**Федеральное государственное автономное образовательное учреждение**

**высшего образования**

**«Национальный исследовательский университет «Высшая школа экономики»**

**Факультет компьютерных наук**

**Образовательная программа Прикладная математика и информатика**

**Направление подготовки 01.03.02 Прикладная математика и информатика**

**бакалавриат**

**О Т Ч Е Т**

**по преддипломной практике**

**Выполнил студент гр.174**

**Тропин Федор Владиславович**

**(ФИО)**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***(подпись)***

**Проверил:**

Доцент, Ильвовский Дмитрий Алексеевич

*(должность, ФИО руководителя практики)*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(подпись)*

*\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(дата)*

**2021 год**

# Квантизация нейронных сетей в задачах обработки естественного языка

Содержание

[Квантизация нейронных сетей в задачах обработки естественного языка 2](#_Toc70083664)

[Аннотация 3](#_Toc70083665)

[Введение 3](#_Toc70083666)

[Обзор литературы 4](#_Toc70083667)

[Методы 5](#_Toc70083668)

[Результаты 7](#_Toc70083669)

# Аннотация

Современные нейронные сети и методы машинного обучения имеют огромное количество параметров. Также на определенных устройствах, в частности смартфонах и планшетах, вычисления значений в типах данных с плавающей точкой производятся крайне медленно, что приводит к полной бесполезности моделей, которые должны работать в реальном времени. В данной работе мы сравниваем эффективность применения различных методов фейковой квантизации. По результатам проделанной работы мы предполагаем получить данные об эффективности примененных алгоритмов на разнообразных задачах и при различных начальных параметрах.

# Введение

С каждым годом нейронные сети становятся все больше и тяжелее, количество слоев и параметров в них растет. Это приводит к замедлению расчетов и увеличению занимаемой весами памяти: как оперативной при вычислениях, так и постоянной при сохранении обученных моделей. Причиной этому являются расчеты, производимые при помощи чисел с плавающей точкой. Также с каждым годом растет как рынок смартфонов, так и приложений, разрабатываемых для них. С ростом качества и эффективности нейронных сетей их все чаще интегрируют для решения определенных задач прямо на устройства. При этом смартфоны имеют недостаток – на их процессорах затруднительно производить вычисления в вещественных числах. Обе проблемы возможно решить при помощи технологии приведения данных к целочисленному виду. Однако грубое округление чисел приводит к большим потерям точности. Существует специальный метод под названием квантизация. Данные сначала масштабируются, потом округляются и при выходе значений за пределы используемой битности округляются до минимумов или максимумов соответственно. При некоторых подходах так же возможно обучение самой квантизирующей части модели со своими функциями потерь. Целью нашей работы является определение эффективности применения алгоритма Qsin во время обучения квантизованных моделей на различных задачах. В ходе нашего исследования, перед нами были поставлены задачи изучения и реализация самих способов квантизации, применение полученных инструментов к модели, запуск и сравнение результатов при разнообразных начальных параметрах, а также при различных типах квантизации.

# Обзор литературы

В качестве основной модели для исследования влияния квантизации на качество результатов была взята авторегрессионная модель XLnet [1] – базовая версия состоит из 12 входных слоев, 768 скрытых и 12 выходных, в то время как большая модель содержит 24 входных, 1024 скрытых и 16 выходных слоев. Мы планируем использовать обе модели для сравнения процентного роста или уменьшения интересующих нас величин. Была выбрана именно эта модель в силу ее достаточно сложного внутреннего устройства, большого количества параметров, а также разнообразия решаемых ею задач. Согласно официальной статистике XLNet показывает результаты лучше, чем BERT в 20 задачах, а добивается эффективности уровня state-of-the-art в 18 задачах. Это позволит получить обширный набор данных для сравнения, продемонстрирует силу сжатия весов и ускорения вычислений, потенциально может помочь реализовать на мобильных устройствах большое количество задач, связанных с обработкой текста и его анализом.

Модель планируется применить к следующим проблемам: классификации текстов и вопросно-ответным системам. Для тестирования обоих задач используются датасеты, результаты которых приведены в официальной статье об xlnet.

Классификация проводится на бенчмарке glue. Это тест на общее понимание текстов языковыми моделями, включающий в себе 12 датасетов. Xlnet тестировался на 8 из них.

