Seq2Seq, Attention

Machine translation task

Machine translation task

$$source = (x_1, x_2, \ldots, x_n)$$

The cat sits on the floor

$$target = (y_1, y_2, \ldots, y_m)$$

Кошка сидит на полу

Machine translation task is to find the most probable translation given source:

$$\widehat{target} = \underset{target}{argmax} \; P(target \mid source, \theta)$$

Machine translation task

$$source = (x_1, x_2, \ldots, x_n)$$

The cat sits on the floor

$$target = (y_1, y_2, \ldots, y_m)$$

Кошка сидит на полу

$$\widehat{target} = \underset{target}{argmax} \boxed{P(target \mid source, \theta)}$$

$$P(target \mid source) = P(y_1 \mid source) \cdot P(y_2 \mid y_1,\ source) \ldots P(y_m \mid y_1, \ldots, y_{m-1},\ source)$$

Machine translation task

$$source = (x_1, x_2, \ldots, x_n)$$

The cat sits on the floor

$$target = (y_1, y_2, \ldots, y_m)$$

Кошка сидит на полу

$$P(target \mid source) = P(y_1 \mid source) \cdot P(y_2 \mid y_1,\ source) \ldots P(y_m \mid y_1, \ldots, y_{m-1},\ source)$$

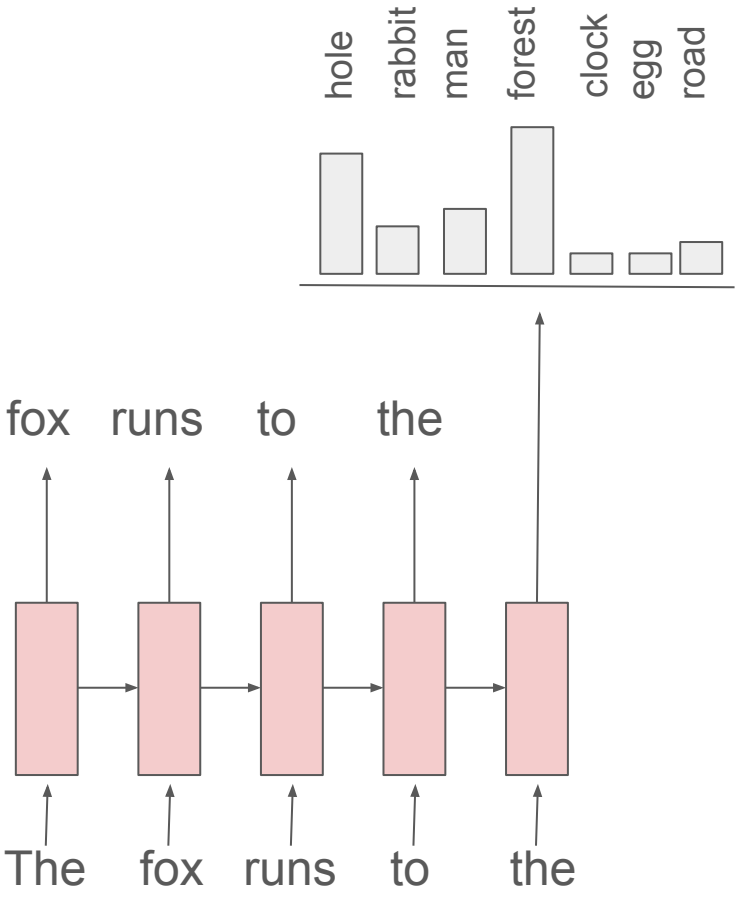
Conditional language model (conditioned on source)

Machine translation techniques

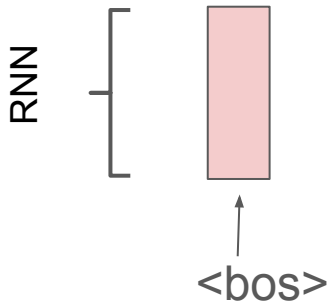
- rule-based (1950th)
- statistics-based
- neural-based (2010th)

