Обработка естественного языка. Лекция 6
BERT и GPT

Михаил Пожидаев

11 октября 2024 г.

BERT

BERT (*Bidirectional Encoder Representations from Transformers*) — двунаправленная многофункциональная модель, построенная как кодирующая часть трансформера. Путём дополнительного слоя позволяет решать широкий круг задач, включая:

- ответы на вопросы (*question answering*;)
- вывод на естественном языке (*NL inference* и др.)

Обучение BERT

Декодера нет!

Поскольку декодер отсутствует, возникает вопрос, как модель обучать. Использовались две задачи, которые можно решать просто на массиве текста:

1. Скрытие и восстановление 15% слов (*masked language modelling*)/
2. Определение следующего предложения (*next sentence prediction*).

Fine tuning

Возможность дообучить модель

BERT производит глубокое кодирование текста с использованием знаний о языке, но как использовать закодированный текст, решает разработчик. Этот подход принято называть «fine tuning».

Новый слой

Предполагается, что модель будет дополнена дополнительным слоем, для которого будет произведена дополнительная процедура обучения с использованием новых обучающих данных (возможно, размеченных разработчиками).

GPT

Generative Pre-trained Transformer — генеративная модель, способная продолжать текст слева направо по заданному началу, разработанная OpenAI.

Никому нельзя!

Подробное описание модели не раскрывается из-за высокого качества её работы, с большим потенциалом применения для выпуска фальшивых новостей в астрономических масштабах.

Особенности GPT

1. Использует кодирование слов в виде пар букв с учётом частотности.
2. Обучается на основе текста, полученного кроулингом доверенных сайтов.
3. Не требует точной донастройки (*fine tuning*), но при необходимости позволяет это сделать для уточнения деталей генерируемого текста.

Архитектура GPT

Отомстим BERT!

GPT-3 представляет собой декодирующую половину модели трансформера, что является противоположностью модели BERT, которая состоит из кодирующей половины.

Увеличим количество слоёв

Разные варианты GPT могут состоять из разного количества слоёв внимания. В самом крупном варианте GPT-3 их количество может достигать 96.

Пример блока

Примерная реализация на PyTorch

```
def forward(self, x):
    x = x + self.attn(self.ln_1(x))
    x = x + self.feedforward(self.ln_2(x))
    return x
```

RuGPT-3.5

Модели от Сбера

1. Модели умеют продолжать тексты на русском и частично на английском языках. Для этого пользователю необходимо сформулировать промпт — фразу, которую модель допишет.
2. Нейросеть обучена на более чем 600 Гб открытых данных: Википедии, художественной литературе, диалогах, программном коде.
3. Нейросеть демонстрирует state-of-the-art возможности для русского языка и умеет продолжать любой текст. Результат, который будет получен с помощью применения модели, не может быть предсказан заранее.

Левый и правый

Разные блоки трансформера

BERT

Представляет собой кодирующую (условно левую) часть оригинального трансформера.

GPT

Представляет собой декодирующую (условно правую) часть оригинального трансформера.

Направленность

Положение предсказываемого слова

BERT

Является двунаправленной, т. е. учитывает влияние текста справа и слева. Таким образом, может предсказать пропущенное слово в любом положении сегмента.

GPT

Является однонаправленной, т. е. учитывает влияние только текста слева. Таким образом, может предсказать только слово в конце текста (но может это делать несколько раз, удлиняя построенный текст).

Возможность дополнения

Как применяется Fine Tuning

BERT

Позволяет производить расширение модели, добавляя новые слои (fine tuning). Таким образом, может применяться для широкого класса задач.

GPT

Предлагает специальный механизм PEFT — Parameter Efficient Fine Tuning, предназначенный для уточнения поведения исходной модели без пересчёта основных весовых коэффициентов.

Размер модели

Большая и ещё больше

BERT

Содержит 345 миллионов параметров и может быть запущена практически на любом домашнем ПК.

RuGPT-3.5

Содержит 13 миллиардов параметров и может быть запущена на домашнем ПК с размером RAM не менее 96G.

Спасибо за внимание!

Всё о курсе: https://marigostra.ru/materials/nlp.html
E-mail: msp@luwrain.org
Канал в Телеграм: https://t.me/MarigostraRu