



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ  
ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ  
(национальный исследовательский университет)»

Институт № 8 «Информационные технологии и прикладная математика» Кафедра 806  
Направление подготовки: 01.03.02 «Прикладная математика и информатика» Группа М8О-407Б-17  
Квалификация бакалавр

УТВЕРЖДАЮ

Зав. Кафедрой 806 Крылов С.С.  
(№ каф.) (подпись) (фамилия, инициалы)

« 09 » февраля 2021 г.

## ЗАДАНИЕ на выпускную квалификационную работу бакалавра

Студенту Бронникову Максиму Андреевичу  
(фамилия, имя, отчество полностью)

Руководитель Ревизников Дмитрий Леонидович  
(фамилия, имя, отчество полностью)  
д.ф.-м.н., профессор, профессор каф. 806 МАИ  
(ученая степень, ученое звание, должность и место работы)

1. Наименование темы: Квантизатор нейронных сетей с эффективным умножением матриц

2. Срок сдачи студентом законченной работы 24.05.2021

3. Техническое задание и исходные данные к работе Разработать инструмент квантизации обученных нейронных сетей, организовать схему работы и хранения квантизованных слоев на базе нейросетового фреймворка Samsung ONE с поддержкой сериализации и десериализации, реализовать вычислительное ядро для эффективного запуска квантизованных сетей, проделать анализ показателей реализованного нейронного слоя.

4. Перечень подлежащих разработке разделов и этапы выполнения работы

п/п	Наименование раздела или этапа	Трудоёмкость в % от полной трудоёмкости дипломной работы	Срок выполнения	Примечание
1	Изучение низкобитовой квантизации, анализ существующих методов и проблем.	5	09.02.2021-18.02.2021	
2	Доработка выбранного метода, анализ и математического обоснование.	10	19.02.2021-28.02.2021	
3	Реализация сериализации и десериализации для квантизованного слоя.	5	01.03.2021-10.03.2021	

4	Описание структуры и схемы взаимодействия слоя с соседними вершинами в промежуточном графовом представлении фреймворка.	10	11.03.2021-20.03.2021	
5	Написание эффективного вычислительного ядра в интерпретаторе моделей, оптимизация под работу с кешом, описание перехода из промежуточного представления в вычислительный граф.	20	21.03.2021-10.04.2021	
6	Создание инструмента получения моделей с квантизованными слоями на основе обученных нейронных сетей.	30	11.04.2021-10.05.2021	
7	Исследование и сравнение показателей исходных и квантизованных моделей.	5	11.05.2021-13.05.2021	
8	Оформление ВКР	15	13.05.2021-24.05.2021	


## 5. Перечень иллюстративно-графических материалов:


№ п/п	Наименование	Количество листов
1	Презентация	20

## 6. Исходные материалы и пособия

1. LQ-Nets: Learned Quantization for Highly Accurate and Compact Deep Neural Networks // Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, Gang Hua; Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 365-382
2. Bqgemm: matrix multiplication with lookup table for binary-coding-based quantized dnns // Yongkweon Jeon, Baeseong Park, Se Jung Kwon, Byeongwook Kim, Jeongin Yun, and Dongsoo Lee; arXiv preprint arXiv:2005.09904, 2020.
3. Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv:1806.08342, 2018.

7. Дата выдачи задания 09.02.2021

Руководитель  (Ревизников Д.Л.)  
(подпись)

Задание принял к исполнению  (Бронников М.А.)  
(подпись)