Seq2Seq

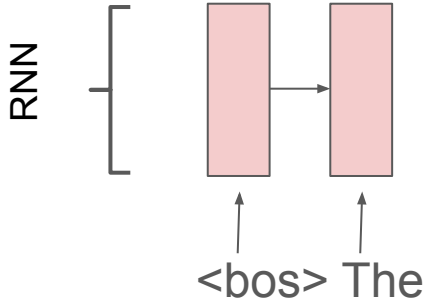
LM (reminder)


hole
rabbit
man
forest
clock
egg
road
fox
runs
to
the
The
fox
runs
to
the

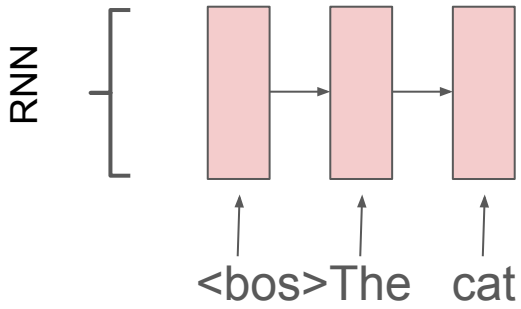
Encoder-Decoder



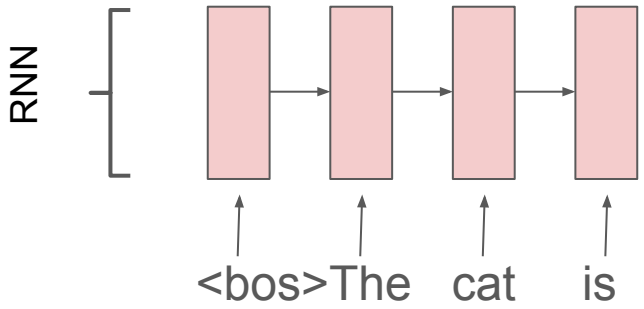
Encoder-Decoder



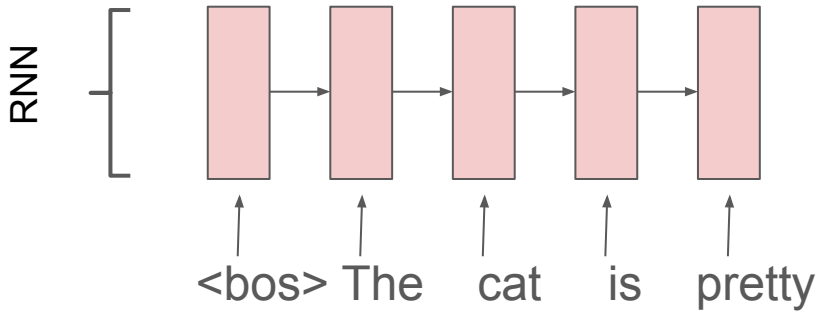
