Отчёт по финальному проекту курса NLP

Май 2024

Содержание

1 Введение

[аналог abstract, рассказать почему работа актуальна, чем уникальна]

Мы будем уменьшать Bert. В эпоху больших языковых моделей актуальны эффективные техники уменьшения моделей и ускорения их инференса. Наша работа интересна самим способом достижения результата. В основном такие модели дистиллируют или квантизуют. Мы применим прунинг. Прунинг мы воспринимаем как правильную иниализацию сети ученика для дальнейшей дистилляции. Уникальность работы в том, что мы получим сеть меньшего размера, но сопоставимого по качеству с оригинальной сетью

1.1 Команда

Список участников проекта представлен в следующей таблице:

Участник	E-mail	Роль
Петров Александр Васильевич	petrov1c@yandex.ru	*{роль}*
Сентюрев Михаил Алексеевич	mixail_sen@outlook.com	*{роль}*

2 Цели

[чего хотим достичь в рамках проекта]

В рамках проекта мы хотим показать как можно уменьшить размер сети одновременно с уменьшением вычислительной сложности, получив в результате более быструю и компактную сеть без заметной потери в качестве

3 Сопутствующая работа

[нужно рассмотреть существующие к достижению поставленной цели, дать ссылки на источники (где сказано про эти подходы)]

Одна из известных работ, посвященных теме уменьшения моделей, это работа по дистилляции оригинального берта. В ней была применена техника дистилляции, поэтому полученная модель получила название DistilBERT. Авторы смогли уменьшить количество параметров на 40%, сохранив при этом 97% качества (согласно бенчмарку GLUE)

4 Исходные данные

[рассказ про саму модель для дальнейшей работы, а также про данные, используемые в процессе (откуда, что из себя представляют)] Мы хотим получить хороший энкодер для предложений русского языка, который можно будет использовать в качестве бекбона для разных классов задач. В качестве исходной модели мы возьмем модель LaBSE от компании Гугл. Гугл создал эту бертподобную модель для целей перевода для более чем 100 языков. Она позволяет переводить тексты с разных языков в единое скрытое пространство, другими словами, одно и тоже сказанное на разных языках отображается в похожие по косинусой близости вектора. Мы возьмем не оригинальную модель, а ее двуязычную версию, которая была получена, путем оставления в токенайзере только русских и английских токенов, в результате чего модель уменьшилась в 4 раза, так как уменьшение количества токенов естественным образом привело к уменьшению слоя эмбеддингов и выходного слоя пуллинга (“ссылка на хабр”) Так же упомянем, что автор LaBSE-en-ru получил из нее дистиллированную версию, которая в русскоязычном сообществе хорошо известна под названием rubert-tinu2 (“ссылка на хабр”). В данной работе мы так-же получим уменьшенную версию LaBSE-en-ru, применив немного другой подход.
В качестве данных для обучения будем использовать () двуязычный парный корпус предложений (ссылка)на 1000000

5 Метод

[подробно расписываем каким путём хотим получить результат (цель), сюда должна войти теория]

Будем уменьшать модель выкинув из нее половину всех голов внимания на первом шаге. И затем будем обучать полученную модель путем дистилляции знаний от оригинальной модели. Далее по тексту мы можем называть оригинальную модель - моделью учителем, а получаемую в рамках данной работы модель - моделью учеником Замерять качество мы будем на бейнчмарке русскоязычных энкодеров Енкодечка (ссылка)

Оригинальная модель имеет 12 слоев внимания по 12 голов в каждом. Мы будем применять разложение в ряд Тейлора, чтобы понять какие головы оказывают меньше всего влияния на выход энкодера. Этот способ хорош по нескольким причинам: во-первых он хорошо теоретически обоснован, во-вторых коэффициенты ряда мы можем получить через накопление градиентов и далее по величене накопленных градиентов определять важность голов. Здесь есть некоторая сложность, заключающаяся в том, что все головы одного слоя соединены в 3 общих линейных слоя. Так сделано для целей оптимизации вычислений, но нам придется фактически “вырезать” головы из этих общих слоев

Итак, сначала мы получаем запруненную модель, сохраняем её, и будет ее также сравнивать с LaBSE, LaBSE-en-ru, rubert-tinu2 Такое сравнение добавит нам наглядности. С одной стороны мы хотим получить модель не сильно хуже, чем LaBSE и LaBSE-en-ru С

другой стороны мы точно не хотим проиграть в качестве rubert-tinu2. Зачем нам получать модель хуже :)

Далее, только что запруненную модель мы начинаем стягивать к выходам модели LaBSE-en-ru. Использовать будем комбинированный лосс, состоящий из лосса, оценивающего похожесть русских и английских предложений и лосса по выходам сети учителя и сети ученика

6 Эксперимент

6.1 Критерии

[критерии оценки качества результата (как валидации, так и теста)] Качество модели мы оцениваем на бенчмарке, энкодечка. Бенчмарк оценивает качество эмбеддингов по следующим типам задач * Semantic text similarity (STS) на основе переведённого датасета STS-B; * Paraphrase identification (PI) на основе датасета paraphraser.ru; * Natural language inference (NLI) на датасете XNLI; * Sentiment analysis (SA) на данных SentiRuEval2016. * Toxicity identification (TI) на датасете токсичных комментариев из OKMLCup; * Inappropriateness identification (II) на датасете Сколтеха; * Intent classification (IC) и её кросс-язычная версия ICX на датасете NLU-evaluation-data, который я автоматически перевёл на русский. В IC классификатор обучается на русских данных, а в ICX – на английских, а тестируется в обоих случаях на русских. * Распознавание именованных сущностей на датасетах factRuEval-2016 (NE1) и RuDReC (NE2). Эти две задачи требуют получать эмбеддинги отдельных токенов, а не целых предложений; поэтому там участвуют не все модели.

6.2 Инструкции

[инструкции для воспроизведения эксперимента]

Инструкции по подготовке окружения, получении данных, подключении системы отслеживания экспериментов находятся в репозитории в файле README.md

Так же в нем указаны скрипты для запуска эксперимента и просмотра результатов. Так для запуска эксперимента в подготовленной среде достаточно ввести команду make train

ВАЖНО. Для работы необходимо не менее 24 ГБТ видеопамяти

6.3 Базовые подходы

[описание некоторого минимально сложного подхода, который давал бы минимально приемлемый по требованиям результат]

7 Результаты

[что получилось в итоге, графики и т.д.]

Здесь (ссылка) можете ознакомиться с полным циклом эксперимента и полученными результатами В результате мы получили сеть, которая в 1.5 раза меньше по количеству параметров работает на 40% быстрее имеет в два раза меньшую вычислительную сложность по среднему проценту качества на всех задачах отличается от оригинальной сети всего на 0.5% на задаче NLI показывает качество выше оригинальной сети

8 Интерпретация

[интерпретация результатов, т.е. как трактовать графики и др.]

9 Заключение

[краткое summary проделанной работы] Данная работа показывает, что можно эффективно уменьшать и ускорять сети. Текущий результат был получен в рамках ограниченности ресурсов как временных, так и аппаратных. Однако он вполне сопоставим с результатами известных работ, описанных в п.3. Более того, ряд техник не были применены из-за

ограниченности по времени. Авторы данной работы будут продолжать этот эксперимент и нацелены уменьшить исходную модель в 10 раз, что более чем согласуется с теорией лотерейных билетов (ссылка)

10 Ссылки

[ссылки на использованную литературу]