NLP Course Template

Владислав Шиманский

May 2023

Abstract

В этой работе представлен метод перевода с английского на русский язык за счет эффективного переноса обучения из предварительно обученный модели типа трансформер. Ключевый является то, что модель типа T5 чаще используется для других задач, таких, как. Please provide a link to your project code right here: `https://github.com/emervlad/nlp_pr`.

1 Introduction

Модель T5 была представлена в статье Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer за авторством Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu.

1.1 Team

Владислав Шиманский

2 Model Description

В последнее время появилось множество техник трансферного обучения для НЛП. Но некоторые методы могут работать почти одинаково — просто с разными наборами данных или оптимизаторами — но они достигают разных результатов, тогда можем ли мы сказать, что метод с лучшими результатами лучше, чем другой? Учитывая текущий ландшафт трансферного обучения для НЛП, Преобразователь преобразования текста в текст (T5) направлен на изучение того, что работает лучше всего, и как далеко мы можем продвинуть уже имеющиеся инструменты. В эту структуру включены многие задачи: машинный перевод, задача классификации, задача регрессии (например, предсказать, насколько похожи два предложения, оценка сходства находится в диапазоне от 1 до 5), другие последовательности для последовательности задач, таких как резюмирование документа (например, обобщающие статьи из ежедневного почтового свода CNN). Структура режима - это просто

стандартный вид ванильного преобразователя кодера-декодера. T5 использует обычный текст, извлеченный из сети. Авторы применяют довольно простую эвристическую фильтрацию. T5 удаляет все строки, которые не заканчиваются конечным знаком препинания. Он также удаляет строку со словом javascript и все страницы с фигурной скобкой (поскольку она часто встречается в коде). Он дедуплицирует набор данных, беря скользящее окно из 3 фрагментов предложений и дедуплицируя его так, чтобы только один из них отображался в наборе данных. Например, выше 3 страниц последний абзац на средней странице удаляется, так как тот же контент отображается на первой странице. Он заканчивается 750 гигабайтами чистого английского текста. Набор данных общедоступен на tensorlow.text.c4. С фреймворком, архитектурой модели и немаркированным набором данных следующим шагом будет поиск неконтролируемой цели, которая дает модели некоторые способы обучения на немаркированных данных. В исходном тексте некоторые слова выпадают с помощью уникального токена-дозора. Слова выпадают независимо равномерно случайным образом. Модель обучена прогнозировать в основном маркеры-дозорные, чтобы очертить выпавший текст. Таким образом, модель, предварительно обученная на основе преобразователя кодера-декодера размера Bert с целью шумоподавления и набором данных C4, обучила $2^{\mathbf{1}}$ шагов на $2^{\mathbf{3}}$ или 348 токенах с графиком скорости обучения обратного квадратного корня. Задачи тонкой настройки включают GLUE, CNN/DM(CNN/Daily Mail), SQuAD, SuperGLUE и задачи перевода: WMT14 EnDe, WMT14 EnFr и WMT14 EnRo.

Figure 1: T5 Architecture.

3 Dataset

Использовался датасет с сайта https://www.manythings.org/anki/rus-eng.zip, содержащий пары предложений на русском и английском длины примерно до 15 слов. Число пар в датасете 470000. Доля валидационной выборки 0.1. В качестве токенесатора использовался T5Tokenizer с максимальным числом токенов, равным 5000.

	Train	Valid
Pairs	43000	4000
Tokens	5000	5000

Table 1: Statistics of Dataset.

4 Experiments

Предобученная модель использовалась с добавлением префикса "translate from English to Russian"

4.1 Metrics

Была использована BLEU метрика.

4.2 Experiment Setup

Размер батча был равен 64, lr=4e-4 с линейным понижением шага.

5 Results

Было получено явное снижение лосса, как на трейне, так и на валидации. В примерах из кода видно, что модель начала двигаться в правльном направлении. Метрика BLEU достигла показателя в 22 процента.

ододите сво сестра Wait for your sister.

Table 2: Output samples.

6 Conclusion

На примере t5 было показано, чт эта модель потенциально может использзоваться для перевода именно на русский язык при надлежащем подборе параметров.

References