Encoder-Decoder



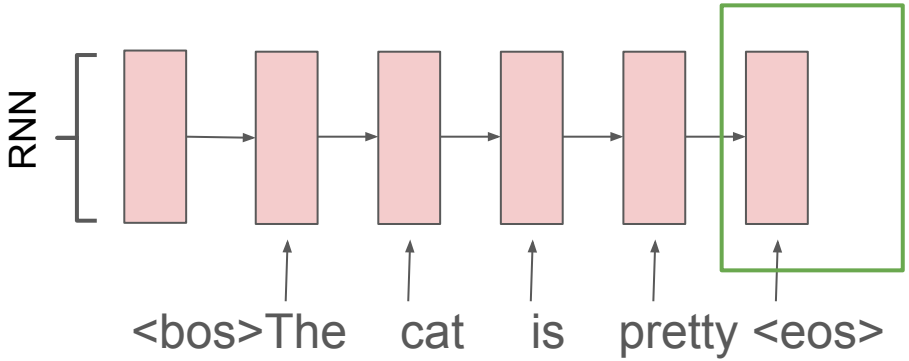
Encoder-Decoder



Encoder-Decoder



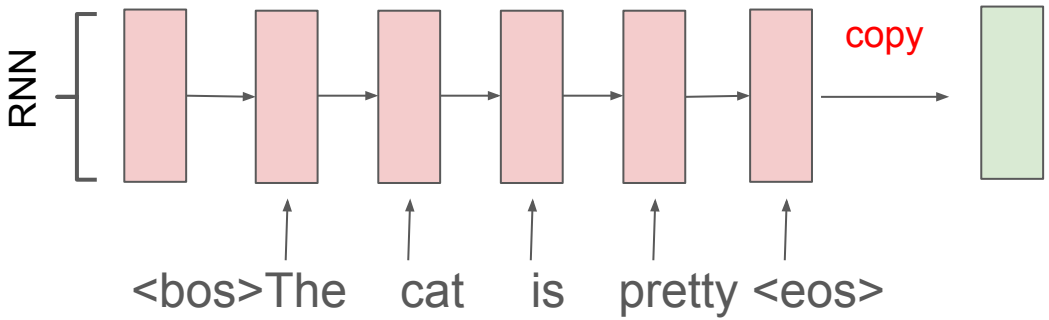
Encoder-Decoder



