# Обзор литературы

# Найти место куда вставить

Хотелось бы отметить, что зачастую необходимо скачивать данные на локальный компьютер для дальнейшего использования при ограничении доступа к сети.

# Методы

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.

Одним из основных показателей, которые были выявленны в процессе исследования, это хорошая адаптивность xlnet к процессу квантизации. На всех протестированных задачах набора датасетов glue, на задаче ner и на модели CNN при квантизации модели в большую из исследованных битностей модель не теряла в качестве и, зачастую, даже поднимала ее