6. Attention-based networks

6.4. Large-scale pretraining with transformers

Manel Martínez-Ramón
Meenu Ajith
Aswathy Rajendra Kurup

Introduction

- In the previous lessons we have seen models trained for specific tasks (English to French translation, image classification).
- Pretraining models are becoming common for better generalized models and generalist models.
- Transformers for machine translation consist of an encoder for the representation of input sequences and one for generating target sequences.
- In general, transformers can be used as encoder-decoder, encoder only, and decoder only.

Encoder-only structures

- Remember the self-attention mechanisms that given a sequence $\mathbf{x}_1, \cdots, \mathbf{x}_N$ outputs a sequence $\mathbf{y}_1, \cdots, \mathbf{y}_n$ with the same shape, where

$$\mathbf{y}_n = \sum_{i=1}^{N} \alpha(\mathbf{x}_i, \mathbf{x}_n)\mathbf{x}_n = \boldsymbol{\alpha}_n^\top \mathbf{x}_n \tag{1}$$

- In additive attention we have

$$\begin{aligned} a(\mathbf{x}_i, \mathbf{x}_n) &= \mathbf{w}_f \tanh(\mathbf{W}_q \mathbf{x}_n + \mathbf{W}_k \mathbf{x}_i) \\ \boldsymbol{\alpha}_n &= \text{softmax}\left(a(\mathbf{x}_1, \mathbf{x}_n), \cdots, a(\mathbf{x}_N, \mathbf{x}_n)\right) \end{aligned} \tag{2}$$

- Therefore, for a sequence of N elements, a self attention mechanism produces a matrix $[\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N] \in \mathbb{R}^{N \times N}$ of self attention scores.

Encoder-only structures

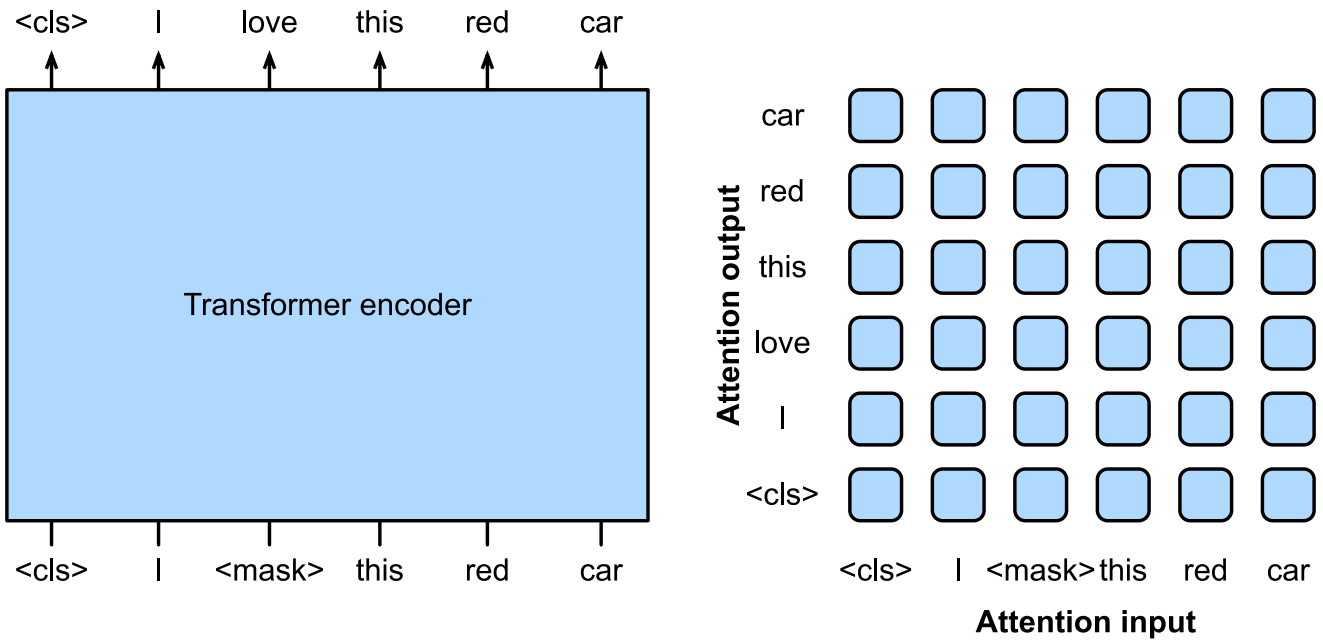
- A multi-head attention mechanism with M heads produces an array of dimensions $N \times N \times M$ attention scores, which are used to construct M representations $\mathbf{h}_{n,m}$ of each element $\mathbf{x}_n$.
- Each representation is then linearly combined with a transformation matrix as

