NYU

Transformers

Yilun Kuang

10/16/2023

Outline

- Encoder-Only / Decoder-Only / Encoder-Decoder Architecture
- Huggingface Transformer

Encoder-Only | BERT

Bidirectional Encoder Representations from Transformers (BERT)

Model Architecture

- Transformer Encoder

What is Special About It

- Mask Language Modeling & Next Sentence Prediction
- Downstream task adaptation

NYU

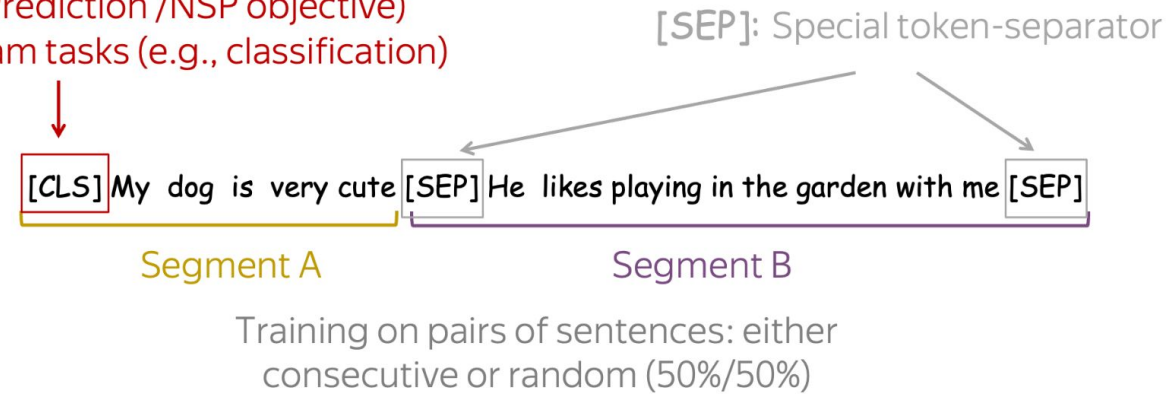
Source:
https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

BERT - Pretrain | Inputs

Training Input: 1) pairs of sentence; 2) [CLS] token; 3) [SEP] token

[CLS]: Special token

- Training time: predict if sentences are consecutive or not (Next Sentence Prediction /NSP objective)
- Test time: downstream tasks (e.g., classification)



NYU

Source: https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

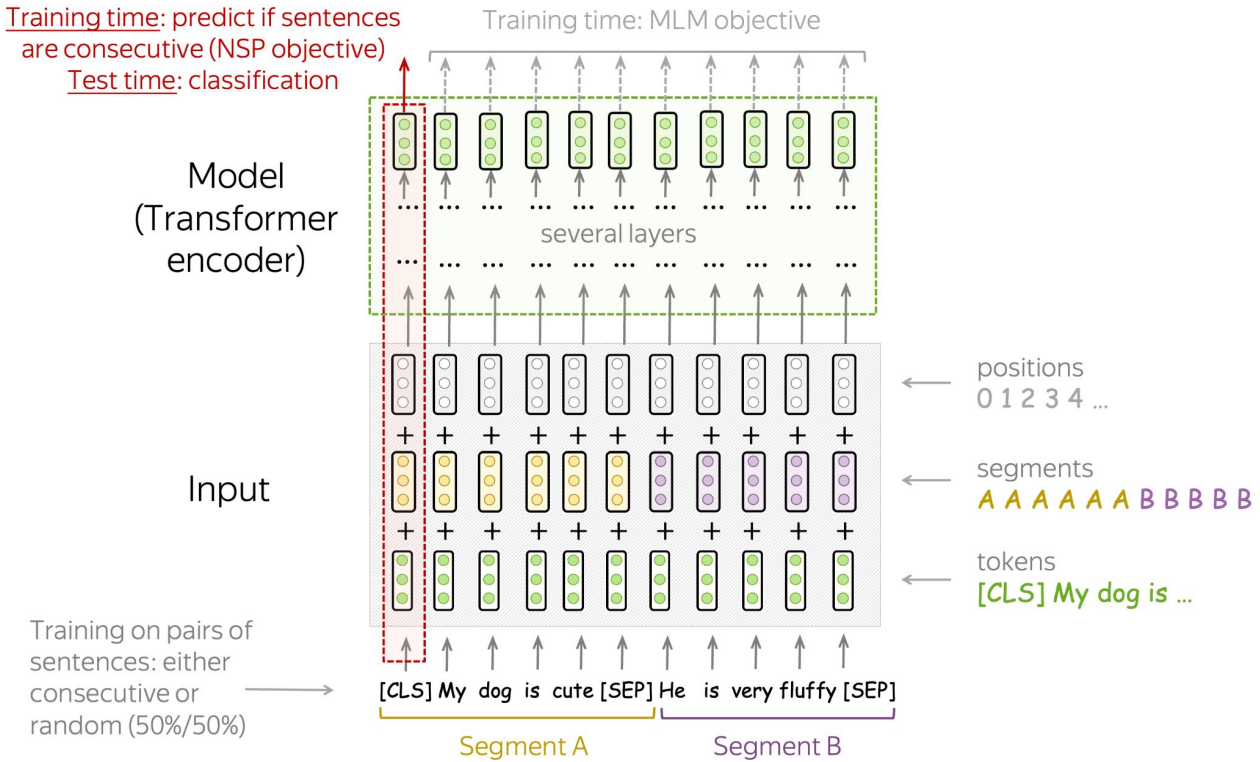
BERT - Pretrain | Inputs

Embedding

1) Token

2) Segment

3) Position



Source: https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

BERT - Pretrain | Objective

Next Sentence Prediction (NSP)

