Attention is all you need

BERT - SOTA 2018

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where

Токенизация

Зачем нам нужны токены?

Предложения –> токены

Любая единица пунктуации считается отдельным сегментом

```
from nltk.tokenize import
wordpunct_tokenize
```

Помогут нам корректно считывать метки сущностей для каждого из токенов в тексте.

После токенизации будем сегментировать на BPE и подавать эти ID токенов в модель.

Пример:

“Я поеду сегодня домой, в Москву."

[“Я”, “поеду”, “сегодня”, “домой”, “,”, “в”, “Москву”, “.”]

Byte pair encoding

Раньше:

Word level segmenttion + one hot encoding = Все слова одинаково не похожи друг на друга

Word level segmentation + word2vec = Что если у меня не все слова в словаре?

Byte pair encoding

Раньше:

Word level segmenttion + one hot encoding = Все слова одинаково не похожи друг на друга

Word level segmentation + word2vec = Что если у меня не все слова в словаре?

Теперь:

Нет OOV (out of vocabulary)!

“Из коробки” учтена совместная встречаемость букв

Примеры:

Мы сделаем что то классное -> “Мы” “сдел” “##аем” “что” “то” “класс” “##ное” - **WordPiece**
Мы_сделаем_что_то_классное -> “Мы” “_с” “дел” “аем” “_что” “_то” “_кла” “сс” “ное” - **SentencePiece**

BPE токенизация

Letter	Count
</w>	18
д	16
о	18
м	15
т	5
ж	3
ь	3

Letter	Count
м</w>	15
о	18
д	16
т	5
ж	3
ь	3

Letter	Count
ом</w>	15
о	3
д	16
т	5
ж	3
ь	3

Letter	Count
дом</w>	10
ом</w>	5
о	3
дь</w>	3
д	3
т	5
ж	3

Letter	Count
дом</w>	10
ом</w>	5
ож	3
дь</w>	3
д	3
т	5

Letter	Count
дом</w>	10
ом</w>	5
ождь</w>	3
д	3
т	5

Исходный словарь:
дом</w>: 10, том</w>: 5, дождь</w>: 3

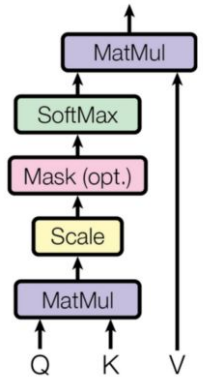
Byte pair encoding - алгоритм сжатия:
Aaabdaaabac –> ZabdZabac -> ZYdZYac -> XdXac

еализации и улучшения:

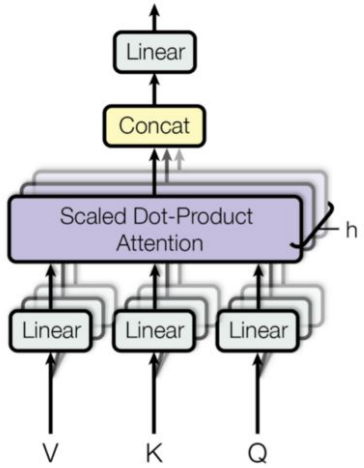
1. BPE
2. WordPiece
3. SentencePiece
4. YouTokenToMe