$$\mathbf{y}_n = \mathbf{W}^\top \mathbf{h}_n \tag{3}$$

where $\mathbf{h}_n$ is the concatenation of vectors $\mathbf{h}_{n,m}$.

Encoder-only structures

- An encoder only is a set of self-attention layers with a fully connected layer that produces a classification.



- This is the pre-trained bidirectional Encoder Representation from transformers (BERT).

Encoder-only structures

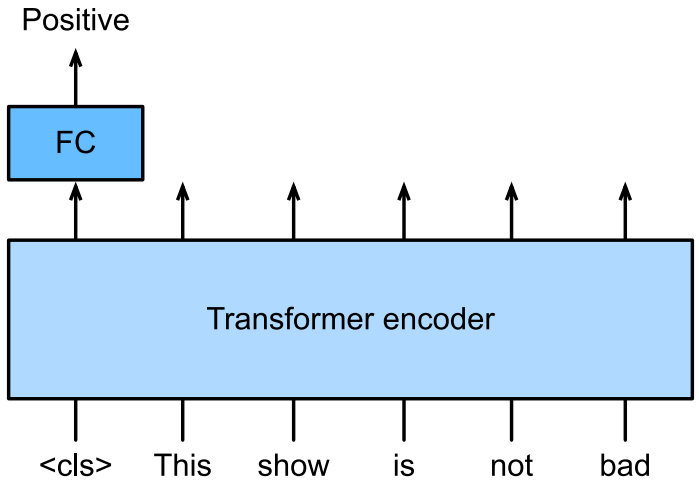
Pretraining BERT

- A BERT is pre-trained by using masked sequences as inputs and using the unmasked sequences as outputs, with the <cls> token included.
- A cross-entropy is used as a criterion.
- It is called bidirectional because the predicted word depends on the previous and posterior tokens.
- This training is unsupervised, as sequences do not have labels. Therefore, large databases (as Wikipedia) can be "easily" used.
- The output token corresponding to the input <cls> is then used to produce a classification after a fine-tuning.

Encoder-only structures

Fine tuning BERT

- The <cls> output is then a global representation of the sequence.
- The BERT can be fine-tuned to perform specific tasks.



- Fully connected layers can be connected to the <cls> to train the whole structure to perform, for example, a sentiment analysis task.

Encoder-only structures

Fine tuning BERT

- Cross entropy and backpropagation is used to train the fully connected layers from scratch while the BERT structure is just updated.
- Based on this idea, and using a 350M parameter structure trained with 250 Billion tokens, BERT has been used for representation spans of text, computer vision, and others.

Encoder-decoder structures

- The encoder-only structure cannot predict text for tasks as machine translation because the output sequence has the same length as the input.
- The encoder-decoder for machine translation first produces a representation of length N.
- The decoder produces a matrix of $N \times m$ attention scores (or equivalently, multi-head scores) to predict the next token.
- This is called cross-attention.

