Отзыв научного руководителя на выпускную квалификационную работу студента Вологина Ильи Олеговича, обучающегося по направлению 02.03.03 (Математическое обеспечение и администрирование информационных систем)
Тема выпускной квалификационной работы:
“Поддержка исполнения моделей машинного обучения в виртуальных машинах JavaScript для платформы KInference”

В последние годы модели машинного обучения находят применение во многих разрабатываемых приложениях, не исключая и веб-приложения. Для запуска моделей, как правило, используются специальные библиотеки, и одной из таких является библиотека KInference, разрабатываемая в компании JetBrains. Первоначально библиотека KInference поддерживала работу только в JVM проектах, поэтому перед Ильей Олеговичем была поставлена задача поддержать исполнение моделей машинного обучения в виртуальных машинах JavaScript для библиотеки KInference.

В первую очередь Илья Олегович поддержал компиляцию KInference для JavaScript с помощью Kotlin Multiplatform, но полученная реализация ожидаемо показала довольно низкую производительность. После этого Илья Олегович изучил оптимизации математических операций, доступных в браузере, а также сравнил реализации аналогичных библиотек для запуска (Tensorflow.js и ONNXRuntime), использующих данные оптимизации. В конце Илья Олегович изучил существующую архитектуру KInference, разделил архитектуру на несколько основных модулей (API модуль и модуль с исполнением графа), разработал и интегрировал новую реализацию KInference для JavaScript с использованием тензорных операций из библиотеки Tensorflow.js и технологии WebGL: реализовал необходимый API для создания реализаций KInference, реализовал поддержку исполнения графа ONNX, реализовал 22 оператора ONNX, необходимых для запуска моделей GPT-2 и BERT.

Апробация платформы KInference на двух существующих моделях GPT-2 и BERT, используемых в проекте Grazie Platform компании JetBrains, показала работоспособность разработанной в рамках ВКР реализации KInference. Кроме того, были достигнуты показатели менее 300 мс — барьера времени отклика пользователю, после которого время работы модели становится заметным глазу на задаче автодополнения текста.

В ходе работы Илья Олегович регулярно взаимодействовал с научным руководителем и консультантом из компании JetBrains, выполнял задачи в срок и оперативно устранял выявленные замечания к работе; за время выполнения ВКР проявил себя как грамотный программный инженер. Считаю, что работа полностью соответствует требованиям к ВКР бакалавра и заслуживает оценки “отлично”.

к.т.н., доцент кафедры системного программирования СПбГУ Т.А. Брыксин

дата: 02.06.2022