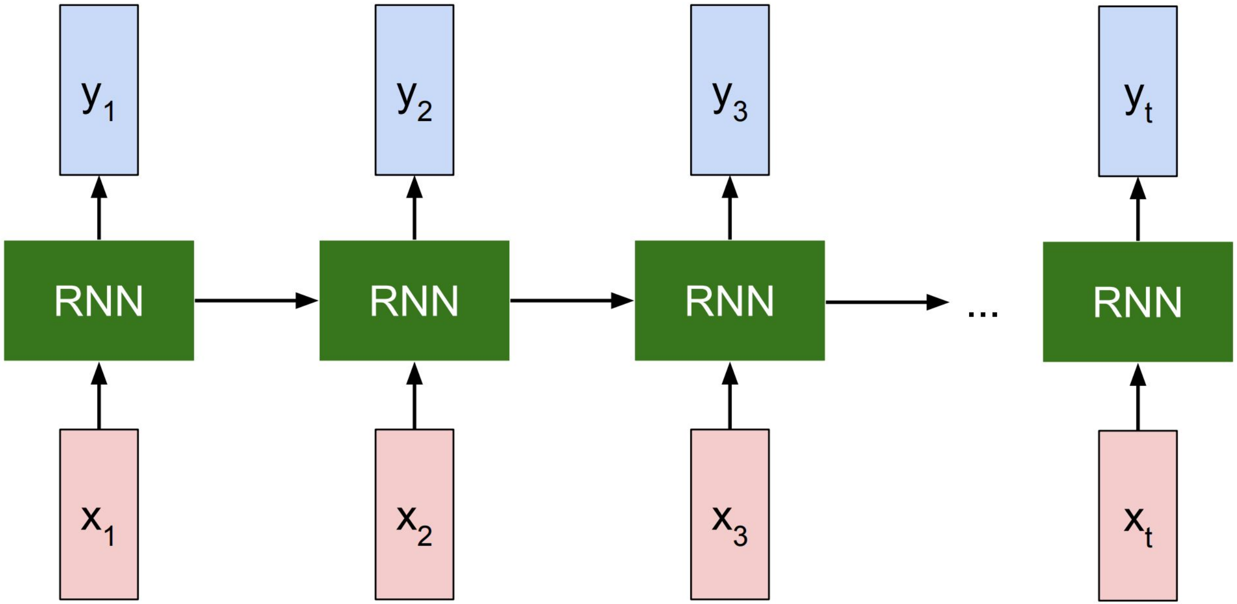
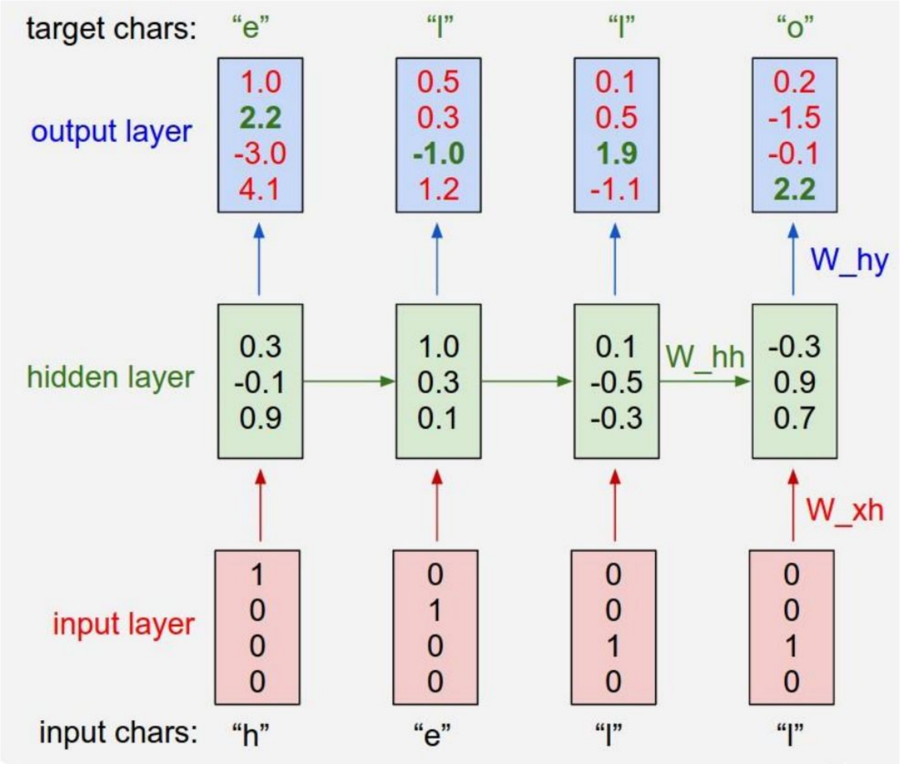
Sequence models

