ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Московский институт электроники и математики
Им. А.Н.Тихонова НИУ ВШЭ

Департамент компьютерной инженерии

Москва 2021 г.

Содержание

1 Введение

Целью данной работы является исследование и измерение качества аффинных преобразований для модели BERT.

Для достижения поставленной цели необходимо решить следующие задачи:

- Исследование моделей векторного представления слов;
- Исследование методов оценки аффинных преобразований;
- Постановка задачи пропорциональной аналогии в терминах аффинного преобразования;
- Разработка метода оценки точности параллельного переноса для контекстуализированных моделей;
- Подготовка экспериментальных данных;
- Проведение экспериментов;
- Оценка полученных результатов.

Исследование проводится на языке python в среде Jupyter Notebok при использовании Google Colab. Jupyter Notebok является наиболее удобной платформой для проведения исследований на python. Google Colab является бесплатной и мощной платформой для запуска кода. При этом дается 12Гб оперативной памяти, доступ к Google диску для доступа к данным, а также есть возможность запускать код с использованием GPU.

2 Содержательная часть

2.1 Описание профессиональных задач студента

Исследование моделей векторного представления слов;

Исследование методов оценки афинных преобразований;

Постановка задачи пропорциональной аналогии в терминах аффинного преобразования;

Разработка метода оценки точности параллельного переноса для контекстуализированных моделей;

Подготовка экспериментальных данных;

Проведение экспериментов;

Оценка полученных результатов.

2.2 Описание выполнения пунктов

2.2.1 Исследование моделей векторного представления слов

1

2.2.2 Исследование методов оценки аффинных преобразований

2

2.2.3 Постановка задачи пропорциональной аналогии в терминах аффинного преобразования

3

2.2.4 Разработка метода оценки точности параллельного переноса для контекстуализированных моделей

4

2.2.5 Подготовка экспериментальных данных

Для получения эмбеддингов слов были взяты тексты из электронной библиотеки КиберЛенинка. Тексты двух жанров: литература и политика.

Для оценки качества аффинных преобразований используется датасет $Googleanalogy$ Данный датасет был переведен на русский язык с сохранением семантических отношений между словами. Не все слова из данного датасета есть в словаре BERT, поэтому часть отношений пришлось убрать.

2.2.6 Проведение экспериментов

Для проведения экспериментов используется язык программирования python в среде Jupyter Notebok с использованием Google Colab. Jupyter Notebok является наиболее удобной платформой для проведения исследований на python. Google Colab является бесплатной и мощной платформой для запуска кода. При этом дается 12Гб оперативной памяти, доступ к Google диску для доступа к данным, а также есть возможность запускать код с использованием GPU. В качестве фреймворка для работы с моделью BERT был выбран pytorch, так как это современная и гибкая библиотека для работы с глубинным обучением.

Для проведения экспериментов необходимо подготовить данные для их обработки в модели BERT. Сначала весь текст разбивается на отдельные предложения, далее происходит их токенизация и индексация. На этом этапе обработанные предложения по-одному отправляются в модель BERT. Данным способом обработаны по 1 миллиону предложений для каждого жанра.

Полученные после обработки объекты представляют из себя четырёхразмерные тензоры, где оси отражают следующую информацию (в скобках представлено количество элементов):

1. Номер слоя (13 слоев);
2. Номер батча (1 предложение);
3. Количество слов/токенов в предложении (количество токенов в предложении);
4. Векторное представление (768 свойств).

По оси слоев первый слой - это эмбеддинг, поступающий на вход модели, остальные 12 слоев отображают выходы 12 энкодеров. Номер батча в нашем случае не важен, так как используется только одно предложение. Следующая ось отображает токены в предложении с сохранением порядка. Последняя ось отвечает за векторное представление каждого токена.

Получить итоговое векторное представление для токена можно несколькими способами (рисунок 1). В нашем случае используется способ с суммированием последних четырех слоев, данный способ показывает хорошее качество. Способ с конкатенацией последних четырех не используется, так как он требует в 4 раза больше ресурсов.

