О Т Ч Е Т

О лабораторной работе № 1
Тема задания: «Учебный проект, посвященный сжатию текста при помощи авторегрессионной модели»

Обучающийся:
Бессонницын Евгений Сергеевич,
№ группы М4145

Описание работы

Данный проект реализует сжатие текста с использованием LSTM модели и адаптивного арифметического кодирования (ААК). В папке ./data/ находится текстовый файл enwik5 для сжатия, являющийся первыми 10^5 байтами enwiki-20060303-pages-articles.xml.

Кодировщик использует LSTM для вычисления вектора вероятностей следующего символа на основе предыдущих. Фактическое значение символа далее кодируется с помощью ААК. Затем веса модели обновляются. Это повторяется для всех символов файла.

Во время декодирования происходит симметричный процесс, декодирование с помощью ААК, предсказание символа при помощи LSTM, обновление модели и т.д. Декодер работает симметрично, поэтому нет необходимости передавать параметры модели.

В качестве ААК используется реализация из
https://github.com/nayuki/Reference-arithmetic-coding

Описание задания к лабораторной работе

Улучшить код так, чтобы он:

- либо на том же сжатии показывал уменьшение времени кодирования и декодирования на 100%;
- либо обеспечивал улучшение коэффициента сжатия на 100% при тех же временных затратах.

Можно улучшать следующие модули:

- Предобработка: заменять слова из предварительно созданного словаря уникальным кодом, использовать идею из Byte-Pair Encoding токенизации и т.д.;

- Нейронная сеть: использование другой архитектуры (GRU, Transformer и др.), изменить количество слоёв, функции активации, связи между слоями и т.д.;
- Арифметический кодер: учесть возможную память источника, оценка вероятностей и т.д.

Требования к реализации:

- Результаты должны быть продемонстрированы на enwik5 из папки ./data/;
- Восстановленый после сжатия файл должен полностью совпадать с оригинальным;
- В результатах приложить таблицу выше, обновив значения базового решения для вашего устройства и добавив строчку с улучшенным решением.

Описание реализации

Для ускорения работы алгоритма были изменены его гиперпараметры, а именно batch_size, seq_length, rnn_units, num_layers и embedding_size. Их уменьшение привело к уменьшению времени кодирования и декодирования более чем на 100%. Результаты прилагаются ниже:

Версия	Исходный размер, байты	Размер после сжатия, байты	Коэффициент сжатия	Затраченное время, с
Baseline	100000	47150	2.12	1654
Modified	100000	49565	2.01	208

Эксперименты проводились на MacBook Air, чип Apple M1, Оперативная память 16 гб.

Исходный код: https://github.com/Turukmokto/LSTM-compresor