ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ " ОРГАНИЗАЦИЯ-2"

УДК 004.896

СОГЛАСОВАНО

Генеральный директор ООО "Организация-1"

Руководитель-1

«___»_____2024 г

УТВЕРЖДАЮ

Генеральный директор ООО " Организация-2"

Руководитель-2

«___»_____2024 г

Подп. и дата | Инв. № дубл. | Взам. инв. № | Подп. и дата | Инв. № подл.

МОДИФИКАЙИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НЕЙРОСЕТЕВОЙ БИБЛИОТЕКИ

PuzzleLib

Техническое задание

(П Р О Е К Т Д О К У М Е Н Т А)

02702700.425000.235-01 90 01

Листов 16

Техническое задание на доработку нейросетевой библиотеки PuzzleLib. В документе описаны задачи по добавлению новых функций, улучшению производительности и поддержке распределённых вычислений. Подходит для программистов, архитекторов и менеджеров, участвующих в проекте.
Подробнее: https://puzzlelib.org/ru/

г. Москва 2024 год

Литера

АННОТАЦИЯ

Настоящий документ является техническим заданием (ТЗ) на внесение модификаций в программное обеспечение (ПО) нейросетевой библиотеки PuzzleLib. ТЗ предназначено для специалистов, занимающихся проектированием, разработкой, тестированием и эксплуатацией ПО, а также для менеджеров проектов и инженеров, ответственных за выполнение и внедрение указанных модификаций.

Документ оформлен в соответствии с требованиями ГОСТ 19.201-78 и включает в себя основные требования к процессу модификации библиотеки PuzzleLib, этапам выполнения проекта, процедурам контроля и приемки, а также перечень необходимых технических и программных характеристик. Настоящий документ направлен на унификацию процесса разработки и внесения изменений в ПО, а также на обеспечение качества на всех этапах выполнения проекта.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

1.1 Наименование праграммы или программного изделия

Модифицированная версия нейросетевой библиотеки PuzzleLib.

1.2 Область применения

Модифицированная библиотека предназначена для разработки и оптимизации нейросетевых моделей в высокопроизводительных вычислительных системах.

2 ОСНОВАНИЯ ДЛЯ РАЗРАБОТКИ

2.1 Документы

Договор между заказчиком и исполнителем, а также требования, выработанные на этапе обсуждения проекта.

2.2 Организация утвердившая документ

- Организация, утвердившая документ: ООО "Нейросети Ашманова".
- Дата утверждения: 05 марта 2024 года.

2.3 Наименование темы разработки

- Тема разработки: «Модификация PuzzleLib».
- Условное обозначение: «PZL-M»

2.4 Сроки разработки

- Дата начала разработки: 05 апреля 2024 г.
- Дата завершения разработки: 31 августа 2024 г.

3 НАЗНАЧЕНИЕ РАЗРАБОТКИ

3.1 Функциональное назначение

Обеспечение дополнительной функциональности и производительности в существующей библиотеке PuzzleLib для ускорения обучения нейронных сетей.

3.2 Эксплуатационное назначение:

Использование в системах с высокой производительностью, поддержка работы на кластерах и графических процессорах.

4 ТРЕБОВАНИЯ К ПРОГРАММЕ

4.1 Требования к функциональным характеристикам

4.1.1 Оптимизация операций GEMM

В программе должна быть реализована оптимизация операций матричного умножения (GEMM - General Matrix Multiply) для повышения производительности на нейронных процессорах (NPU), что должно обеспечить:

1) Уменьшение времени выполнения операций матричного умножения.
2) Поддержка многопоточности и распределённых вычислений.
3) Автоматическое распределение нагрузки на процессоры и ускорители.

4.1.2 Добавление поддержки новых типов слоев

Необходимо расширить функциональность PuzzleLib путём добавления поддержки новых типов слоев для нейронных сетей:

1) Поддержка новых типов свёрточных слоев для улучшения анализа изображений.
2) Добавление специализированных слоев для работы с временными рядами и текстами (например, LSTM, GRU).
3) Реализация более эффективных алгоритмов для обработки данных в режиме реального времени.

4.1.3 Улучшение производительности существующих слоев

Существующие слои, такие как Fully Connected, Convolutional и Pooling, должны быть доработаны для повышения эффективности их работы на современных архитектурах:

1) Использование новых методов вычислений с минимальными задержками.
2) Увеличение скорости вычислений за счёт уменьшения использования ресурсов памяти.
3) Адаптация слоев для работы с большими наборами данных.

4.1.4 Поддержка распределённых вычислений

В модифицированной версии PuzzleLib должна быть предусмотрена возможность распределённых вычислений:

1) Обеспечение взаимодействия между несколькими узлами в кластерах с использованием стандартов распределённых систем (например, MPI, NCCL).

2) Возможность горизонтального масштабирования нейронных сетей на кластерах с поддержкой GPU и NPU.

3) Поддержка выполнения вычислений на гетерогенных системах (например, на комбинации CPU, GPU, NPU).