Описанным ранее методом обрабатываются все подготовленные предложения. Обработка происходит пачками по 10 тысяч предложений. Векторные представления токенов каждой пачки сохраняются на Google диск. Сделано это из-за ограничений оперативной памяти устройства.

После того как получены векторные представления для всего текста, считаются средние эмбеддинги для всех токенов. Из-за ограничений оперативной памяти нельзя посчитать сразу все векторные представления, поэтому они считаются порциями с сохранением промежуточных результатов.

Далее проверялось семантические отношения полученных эмбеддингов на переведенном датасете $Googleanalogytestset$.

2.2.7 Оценка полученных результатов

На рисунках 2 и 3 представлены результаты

3 Заключение

4 Приложения

С кодом можно ознакомиться по ссылке:
https://github.com/andrsolo21/hse_Af_Tr_BERT.

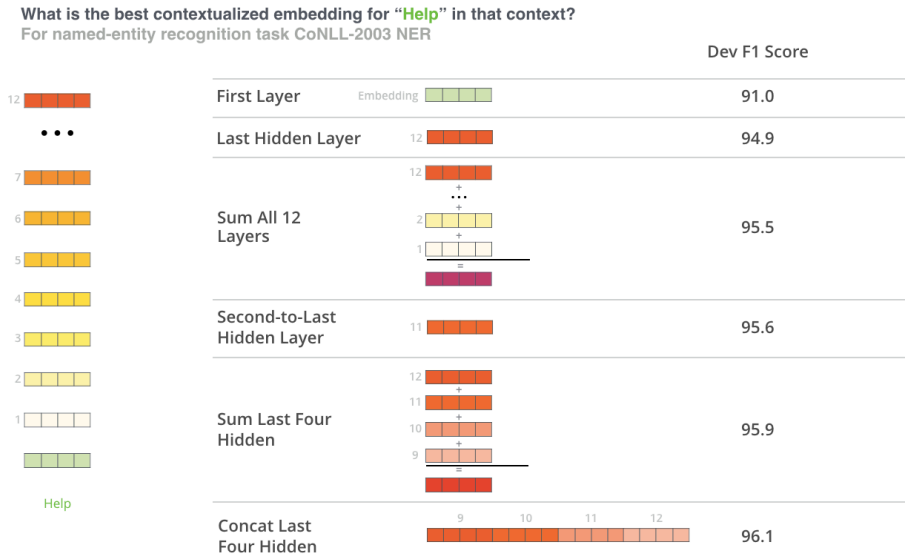

Рис. 1: Возможные варианты получения векторного представления

	result		cosine			count		
metric	3CosAvg	3CosMul	3CosAvg	3CosMul	3CosAdd	3CosAvg	3CosMul	3CosAdd
type								
Adjective adverb	0.222222	0.222222	0.857969	0.930546	1.000000	306	306	306
Capital-countries	0.666667	0.666667	0.832437	0.832437	0.921754	3	3	3
Comparative	0.206522	0.215580	0.845331	0.907858	1.000000	552	552	552
Family	0.659091	0.643939	0.917592	0.937557	1.000000	132	132	132
Nationality adjective	0.333333	0.333333	0.899240	0.968291	1.000000	6	6	6
Opposite	0.000000	0.000000	0.793130	0.918517	1.000000	20	20	20

Рис. 2: Результаты тестирования для текста с литературой

	result		cosine			count		
metric	3CosAvg	3CosMul	3CosAvg	3CosMul	3CosAdd	3CosAvg	3CosMul	3CosAdd
type								
Adjective adverb	0.271930	0.269006	0.856142	0.929608	1.000000	342	342	342
Capital-countries	0.000000	0.000000	0.892919	0.892919	0.928382	3	3	3
Comparative	0.266082	0.245614	0.841756	0.896573	1.000000	342	342	342
Family	0.309524	0.309524	0.888865	0.923934	1.000000	42	42	42
Nationality adjective	0.333333	0.333333	0.906793	0.975569	1.000000	12	12	12
Opposite	0.000000	0.000000	0.792315	0.894300	1.000000	2	2	2

Рис. 3: Результаты тестирования для текста с политикой