Input: [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label: isNext

Input: [CLS] the man went to [MASK] store [SEP] penguin [MASK] are flight ##less birds [SEP]

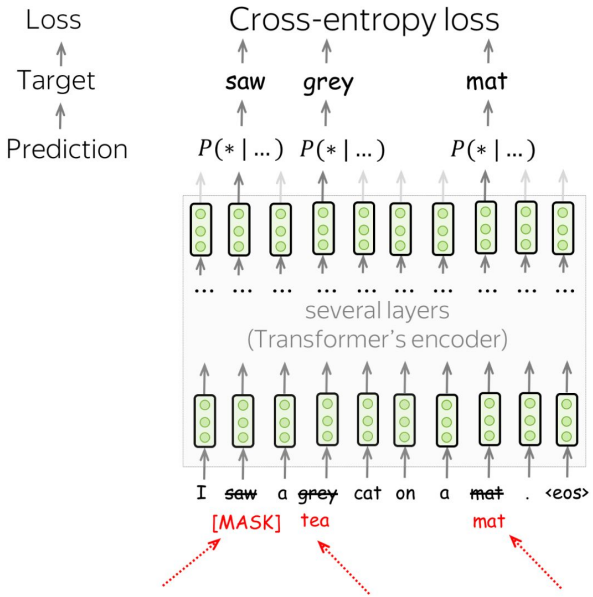
Label: notNext

NYU

Source: https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

BERT - Pretrain | Objective

Masked Language Modeling (MLM)



At each training step:

- pick randomly 15% of tokens
- replace each of the chosen tokens with something
- predict original chosen tokens

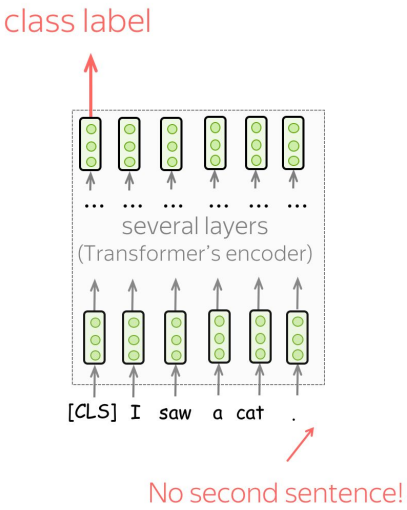
- [MASK], with p = 80%
- Random token, with p = 10%
- Original token, with p = 10%

NYU

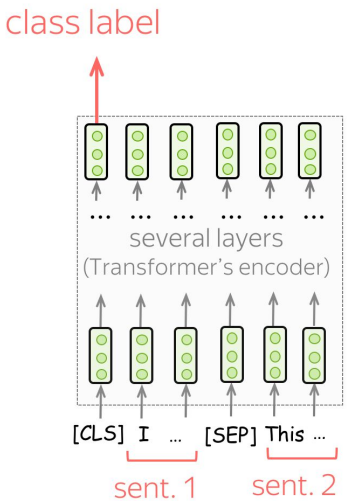
Source: https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

BERT - Finetune | Tasks

Single sentence classification



Sentence Pair Classification

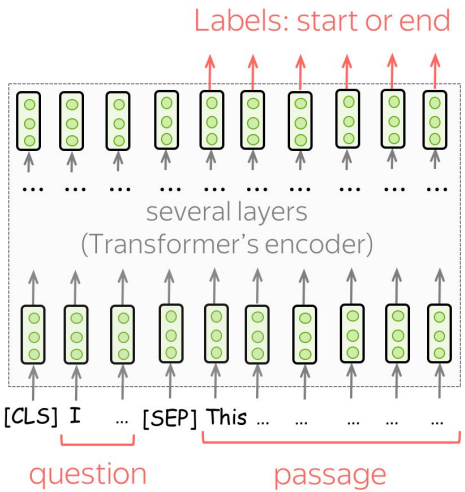


NYU

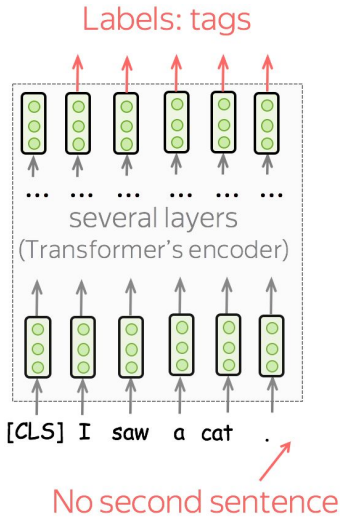
Source:
https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