Recurrent Neural Networks (RNN)



https://cs231n.github.io/rnn/

RNN



https://cs231n.github.io/rnn/

RNN Problems

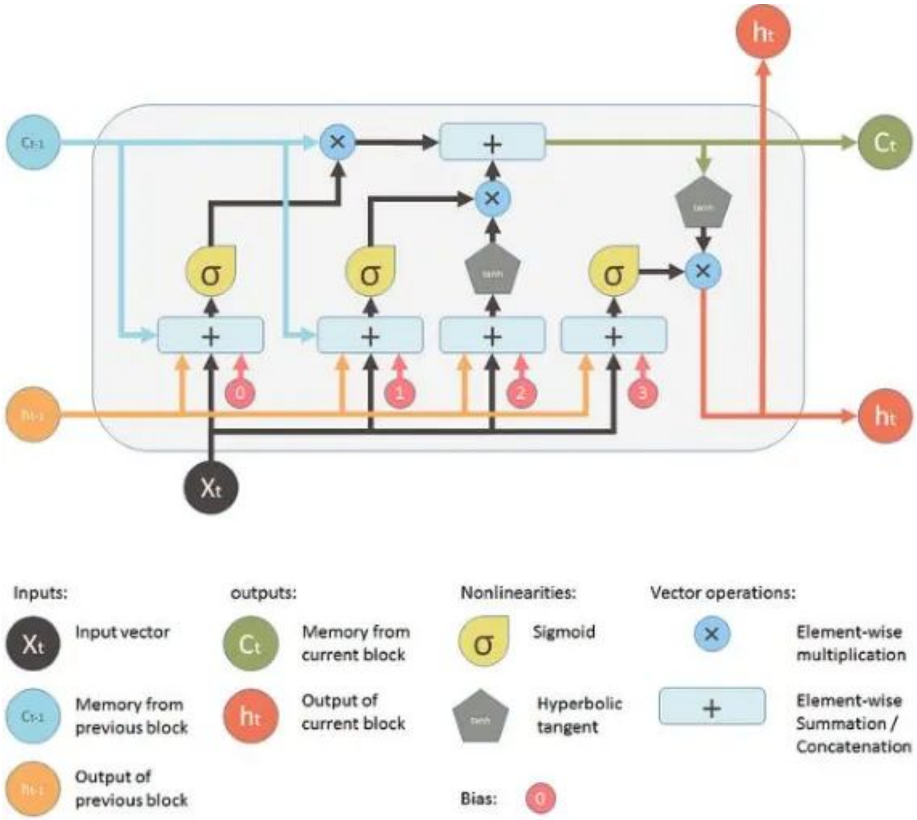
- Vanishing gradient
- Exploding gradient

