Белорусский государственный университет

Факультет прикладной математики и информатики

Кафедра биомедицинской информатики

ЗАДАНИЕ НА КУРСОВУЮ РАБОТУ

Студент Бинцаровский Леонид Петрович

1. Тема: Анализ эффективности нейросетевых вычислений с учетом аппаратных возможностей платформ

2. Срок представления курсовой работы к защите: 15.05.2024 г.

3. Исходные данные для научного исследования (проектирования)

Marat Dukhan and Frank Barchard. (2023). Half-precision Inference Doubles On-Device Inference Performance [Электронный ресурс]: The TensorFlow Blog. – Redmond, WA: Google. – Режим доступа: https://blog.tensorflow.org/2023/11/half-precision-inference-doubles-on-device-inference-performance.html.

ONNX Runtime Execution Providers [Электронный ресурс]: ONNX Runtime. – Redmond, WA: Microsoft Corporation, 2021. – Режим доступа: https://onnxruntime.ai/docs/execution-providers/.

Lee, J., Chirkov, N., Ignasheva, E., Pisarchyk, Y., Shieh, M., Riccardi, F., Sarokin, R., Kulik, A., & Grundmann, M. (2019). On-Device Neural Net Inference with Mobile GPUs [Электронный ресурс]: Google Research. – Режим доступа: https://ar5iv.labs.arxiv.org/html/1907.01989.

Qualcomm Technologies, Inc. (2024). Qualcomm Neural Processing SDK for Windows on Snapdragon [Электронный ресурс]: Qualcomm Developer Network. – Режим доступа: https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk/windows-on-snapdragon.

Ashfaq, S., AskariHemmat, M., Sah, S., Saboori, E., Mastropietro, O., & Hoffman, A. (2022). Accelerating Deep Learning Model Inference on Arm CPUs with Ultra-Low Bit Quantization and Runtime [Электронный ресурс]: Deeplite Inc. – Montreal, Canada: Deeplite Inc., 2022. – Режим доступа: https://arxiv.org/pdf/2207.08820.pdf.

4.1. Ознакомление с функциональностью фреймворков ONNXRuntime, Tensorflow Lite, NCNN, OpenCV DNN.

4.2. Описание методики тестирования

4.3. Разработка общей кросс-платформенной части приложения для замера скорости вычислений нейросетей на языке C++.

4.4. Реализация платформо-зависимых компонент (ONNRuntime DirectML и CoreML, CMake-конфиги для подключения фреймворков для CPU и GPU инференса под Windows, Linux, MacOS.

4.5. Тестирование фреймворков

4.6. Анализ эффективности квантизации сетей

4.7. Подготовка отчета

Руководитель курсового проекта _____

подпись, дата инициалы, фамилия

Задание принял к исполнению _____ 09.02.2024

подпись, дата