4.1.5 Оптимизация использования памяти

Введение механизмов для автоматического управления памятью во время выполнения сложных операций.

1) Оптимизация работы с памятью для сокращения объёмов потребляемых данных и увеличения эффективности обработки на высокопроизводительных устройствах.

2) Реализация функций для динамического распределения и освобождения памяти во время работы программы.

4.1.6 Поддержка кастомных операций и расширений

Предусмотреть возможность добавления пользователями собственных операций и модулей:

1) Внедрение API для интеграции кастомных операторов, не предусмотренных базовой библиотекой.

2) Поддержка пользовательских слоев с использованием языков C++ и Python.

4.2 Требования к надёжности

4.2.1 Обеспечение устойчивого функционирования

Модифицированная библиотека должна корректно функционировать при высоких нагрузках и интенсивной работе с большими данными, избегая сбоев и перегрузок.

4.2.2 Контроль входной и выходной информации

Встроенные механизмы валидации входных данных для исключения ошибок при запуске моделей. Обеспечение контроля корректности выхода нейронных сетей и их прогнозов.

4.2.3 Время восстановления после отказа

Программа должна иметь возможность быстро восстанавливать выполнение задач после отказа системы или оборудования:

1) Автоматическое сохранение состояния работы модели (checkpoint) и её параметров.

2) Поддержка возможности перезапуска процессов с момента последнего сохранения без потери данных.

4.3 Условия эксплуатации

4.3.1 Рабочая температура

Система должна корректно работать в условиях серверного оборудования с рабочей температурой от +10°C до +35°C.

4.3.2 Влажность и электромагнитная совместимость

1) Относительная влажность окружающего воздуха: от 30% до 70% без конденсации.

2) Программа должна быть устойчива к электромагнитным помехам, типичным для дата-центров и серверных помещений.

4.3.3 Требования к персоналу

Для работы с программой необходим обслуживающий персонал с уровнем квалификации:

1) Опыт работы с нейронными сетями и высокопроизводительными вычислениями.

2) Знание языков программирования Python и C++.

3) Опыт настройки и эксплуатации кластерных систем с использованием GPU и NPU.

4.4 Требования к составу и параметрам технических средств

4.4.1 Минимальные технические требования

1) Процессор: многоядерный процессор с поддержкой параллельных вычислений (Intel Xeon или эквивалент).

2) Оперативная память: не менее 64 ГБ.

3) Графический процессор: Nvidia GPU с поддержкой CUDA (например, Nvidia A100) или аналогичные ускорители.

4) Жёсткий диск: не менее 1 ТБ SSD для хранения данных и временных файлов.

4.4.2 Рекомендуемые технические средства

1) Поддержка гетерогенных вычислительных платформ (CPU, GPU, NPU).

2) Многоузловая кластерная система для распределённой обработки больших объёмов данных.

3) Использование облачных вычислительных платформ (AWS, Microsoft Azure, Google Cloud) для динамического масштабирования ресурсов.

4.5 Требования к информационной и программной совместимости

4.5.1 Совместимость с информационными структурами

1) Входные и выходные данные должны поддерживать форматы JSON, CSV, HDF5 для структурированных данных.

2) Для хранения и передачи моделей используется формат ONNX (Open Neural Network Exchange).

4.5.2 Совместимость с программным обеспечением

1) Программа должна быть совместима с популярными фреймворками глубокого обучения, такими как TensorFlow и PyTorch.

2) Интеграция с библиотеками для параллельных вычислений, такими как OpenMP, MPI, и NCCL для оптимизации вычислительных процессов.

4.5.3 Требования к исходным кодам

1) Все модификации должны быть реализованы на языках программирования Python и C++.

2) Код программы должен быть модульным, с возможностью легкого расширения и поддержки.

3) Строгое соответствие стилям кодирования и документации PEP-8 для Python и Google C++ Style Guide.

5 ТРЕБОВАНИЯ К ДОКУМЕНТАЦИИ

- Инструкции по установке и эксплуатации должна быть дополнена информацией по использованию новых функций и модулей.

6 ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

6.1.1 Экономическая эффективность

Модифицированная библиотека обеспечит улучшение производительности моделей на 30-40%, что снизит время обучения и эксплуатационные расходы.

7 СТАДИИ РАЗРАБОТКИ

- Анализ требований.
- Разработка и тестирование модификаций.
- Передача заказчику.

8 КОНТРОЛЬ И ПРИЕМКА

8.1 Виды испытаний

- Функциональное и нагрузочное тестирование.
- Интеграция с существующими системами заказчика.

8.2 Приемка

Проводится после успешного выполнения всех тестов и устранения выявленных ошибок.

Лист регистрации изменений

Изм.	Номера листов (страниц)				Всего листов (страниц) в документе	Номер документа	Входящий номер сопроводительного документа и дата	Подпись	Дата
	изменённых	заменённых	новых	аннулированных					