Промежуточный отчет

Петров С.
Факультет компьютеных наук
Высшая школа экономики
Москва
stasdp@mail.ru

Аксенов К.
Факультет компьютеных наук
Высшая школа экономики
Москва
akskir@gmail.com

Abstract

В данном отчеты мы рассматриваем оригинальную статью с конференции ICML 2018 GroupReduce: Block-Wise Low-Rank Approximation for Neural Language Model Shrinkin. Мы посторяем предложенный в статье метод и пытаемся воспроизвести полученные результаты. На сегоднящний день много сил тратится исследователями на то, что бы найти новые способы сжатия глубоких нейронных сетей для использования в реальном времени на устройствах с ограниченной памятью. В оригинальной статье предложен метод сжатия весов слоев Embedding и софтмакс.

1 Краткое описание задачи и алгоритмов решения

В данной статье авторы фокусируются на задачи сжатия нейронных сетей для обработки естесвенного языка, рассматриваются задачи предсказания следующего слова и машинного перевода. Несмотря на то, что предложено много методов для сжатия нейросетей, большинство из них сфокусированы на сжатии сверточных слоев и работают в несколько раз хуже для сжатия слоев Embedding. В слое Embedding каждая строка матрицы представляет собой токен, каждый токен встречается с определенной частотой, которая описывается законом Ципфа.

1.1 Сингулярное разложение

Обычное сингулярное разложение с выбором ранга матрицы, который мы ожидаем получить на выходе

1.2 Взвешенное сингулярное разложение

Мы не зря упоминули закон Ципфа в постановке задачи, потому что в модели языка частота токена важна. Поэтому мы будем применять сингулярное разложнение к матрице QA, где $Q = diag(\sqrt{q_1}, ..., \sqrt{q_n},)$ где q_i - частота слова в обучающей выборке, а A - исходная матрица Embedding.

1.3 Взвешенное сингулярное разложение по блокам

Следующим шагом можно разбить матрицу QA, на блоки в соответсвии с частотой слов в блоке и применить к ним взвешенной сингулярное разложение. Ранг для разложения фиксирован.

Preprint. Work in progress.

Таблица 1: Результаты применения методов сжатия

Авторство/Метод	SVD	weighted SVD	block weighted SVD	dymanic rank	GroupReduce
Оригинальные	161.44	155.10	135.19	129.63	127.26
Воспроизведенные	149.09	150.65	150.51	146.91	148.34

1.4 Взвешенное сингулярное разложение по блокам с динамическим рангом

В предыдущем методе мы никак не изменяли ранг, но это не совсем правильно, поскольку мы хотим, лучше представлять информацию о частотных словах в угоду плохому представлению менее частотных. Мы определим зависимость ранга блока от средней частоты слов в этом блоке. Таком образом для кластера p ранг определяется как $\frac{f_p}{f_c}r$, где f_p - средняя частота слов в кластере p, f_c - средняя частота слов в кластере с наименьшим средней частотой слов, r - минимальный ранг, который определяет наш "бюджет"

1.5 Предложенный алгоритм GroupReduce

Алгоритм предложенный в статье собирает в себе все вышеперечисленные методы и минимизирует ошибку реконструкции. Для расчета ошибки мы спроецируем A на базис V^p и выясним насколько ошибка уменьшилась. Итак, если $\|A_i - V^p(V^p)^T A_i\| > \|A_i - V^{\overline{p}}(V^{\overline{p}})^T A_i\|$ тогда мы переносим токен i в из кластера p в кластер $\overline{p}$. В конце мы возвращаем $[U^1(V^1)^T, ...U^c(V^c)^T]$

2 Эксперименты и результаты

Базовая модель PTB-small. Она была обучена на 12 эпохах. В оригинальной статье, авторы не приводят количество эпох. Авторы последовательно применяли все предложенные выше техники и оценивали перплексию. Мы повторили эксперимет авторов. В таблице 1 приведны результаты авторов оригинальной статьи и авторов настоящего отчета.

3 Анализ результатов и выводы

Как можно видеть по таблице 1 , не все результаты были воспроизведены. Причин на то, может быть несколько:

- В оригинальной статье не во всех деталях описан алгоритм GroupReduce. Большим вопросом в этом является добавление и удаление слов из кластера в кластер. Каким образом его производили, и каким образом возвращались к изначальной матрице, поскольку при изменении порядка строк в матрице Embedding меняется ее структура, где на месте i-ого слова стоит слово с индексом j.
- В статье не приведены конкретные параметры алгоритма GroupReduce
- В статье не приводятся ранги к которым приводят матрицы Embedding.

В дальнейшем мы сравним полученные результаты с другими техниками сжатия и постараемся найти способ для вопсроизведения результатов.