BERT - Finetune | Tasks

Question Answering



Single sentence tagging



NYU

Source:
https://lena-voita.github.io/nlp_course/transfer_learning.html#bert

Decoder-Only | GPT

Generative Pretrained Transformer (GPT)

Model Architecture

- Transformer Decoder

What is Special About It

- Autoregressive Language Modeling
- Downstream task adaptation

NYU

Source:
https://lena-voita.github.io/nlp_course/transfer_learning.html#gpt

GPT - Pretrain | Inputs

Training Input: 1) sentences; 2) [PAD] / [EOS] token

Example:

My dog is very cute. He likes playing in the garden with me [EOS] [EOS] ... [EOS]

Embedding:

Token embeddings + positional embeddings

NYU

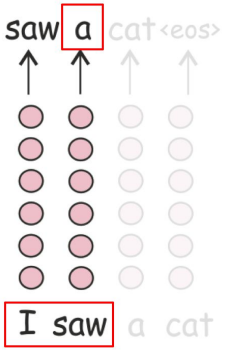
Source:
https://lena-voita.github.io/nlp_course/transfer_learning.html#gpt

GPT - Pretrain | Objective

Autoregressive Language Modeling

Language Modeling

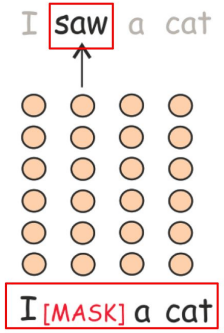
- Target: next token
- Prediction: $P(* \mid \texttt{I saw})$



left-to-right, does not see future

Masked Language Modeling

- Target: current token (the true one)
- Prediction: $P(* \mid \texttt{I [MASK] a cat})$



sees the whole text, but something is corrupted

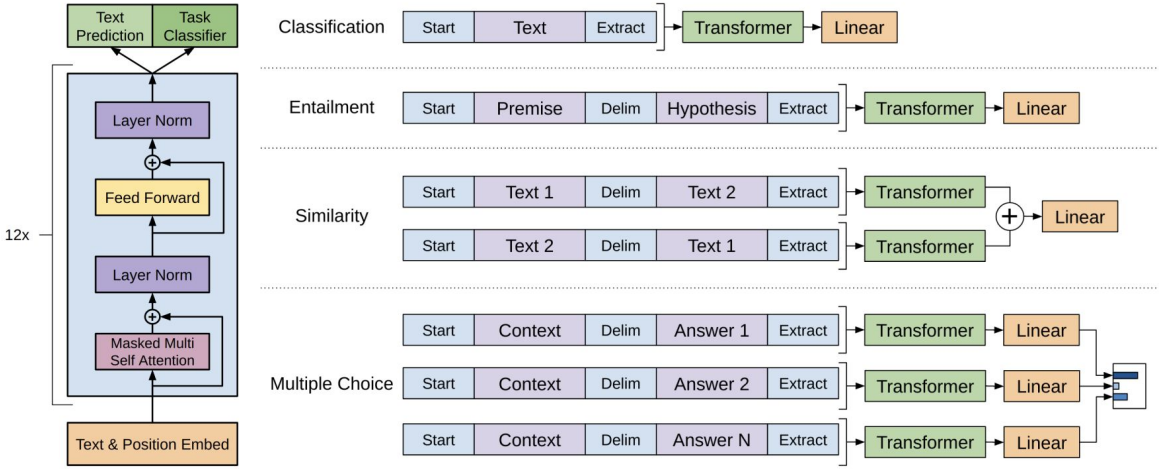
NYU

Source: https://lena-voita.github.io/nlp_course/transfer_learning.html#gpt

GPT - Finetune | NLU Tasks

GPT for Natural Language Understanding (NLU) Tasks

- You can train an additional linear head on top of the final transformer block activation vector, but this is only used in GPT-1.



NYU

Source: Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Encoder-Decoder | T5

Text-to-Text Transfer Transformer (T5)

Model Architecture

- Transformer Encoder & Decoder

What is Special About It

- pretraining on multi-task mixture of unsupervised and supervised tasks (converted to Text-to-Text format)

NYU

Source:
https://arxiv.org/pdf/1910.10683.pdf

Resources

- Encoder-Only (BERT)
  - https://github.com/JonasGeiping/cramming
- Decoder-Only (GPT)
  - https://github.com/karpathy/nanoGPT
- Encoder-Decoder (T5)
  - https://github.com/PiotrNawrot/nanoT5

Summary

- Encoder-Only (BERT)
- Decoder-Only (GPT)
- Encoder-Decoder (T5)

Acknowledgement

This presentation is adapted from Elena (Lena) Voita's NLP Course | For You (https://lena-voita.github.io/nlp_course.html)