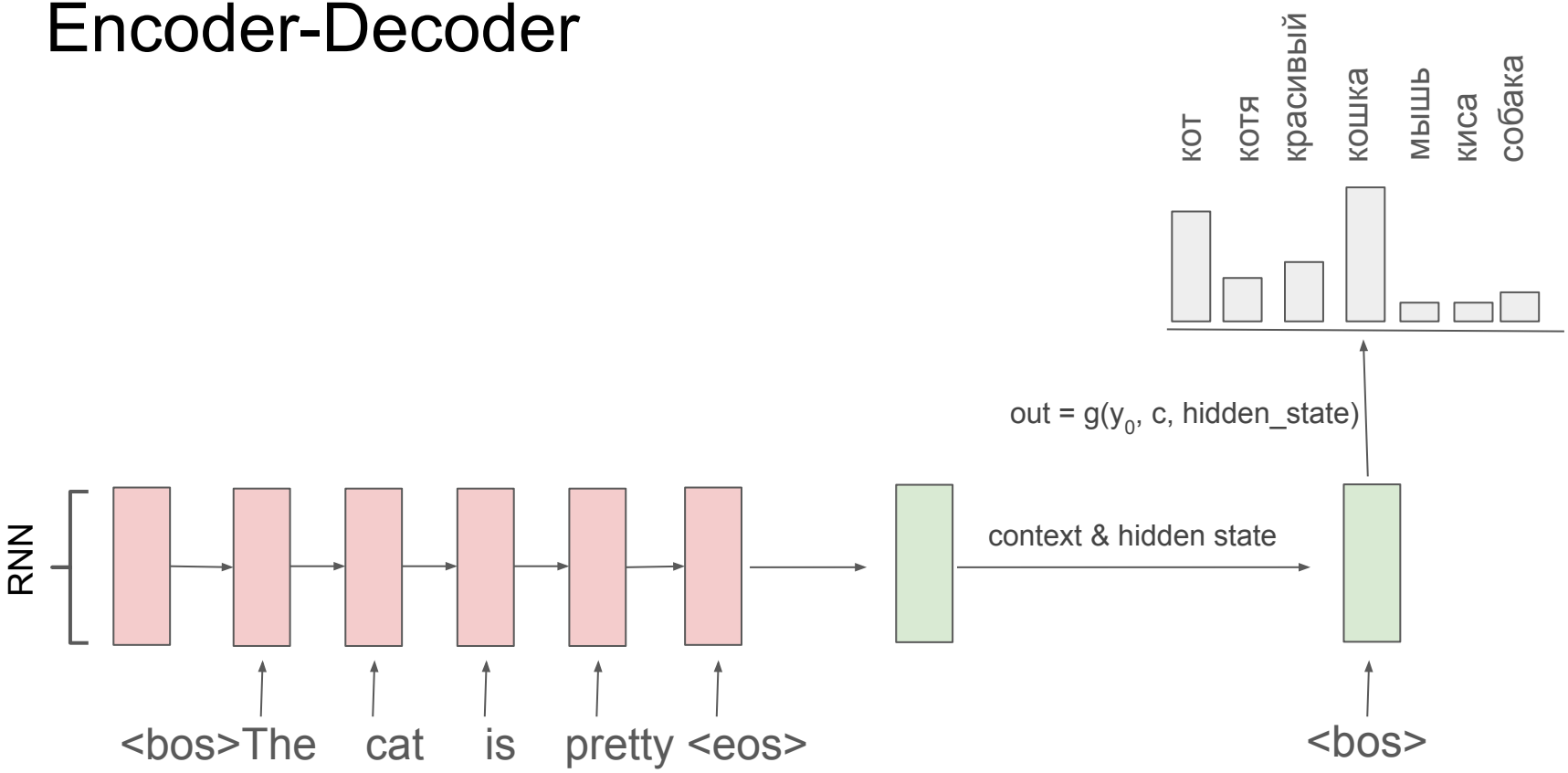
Encoding of a source sentence

Contains information about the whole sentence

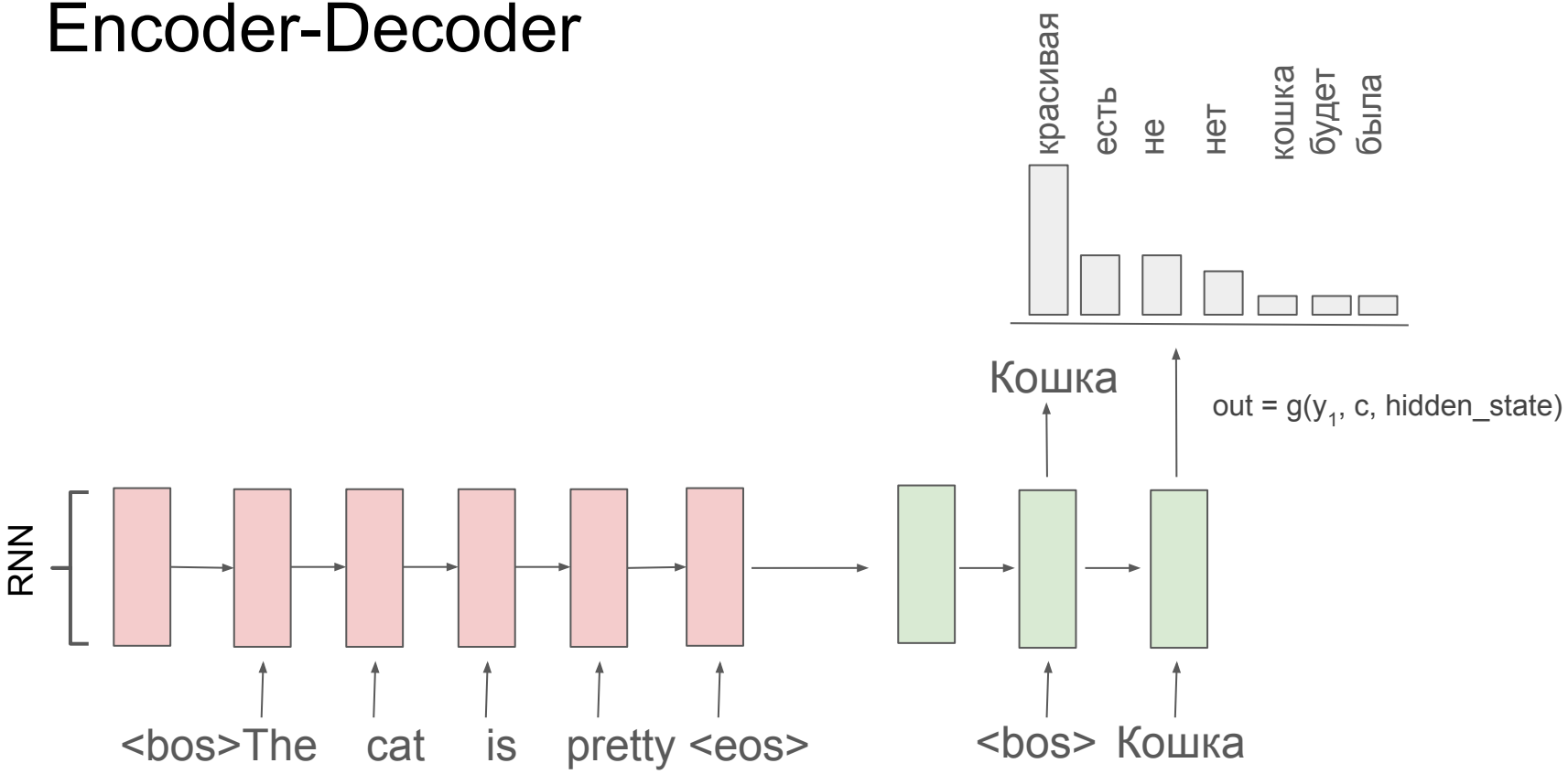