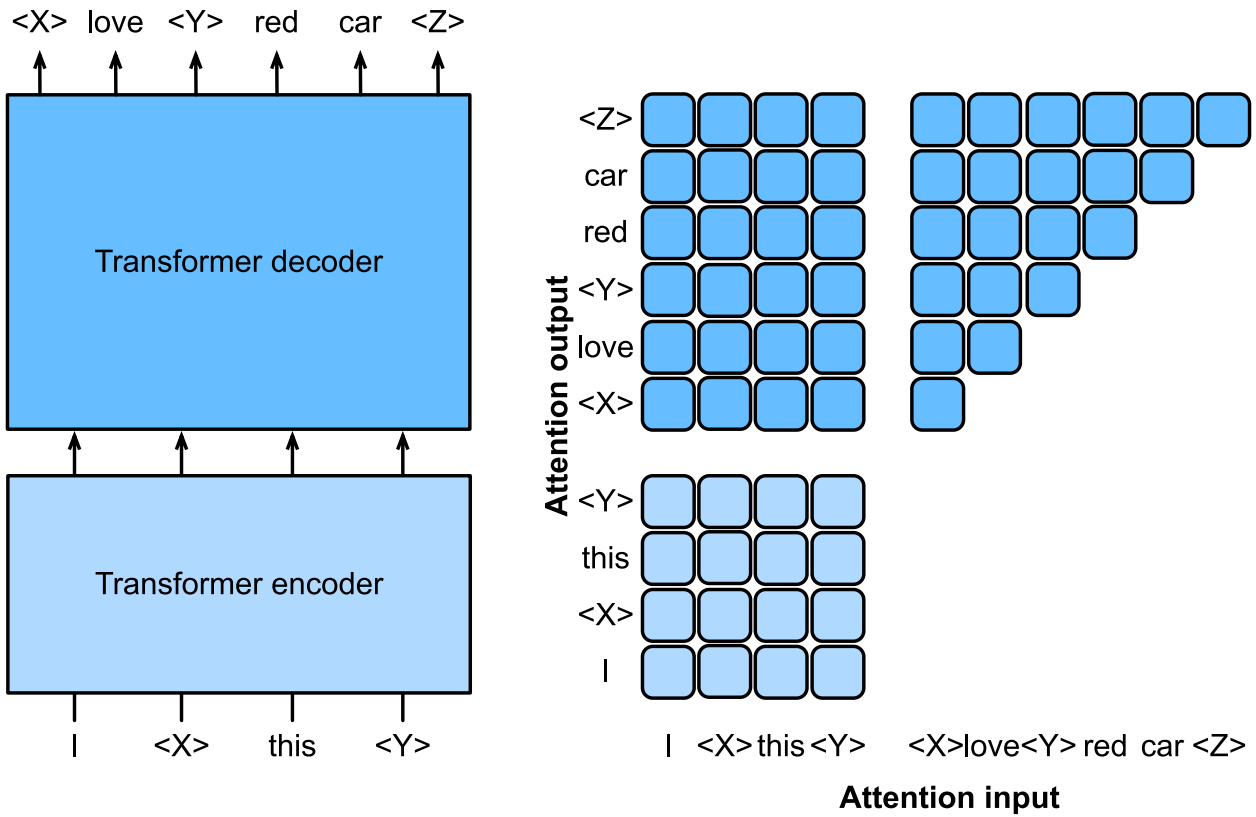
Encoder-decoder structures

- BART and T5 are two encoder-decoder schemes that are pre-trained with large text corpora.
- Both reconstruct text during the pre-training. BART uses deletion, permutation, and rotation besides masking.