Вопросно-ответные задачи были указаны на официальном гитхабе как одни из тех, на которых при помощи XLNet удается достичь качества уровня state-of-the-art. Исходя из этого мы решили использовать инструменты доступные через библиотеку huggingface. Данный ресурс предоставляет как предобученные модели, токенайзеры, датасеты, так и рабочие тетради с удобным интерфейсом.

# Методы

Глобально процесс квантизации заключается в следующих трех процедурах: сначала тензор скалируется, далее происходит округление и в качестве завершающей функции используется обрезание значений вылезающих за пределы допустимых границ. В процессе каждой операции используется поэлементное действие. Во всей этой последовательности действий стоит выделить два важных параметра – значение используемой битности и величина, при помощи которой производится скалирование.

Битность задает количество байт, при помощи которой можно закодировать целое число. Исходя из данной величины производится финальное округление. Так же стоит отметить, что обычно битность задается сразу для всей модели, а значит и для всех слоев одинаковой. Зачастую эта константа равняется двойке, возведенной в определенную степень, но не больше, чем 2 в 16.

Способ же расчета и применение делителя тензоров является центральной в этой работе и определяет сам тип квантизации. На данном этапе важно обозначить, что существует два отдельных тензора на каждом слое, которые подлежат квантизации: сами входные данные, которые будут обрабатываться и взаимодействующие с ними веса. Для обеих этих частей статистика ведется отдельно.

На данный момент существует несколько подходов к квантованию данных, глобально делящиеся на статическое и динамическое. Динамическая квантизация использует крайние значения из набора поступивших данных, чаще всего – максимум модуля значений тензора по батчу, в то время как статическая определенным образом сохраняет нужное значение и постоянно хранит его в памяти. В свою очередь статистические квантизации сами различаются по своим подходам на необучаемые и с дообучением.

Скейлеры необучаемого квантования считаются как определенная функция от максимальных значений по модулю батча. В зависимости от задачи и типов данных это может быть среднее определенного типа, медиана, максимум по всем батчам или любая другая подходящая под проблему функция от большого количества данных. Это значение рассчитывается отдельно от циклов обучения и валидации модели, т.к. во время расчетов необходимо не только пропускать данные через всю модель без дальнейшего изменения весов, но и производить сбор, обновление и сохранение статистики на каждом слое.

Если же применяется процесс дообучения после применения квантизации, то сразу возникает несколько проблем. Самая незначительная из них это то, что для того, чтобы значения скаляров менялись, их надо обучать особым образом. Первая же тонкость, в основе которой лежит математика, возникает на уровне округления. Дело в том, что во время квантизации именно этот момент является узким горлышком для процесса обратного распространения – у данной функции нет градиента. Это значит, что функцию придется переписать особым образом с наследованием от необходимых классов используемого фреймворка, определив градиент равным единице во всех местах, иначе говоря, при пересчете весов во время обратного распространения на данном этапе функция возвращает то же значение, что и пришло в нее. В целом уже после этого можно наблюдать рост качества модели и изменения значений делителей. Но, помимо этого, нам интересно что бы мы как можно меньше теряли информации во время самого округления. Этого можно добиться, наложив отдельную функцию потерь веса и активации. Данные имеют разные реализации, однако перед нами в этой работе была поставлена задача исследовать Qsin. Данный алгоритм принимает вектор, поделенный на необходимый скаляр и поэлементно, в зависимости от попадания в определенные сегменты, применяет свой набор преобразований. Итоговая сумма значений со всех слоев добавляется к стандартному значению потерь с определенным коэффициентом. Во время распространения обратной ошибки данное значение так же учитывается и имеет свой граф, в силу этого у нас не только уменьшается ошибка при скалировании, но и так же подбираются лучшие значения для приведения значений в тензорах к максимально близким к целым числам.

# Результаты

На данный момент уже реализованы все способы квантизации, в том числе и функция потерь Qsin. Более того, все алгоритмы протестированы на простой сети, решающей задачу классификации комментариев с IMDb. Все подходы показали свою работоспособность. Качество квантизации оценить тяжело в силу простоты решаемой задачи и ни один метод не отклонился от эталонного решения больше чем на порядок шума.

В качестве результатов предполагается объединить все инструменты квантизации в единый класс с удобным пользовательским интерфейсом. Следующим шагом станет добавление либо расширение полученным классом слоев XLNet. Далее планируется запуск на обозначенных задачах с разными типами квантизации и при разнообразных стартовых параметрах. Полученные результаты, включающие в себя написанный код, модифицированные базовые решения, текстовый отчет и таблицы качества моделей планируется выложить в открытый доступ на github.