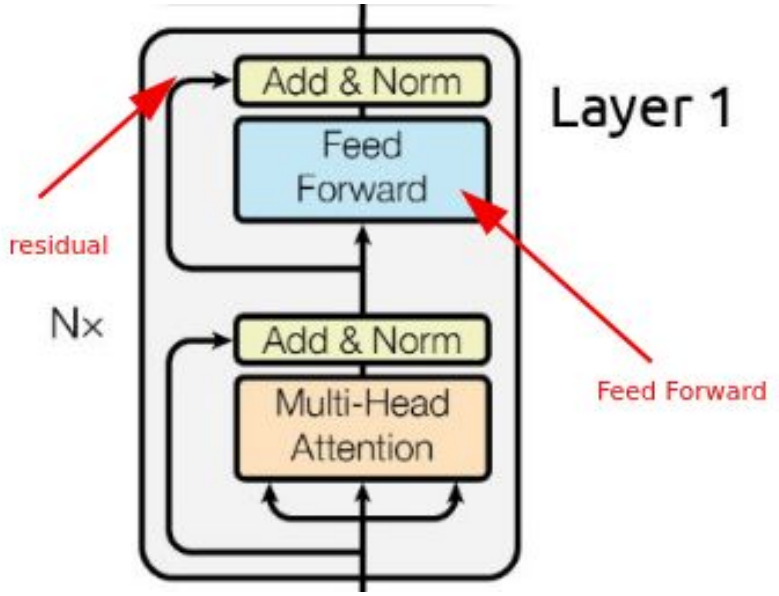
Transformer layer

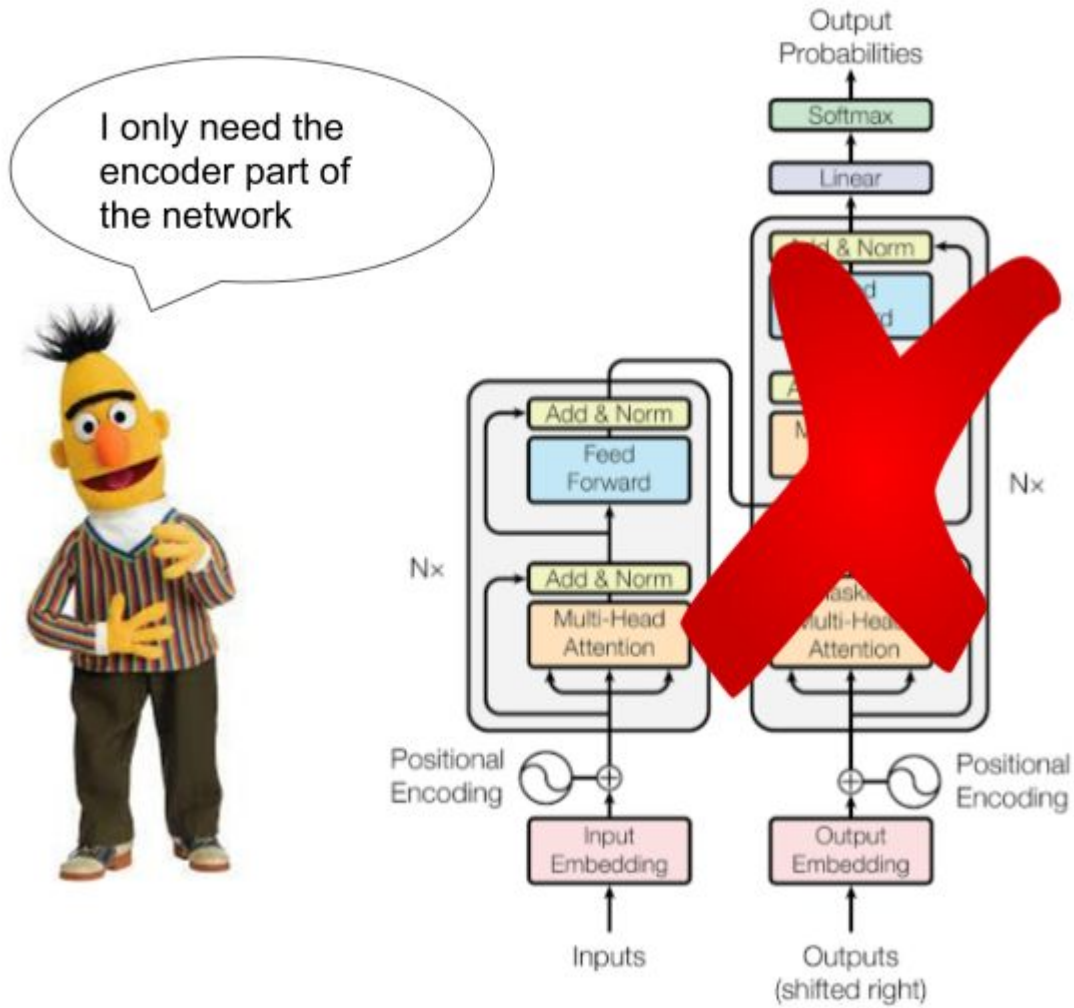
Scaled Dot-Product Attention



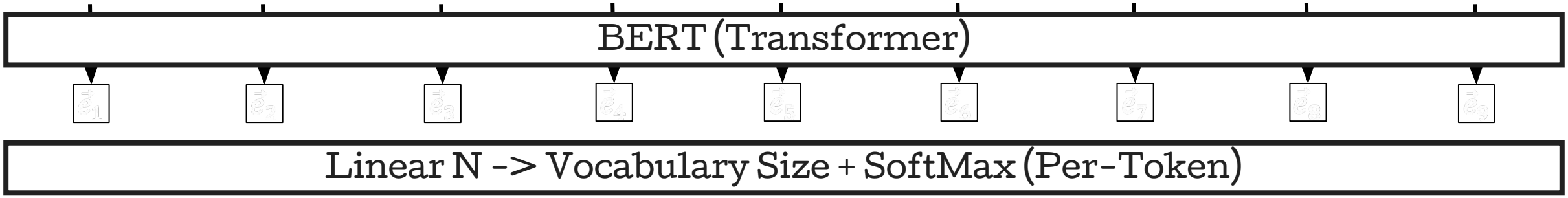
Multi-Head Attention






I only need the encoder part of the network
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Add & Norm
Attention
Nx
Positional Encoding
Output Embedding
Outputs (shifted right)

[BERT] Masked Language Modeling



Предсказывание замаскированных слов

15% слов нужно предсказать, из них:

- 80% заменены на [MASK]
- 10% заменены на случайный токен
- 10% оставлены нетронутыми

Лосс Функция: Кросс-Энтропия

[BERT] Masked Language Modeling

Предсказывание следующего предложения
- 50% действительно следующее предложение
- 50% случайные предложения из корпуса

Обе задачи обучаются одновременно
Это позволяет
- Понимать общий смысл предложений
- Понимать семантику (взаимодействие) слов

Предобучение очень долгое, для предобучения нужно:
- Датасет:
  - BooksCorpus 800M слов
  - English Wikipedia 2,500M слов
- Мощности: 64 TPU
- Время: 4 дня

Лосс Функция: Бинарная Кросс-Энтропия

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Docs | Solutions | Pricing | Log In | Sign Up

Tasks

Image Classification | Translation | Image Segmentation | Fill-Mask | Automatic Speech Recognition | Token Classification | Sentence Similarity | Audio Classification | Question Answering | Summarization | Zero-Shot Classification | + 23 Tasks

Libraries

PyTorch | TensorFlow | JAX | + 32

Datasets