Encoder-decoder structures

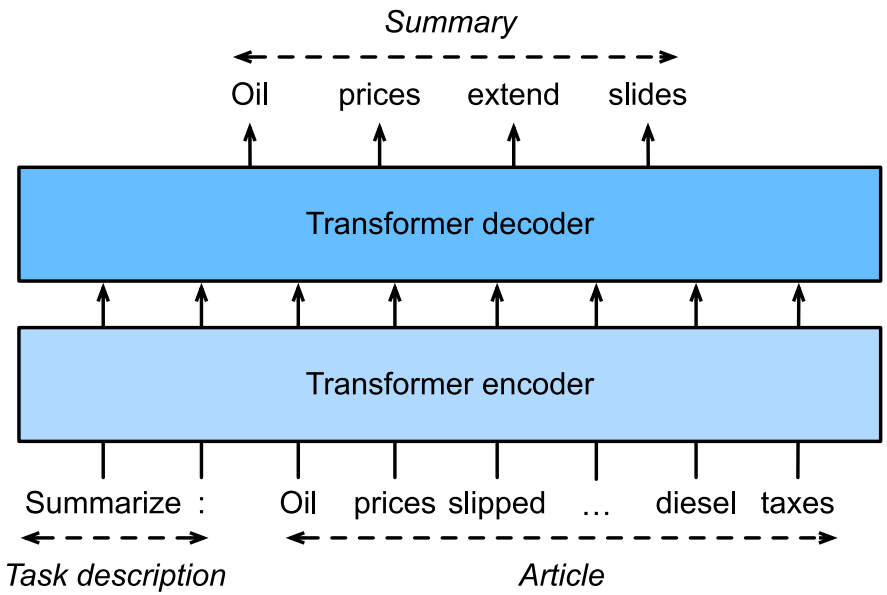
Pretraining T5

- The Text-to-Text Transfer Transformer (T5) is trained to predict consecutive spans with special tokens.



- The decoder self-attention is causal (it attends only to past tokens.)

- T5 can be fine-tuned to perform various specific tasks in the same text-to-text problem.
  - T5 includes the task description at the input.
  - T5 can produce sequences with arbitrary lengths.
  - No additional layers are required.



Encoder-decoder structures

Fine-tuning T5

- The 11-billion-parameter T5 (T5-11B) achieved state-of-the-art results on multiple encoding (e.g., classification) and generation (e.g., summarization) benchmarks.
- The Imagen structure, text is input to a frozen 4.6 Billon parameter T5 encoder (T5-XXL) to produce images.

Teddy bears swimming at the Olympics 400m Butterfly event.

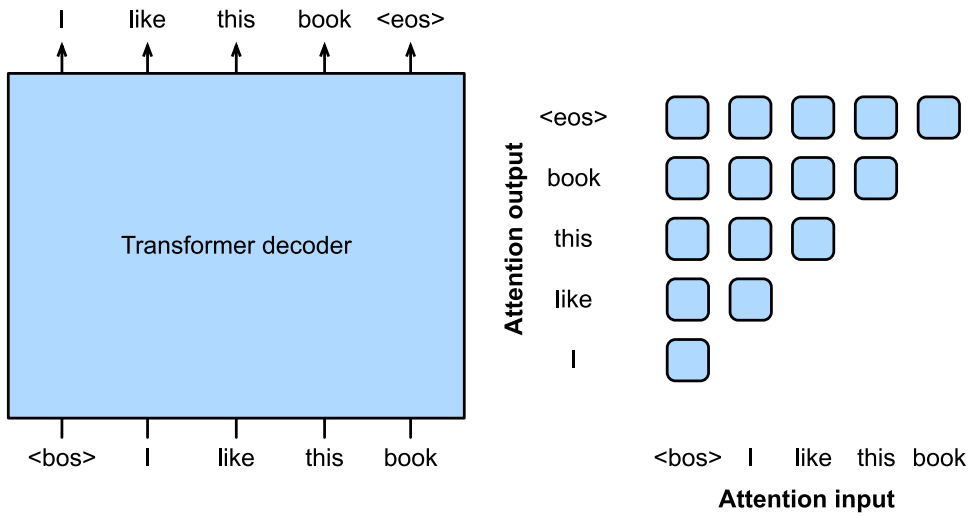
A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Decoder-only structures

GPT and GPT-2

- The decoder-only structures remove the encoder and the encoder-decoder cross attention.
- This allows to train large models with any text databases.
- The Generative Pre-Training model is based on a decoder.



