**ИТМО**

**Факультет Програмной Инженерии и Компьютерных Технологий**

**Лабораторная работа № 3**

“Исследование и применение инструментов для развертывания моделей машинного обучения”

Выполнили: Распопин Даниил Игоревич, Сыюань Хуан, Р4155

Преподаватель: Старобыховская Анастасия Александровна

Санкт-Петербург

2024

**СОДЕРЖАНИЕ**

[Задание 3](#_gjdgxs)

[Основные этапы 3](#_30j0zll)

[Выводы 5](#_1fob9te)

# Задание

В работе была поставлена следующая задача: взяв лучшую модель, полученный в результате выполнения предыдущей лабораторной работы, замерить скорость и качество обучения и прогнозировния модели как на процессоре, так и на графической карте. После этого, экспортировать модель в формате ONNX. Загрузив модель в формате ONNX, повторно провести замеры. Наконец, экспортировать модель в виде сервиса BentoML. Измерить скорость работы запущенного сервиса.

# Основные этапы

На начальном этапе работы группа распределила задачи следующим образом:

* Сыюань Хуан – экспорт моделей и замер быстродействия ONNX.
* Даниил Распопин – экспорт моделей и замер быстродействия BentoML.

На первом этапе необходимо было замерить скорость и качество работы модели на CPU и на GPU. Полученные показатели:

* Время обучения на CPU: 0.9745 секунд
* Время прогнозирования на CPU: 0.0005 секунд
* Время обучения на GPU: 0.8386 секунд
* Время прогноза на GPU: 0.0004 секунд

На втором этапе необходимо было экспортировать модель в формате ONNX, загрузить полученную модель и повторно замерить быстродействие прогноза. При помощи установки параметра providers=['CPUExecutionProvider'] запуск CPU-модели гарантировано осуществлялся на CPU.

Результаты замера:

* Время прогноза на CPU: 0.0026 секунд
* Время прогноза на GPU: 0.0056 секунд

На третьем этапе работы необходимо было запустить модель в виде сервиса BentoML и замерить быстродействие сервиса. Быстродействие замерялось по следующему алгоритму:

1. В терминале ОС Linux запускался сервис BentoML: bentoml serve service:svc
2. Вызывался и исполнялся Python-код, отправляющий POST-запрос по адресу bentoml-сервиса (0.0.0.0:3000/predict), в качестве данных для POST-запроса передавался массив, на котором предполагалось выполнять прогноз
3. Время выполнения запроса и получения ответа от сервера замерялось и записывалось.

Результат замера: время выполнения запроса к сервису и получения ответа составило 0.5833 секунды.

# Выводы

Обучаемые модели машинного и глубокого обучения могут быть тяжеловесны, иметь разные форматы сохранения, что препятствует быстрому внедрению этих моделей в production-среду. Формат ONNX и BentoML – это технологии, которые позволяют стандартизировать запуск моделей. В рамках работы нами было отмечено удобство работы с этими технологиями, хорошая задокументированность, наличие обучающих материалов и примеров кода.