

Transformers

PA154 Language Modeling (11.1)

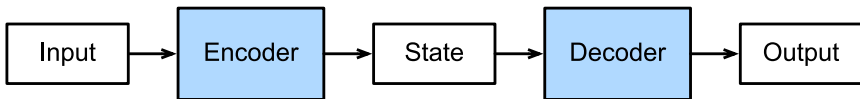
Pavel Rychlý

pary@fi.muni.cz

May 7, 2024

Encoder-Decoder

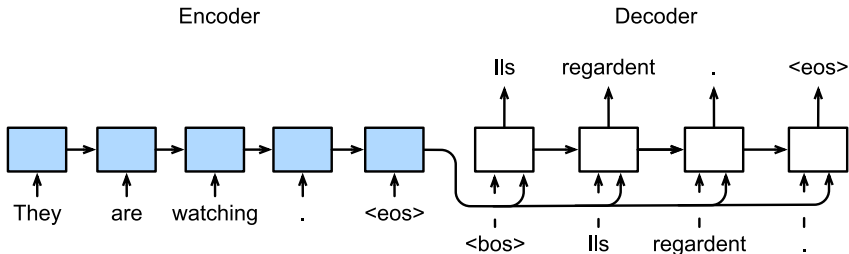
- variable input/output size, not 1-1 mapping
- two components
- Encoder: variable-length sequence \rightarrow fixed size state
- Decoder: fixed size state \rightarrow variable-length sequence



Sequence to Sequence

■ Learning

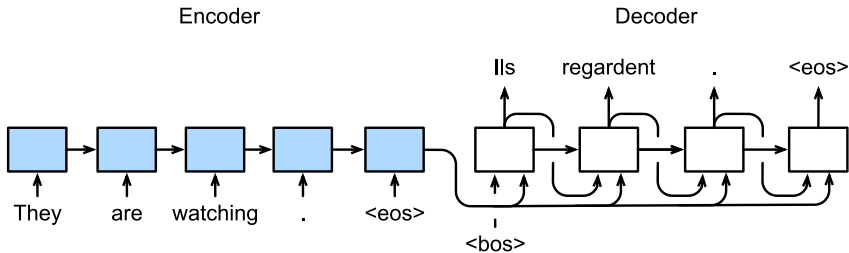
- Encoder: Input sequence \rightarrow state
- Decoder: state + output sequence \rightarrow output sequence



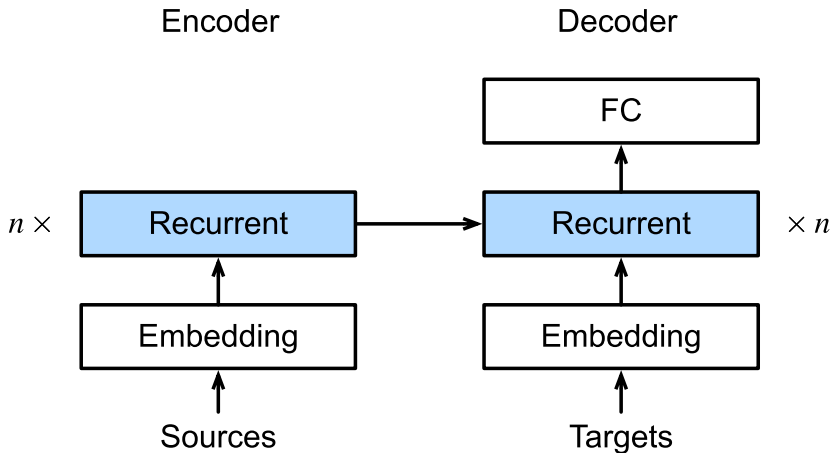
Sequence to Sequence

■ Using

- Encoder: Input sequence \rightarrow state
- Decoder: state + sentence delimiter \rightarrow output

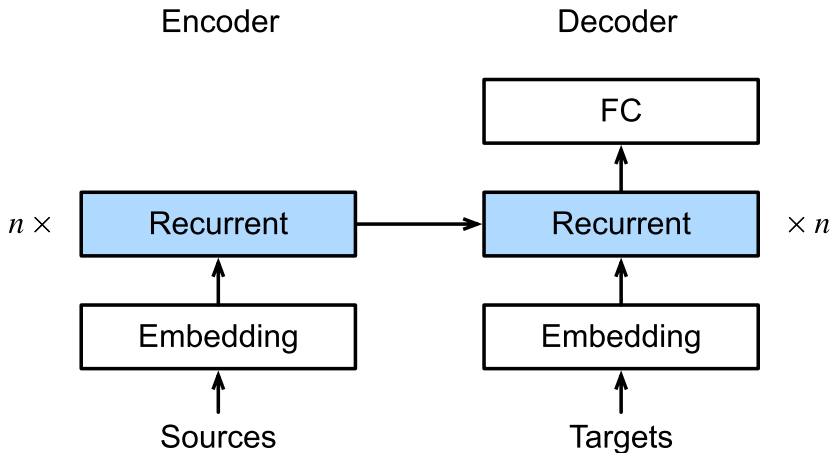


Multi-layer encoder/decoder



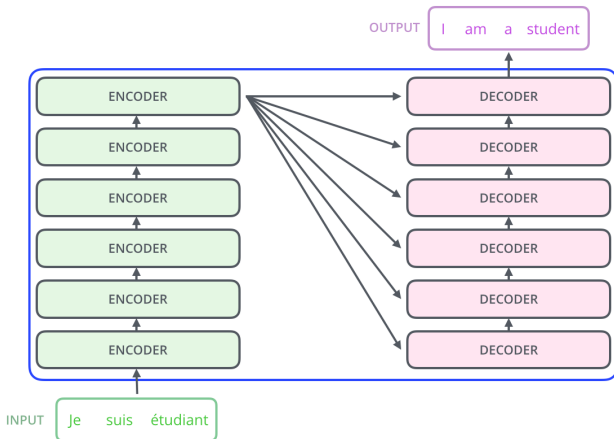
Multi-layer encoder/decoder

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



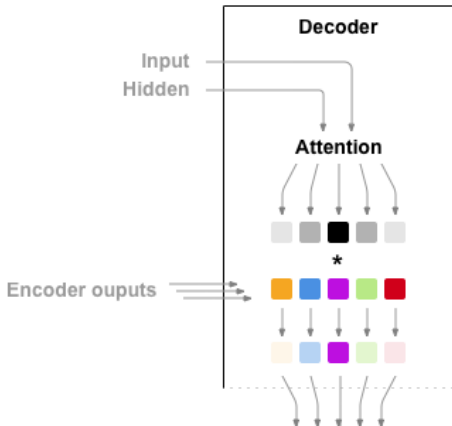
Attention

- 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 sequential processing
- attention to previous (and following) tokens
- fully parallel processing during training

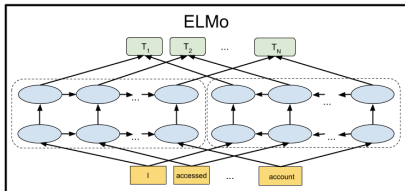
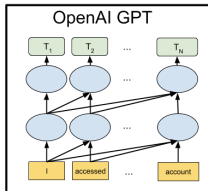
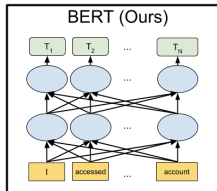
Transformes

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

http:
[//jalammr.github.io/illustrated-transformer/](http://jalammr.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 variants: 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