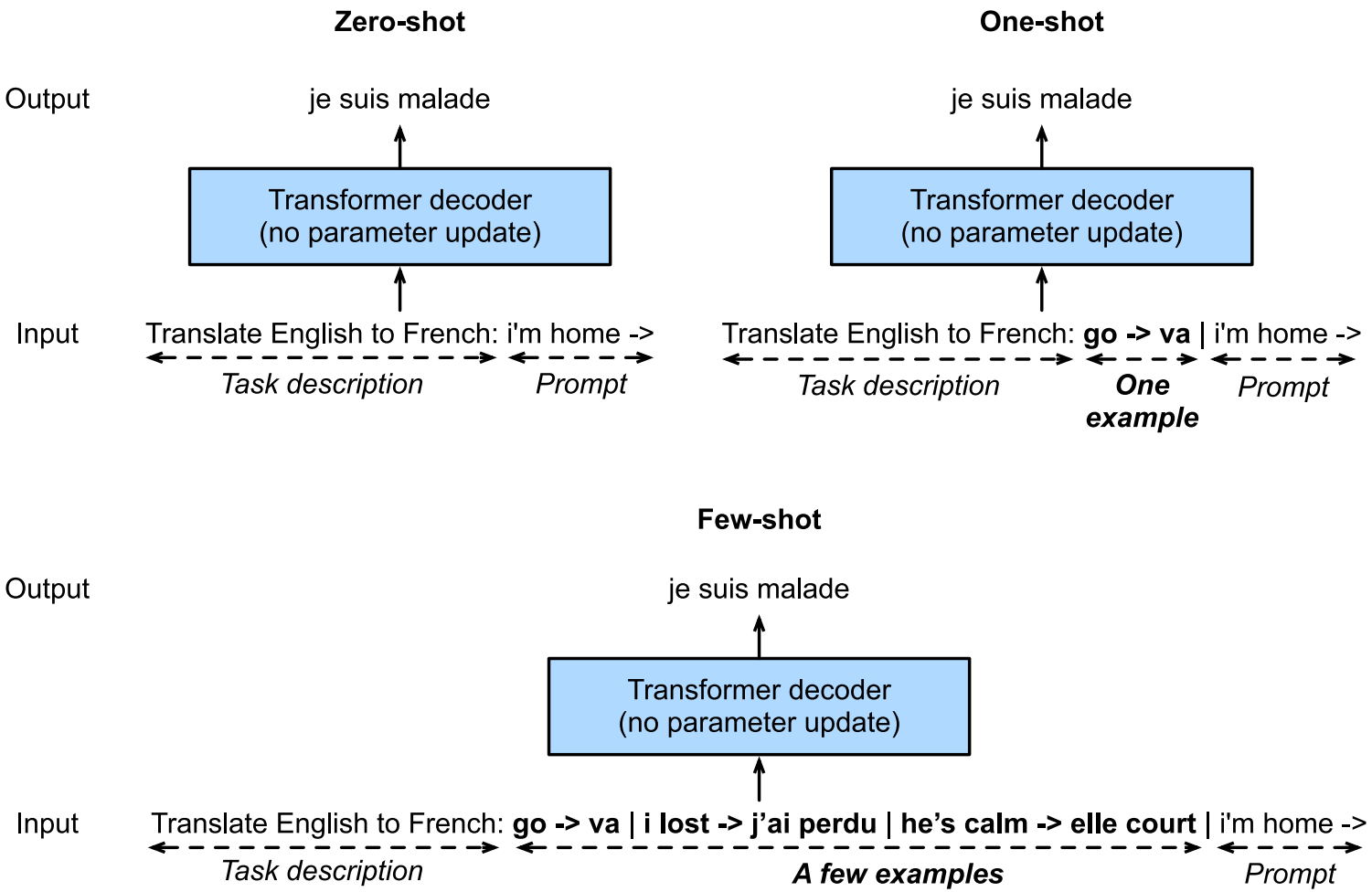
- The GPT is trained with the output as the input sequence shifted.

Decoder-only structures

GPT and GPT-2

- GPT (2018) has 100 million parameters and needs to be fine-tuned for individual downstream tasks.
- GPT-2 (2019) uses pre-normalization and improved initialization and weight-scaling
- GPT-2 was pre-trained on 40 GB of text and had 1.5 billion parameters.
- GPT-2 obtained state-of-the-art results on language modeling benchmarks and promising results on multiple other tasks without updating the parameters or architecture.

GPT3


Zero-shot
Output
je suis malade
Transformer decoder
(no parameter update)
Input
Translate English to French: i'm home ->
Task description
Prompt
One-shot
je suis malade
Transformer decoder
(no parameter update)
Translate English to French: go -> va | i'm home ->
Task description
One example
Prompt
Few-shot
Output
je suis malade
Transformer decoder
(no parameter update)
Input
Translate English to French: go -> va | i lost -> j'ai perdu | he's calm -> elle court | i'm home ->
Task description
A few examples
Prompt