

- ChromaDB / Weaviate: эти векторные базы данных были выбраны за их открытость, производительность и богатый функционал для семантического поиска. Они предоставляют эффективные методы индексации, такие как HNSW, что критически важно для работы с высокоразмерными данными [2].

- FastAPI: для создания высокопроизводительного API был выбран фреймворк FastAPI. Этот фреймворк обеспечивает простоту разработки, автоматическую генерацию документации и асинхронную обработку запросов, что снижает задержки при обслуживании клиентов.
- Docker & Kubernetes: технологии контейнеризации и оркестрации были выбраны за свою способность обеспечить масштабируемость, отказоустойчивость и простоту развертывания системы в различных средах [3].

Интерфейс приложения для управления нейроданными разработан с акцентом на удобство и интуитивность использования. Реализованы функции загрузки моделей (весов и эмбеддингов), семантического поиска по векторным представлениям и интерактивная панель мониторинга для отображения метрик производительности системы, статистики использования и результатов поисковых запросов.

Полученные результаты показывают, что использование современных векторных баз данных и микросервисной архитектуры позволяет создать масштабируемое и производительное решение для хранения и интеллектуального поиска по нейроданным, что является важным шагом в развитии инфраструктуры для машинного обучения и MLOps. Архитектура системы представлена на рисунке 1:

Заключение. Разработана высокопроизводительная и масштабируемая система для хранения и семантического поиска по нейроданным. Реализованная архитектура демонстрирует возможность создания эффективных решений для управления большими объемами векторных данных в конвейерах машинного обучения. Перспективным направлением дальнейших исследований является оптимизация алгоритмов индексации и повышение эффективности распределенного поиска.

Литература

1. *Костюков И.В.* Современные системы управления базами данных. М.: Наука, 2020.
2. *Шафеев А.П., Соловьев Д.В.* Векторная семантика и поиск в больших данных. СПб.: БХВ-Петербург, 2022.
3. *Лебедев С.Н.* Контейнеризация и оркестрация приложений с использованием Docker и Kubernetes. М.: ДМК Пресс, 2021.

Сведения об авторе:

Травничева Полина Владимировна – старший преподаватель кафедры прикладного и системного программирования Витебского государственного университета имени П.М. Машерова, магистр, e-mail: travnichevapolina@gmail.com