- Generally forgets stuff that is many iterations ago

LSTM

- Tries to answer the problems of RNNs

- More gates/paths
  - Input
  - Output
  - Forget



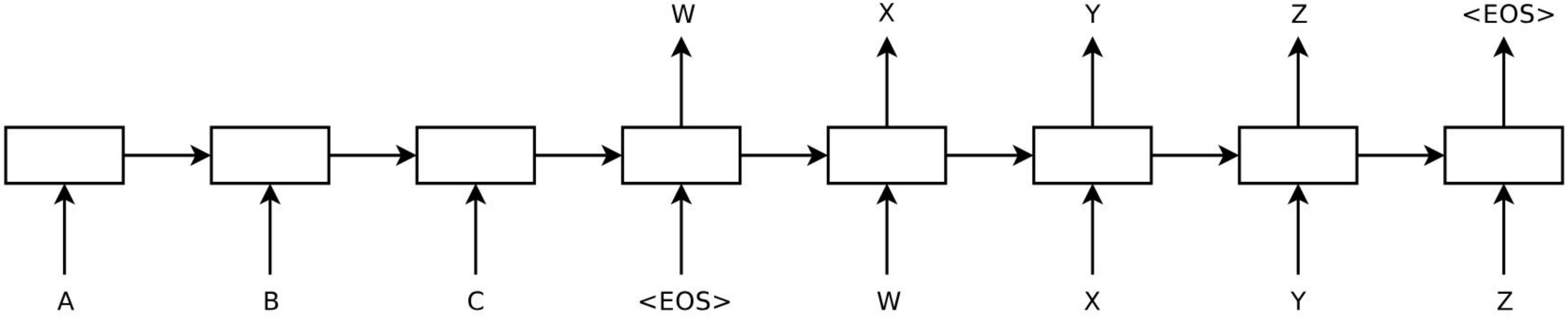
Language as a sequence

- Sequence2Sequence models
- Input as a sequence of tokens / words / char
- Output as a sequence of tokens / words / char

Encode and Decoder

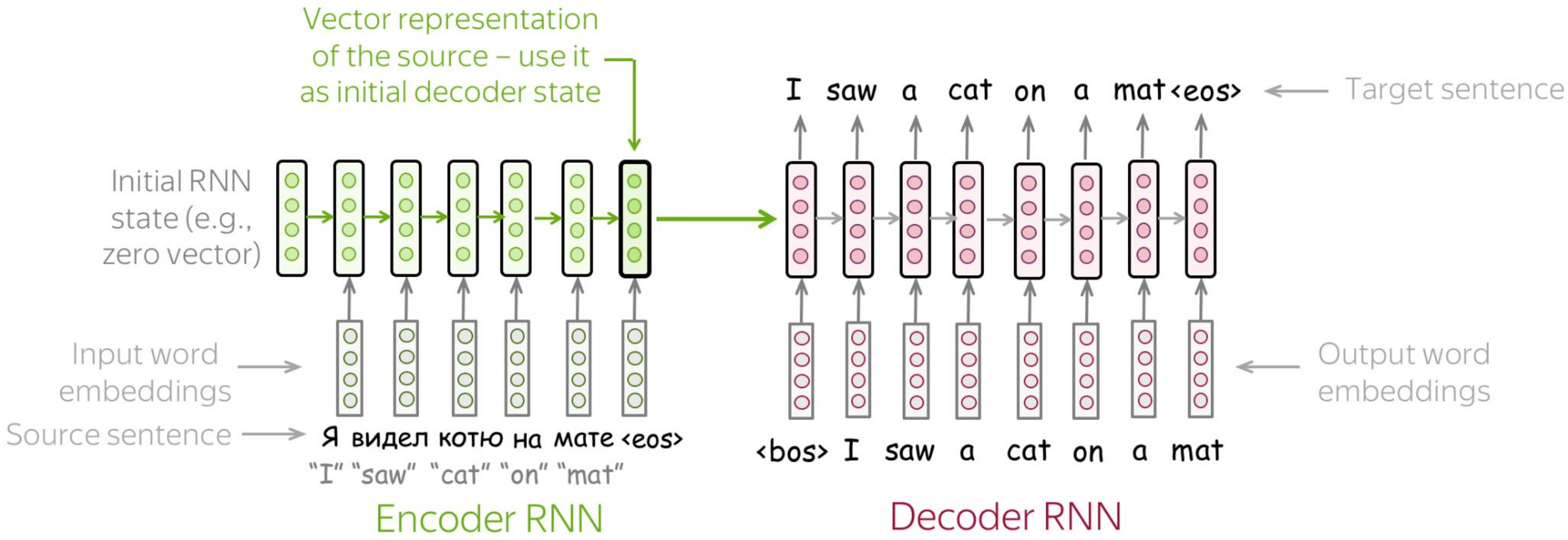
- Long Short Term Memory cells (LSTM)

  Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks (2014)



