MUNI FI

Transformers

PA154 Language Modeling (10.1)

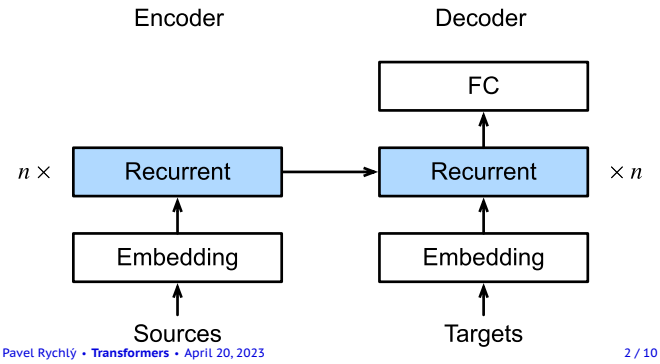
Pavel Rychlý

pary@fi.muni.cz

April 20, 2023

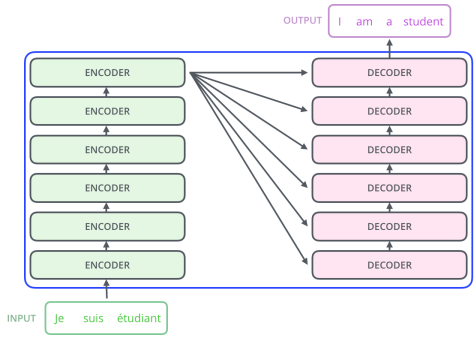
Multi-layer encoder/decoder

- Encoder: Input sequence → state
- Decoder: state + sentence delimiter → output
- Problem: fix size state



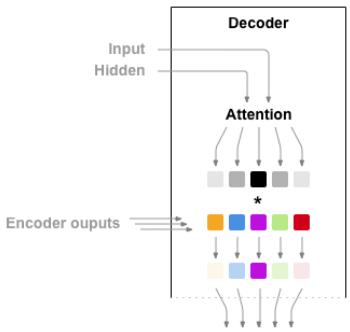
Attention

- each decoder layer has access to all hidden states from the last encoder
- use attention to extract important parts (vector)



Attention

- use attention to extract important parts (vector)
- important = similar to “me”



Self-Attention

- instead of sequencial processing
- attention to previous (and following) tokens
- fully parallel processing during training

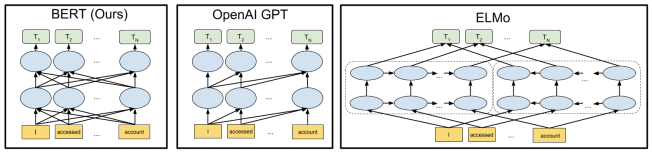
Transformes

- *Attention is All You Need*
- self-attention in both encoder and decoder
- masked cross-attention in decoder

`http://jalammar.github.io/illustrated-transformer/`

Transformers variants

- using context to compute token/sentence/document embedding
- BERT = Bidirectional Encoder Representations from Transformers
- GPT = Generative Pre-trained Transformer
- many varians: tokenization, attention, encoder/decoder connections



BERT

- Google
- encoder only
- pre-training on raw text
- masking tokens, is-next-sentence
- big pre-trained models available
- domain (task) adaptation

Input: `The man went to the [MASK]1 . He bought a [MASK]2 of milk .`
Labels: `[MASK]1 = store; [MASK]2 = gallon`

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Using pre-trained models

- (BERT) trained on huge amount of data
- finetuned on task specific data
- using output of BERT as an input to task specific model (without modification of BERT)

GPT

- Open AI
- decoder only
- pre-training on raw text
- trained on prediction of next token