mozilla-foundation/common_voice_7_0 | squad | wikipedia | common_voice | glue | emotion | xtreme | bookcorpus | + 313

Languages Clear All

English × | French | Spanish | German | Chinese | Japanese | Portuguese | Russian | + 200

Licenses

apache-2.0 | mit | afl-3.0 | + 54

Other

AutoTrain Compatible | Eval Results | Has a Space

Models 431 | bert | Sort: Most Downloads

bert-base-uncased
Updated 6 days ago • ↓ 24.4M • ♡ 338

bert-base-multilingual-cased
Updated 6 days ago • ↓ 7.45M • ♡ 67

bert-base-cased
Updated 6 days ago • ↓ 5.93M • ♡ 58

prajjwal1/bert-tiny
Updated Oct 27, 2021 • ↓ 2.39M • ♡ 14

bert-large-cased
Updated May 18, 2021 • ↓ 749k • ♡ 3

dslim/bert-base-NER
Updated Sep 5, 2021 • ↓ 741k • ♡ 86

bert-large-uncased
Updated 8 days ago • ↓ 717k • ♡ 12

finiteautomata/bertweet-base-sentiment-analysis
Updated Jun 23 • ↓ 623k • ♡ 36

bert-large-uncased-whole-word-masking-finetuned-sq…
Updated May 18, 2021 • ↓ 498k • ♡ 31

bert-base-multilingual-uncased
Updated Aug 29 • ↓ 489k • ♡ 24

nlpaueb/legal-bert-base-uncased
Updated Apr 28 • ↓ 263k • ♡ 32

nlptown/bert-base-multilingual-uncased-sentiment
Updated Apr 18 • ↓ 215k • ♡ 74

nlpaueb/legal-bert-small-uncased
Updated Apr 28 • ↓ 205k • ♡ 5

Davlan/bert-base-multilingual-cased-ner-hrl
Updated Jun 25 • ↓ 185k • ♡ 19

dslim/bert-large-NER
Updated Jun 28 • ↓ 174k • ♡ 24

deepset/bert-large-uncased-whole-word-masking-sq…
Updated 4 days ago • ↓ 141k • ♡ 11

emilyalsentzer/Bio_Discharge_Summary_BERT
Updated Feb 27 • ↓ 100k • ♡ 11

prajjwal1/bert-small
Updated Oct 27, 2021 • ↓ 99.8k • ♡ 7