Encoder and Decoder

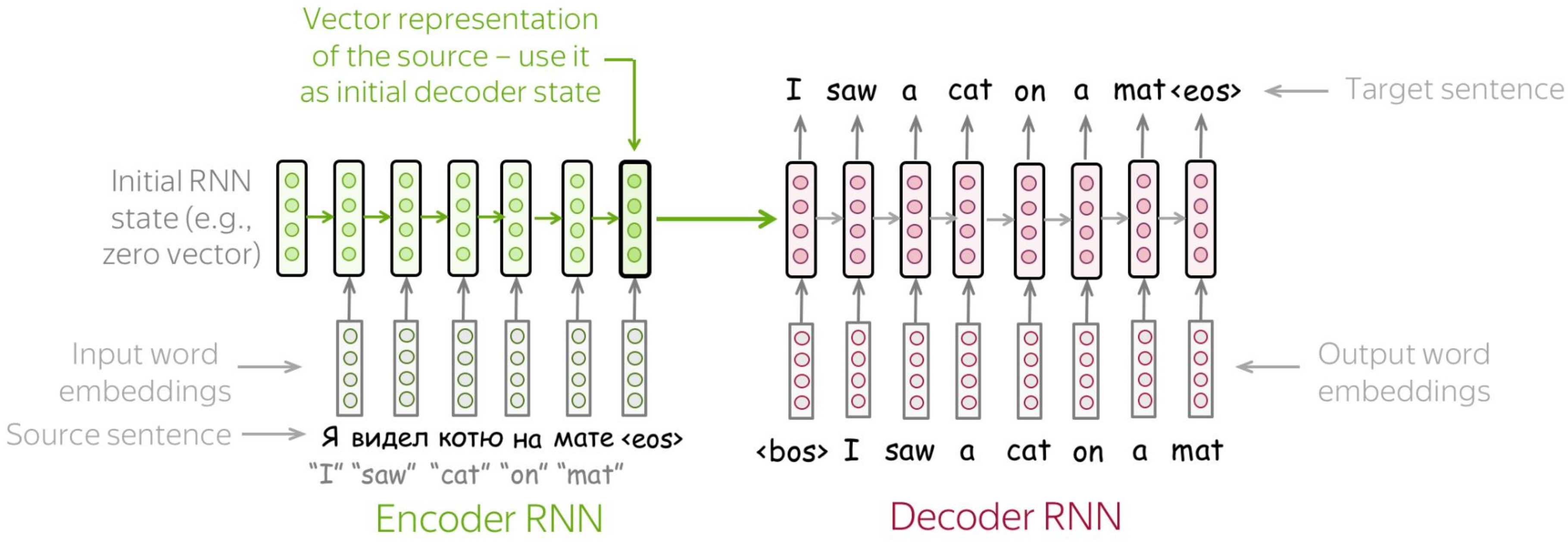
- Encoder uses word embeddings as inputs
- Encoder produces a vector representation of a full sentence (or paragraph)


Vector representation of the source – use it as initial decoder state
I saw a cat on a mat <eos>
Target sentence
Initial RNN state (e.g., zero vector)
Input word embeddings
Output word embeddings
Source sentence
Я видел котю на мате <eos>
"I" "saw" "cat" "on" "mat"
<bos> I saw a cat on a mat
Encoder RNN
Decoder RNN

https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Encoder and Decoder

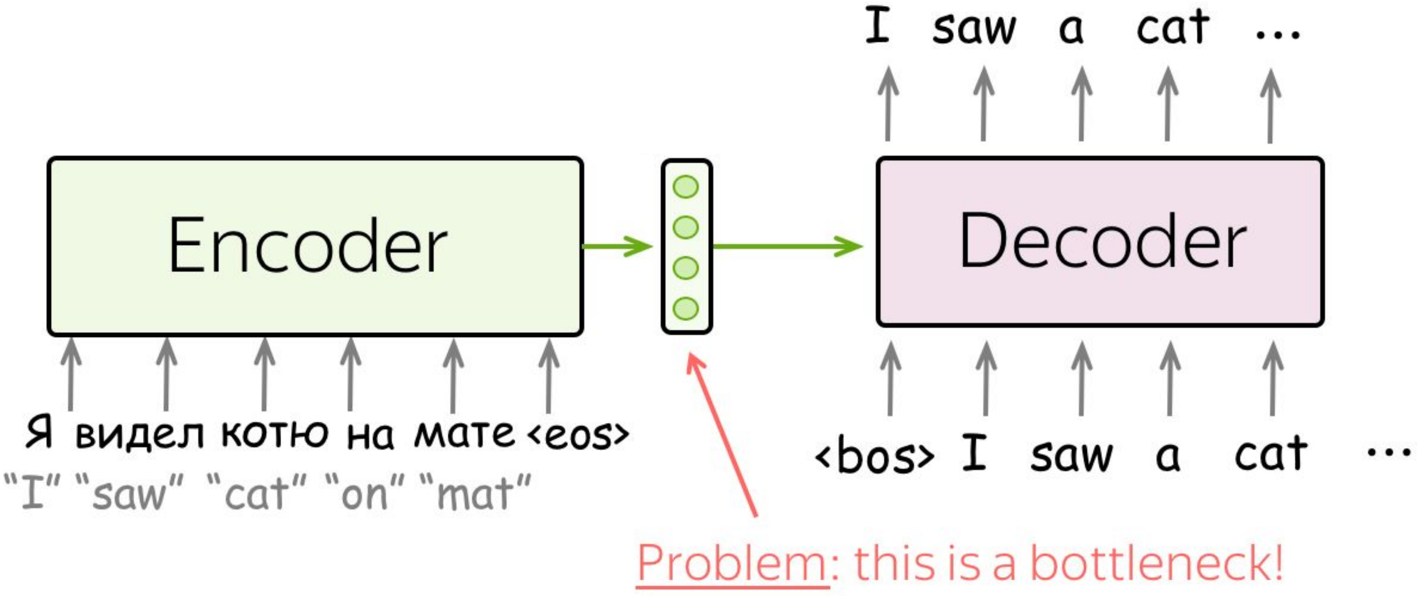
- Decoder uses this as a “starting point”
- Decoder predicts the next word, until it predicts <EOS>



https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Problems

- Input sentence has to be compressed into a single vector.
- Despite use of LSTMs, longer inputs lead to model forgetting earlier stuff.


I saw a cat ...
Encoder
Decoder
Я видел котю на мате <eos>
"I" "saw" "cat" "on" "mat"
<bos> I saw a cat ...
Problem: this is a bottleneck!

https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

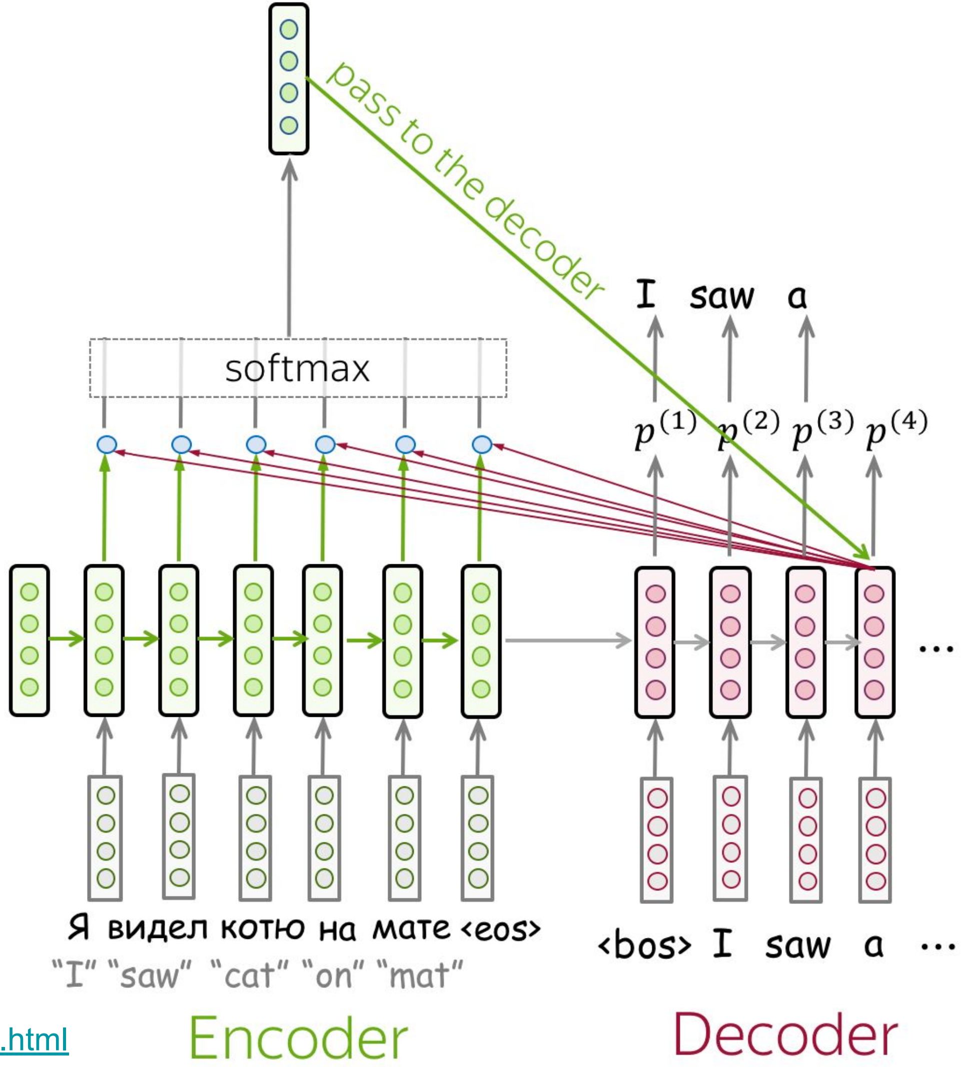
Attention

- Provide representations of each input token to the decoder.
- Model can learn which part of the input is important for predicting the next output.

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. (2014)

Attention

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. (2014)

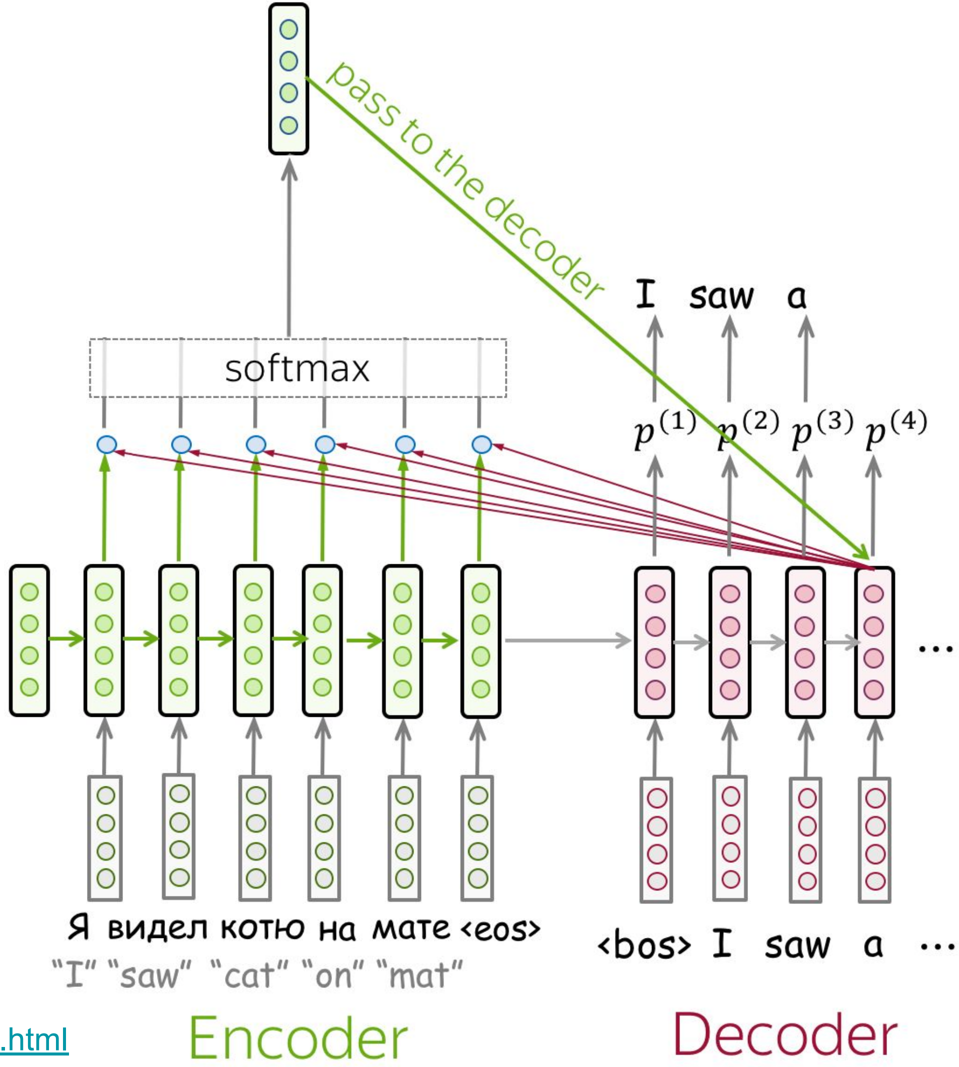


https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention

- Provide representations of each input token to the Decoder.

- Model can learn which part of the input is important for predicting the next output.


pass to the decoder
softmax
I saw a
$p^{(1)}$ $p^{(2)}$ $p^{(3)}$ $p^{(4)}$
...
Я видел котю на мате <eos>
"I" "saw" "cat" "on" "mat"
<bos> I saw a ...
Encoder
Decoder

https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html