Encoder-Decoder



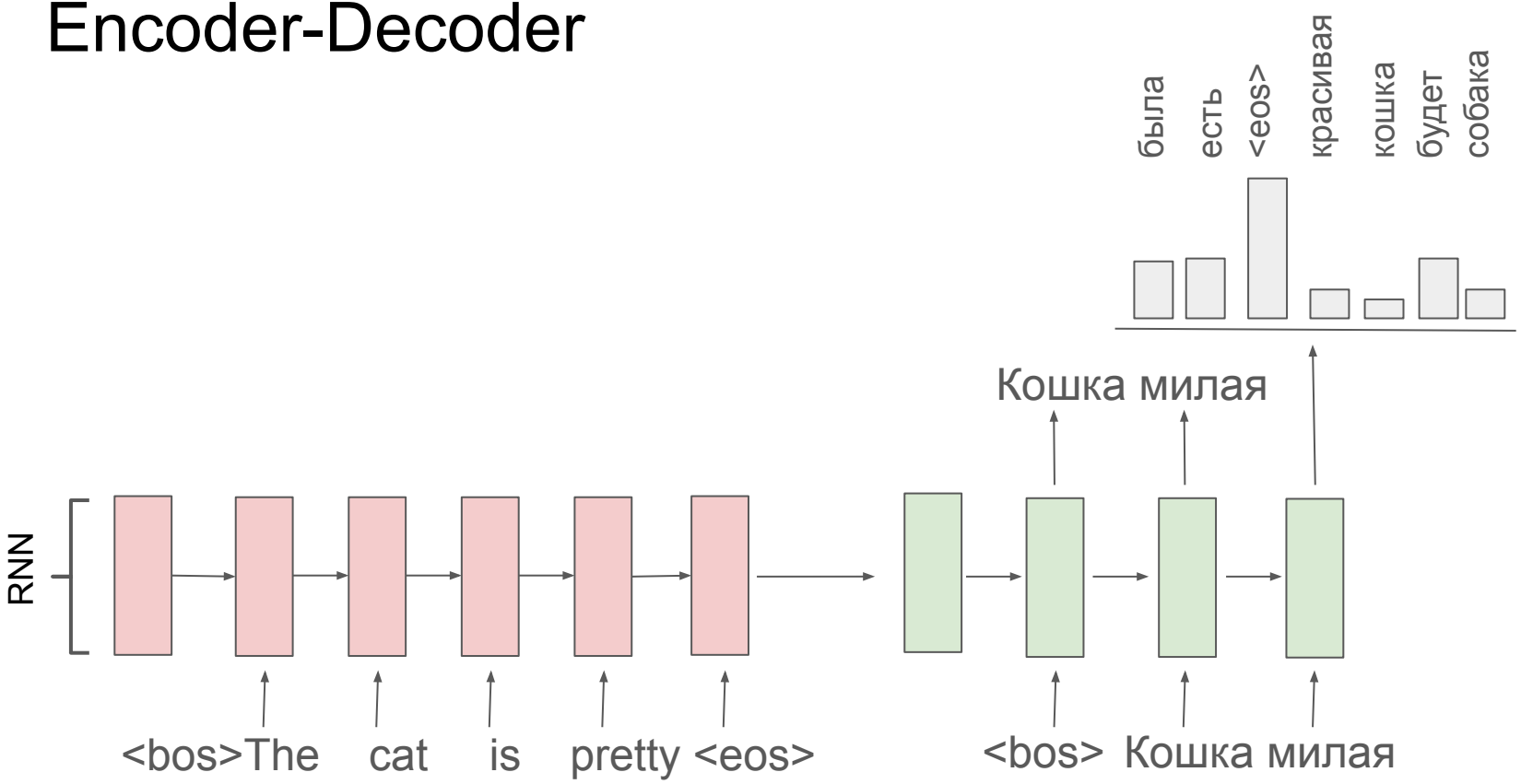
Encoder-Decoder



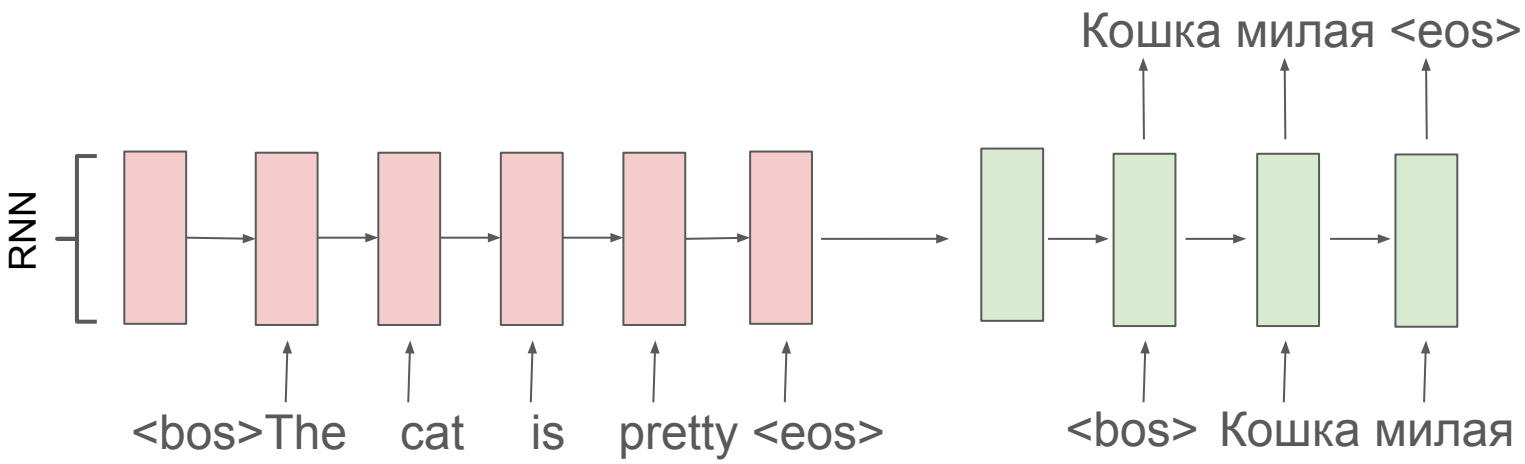
Encoder-Decoder



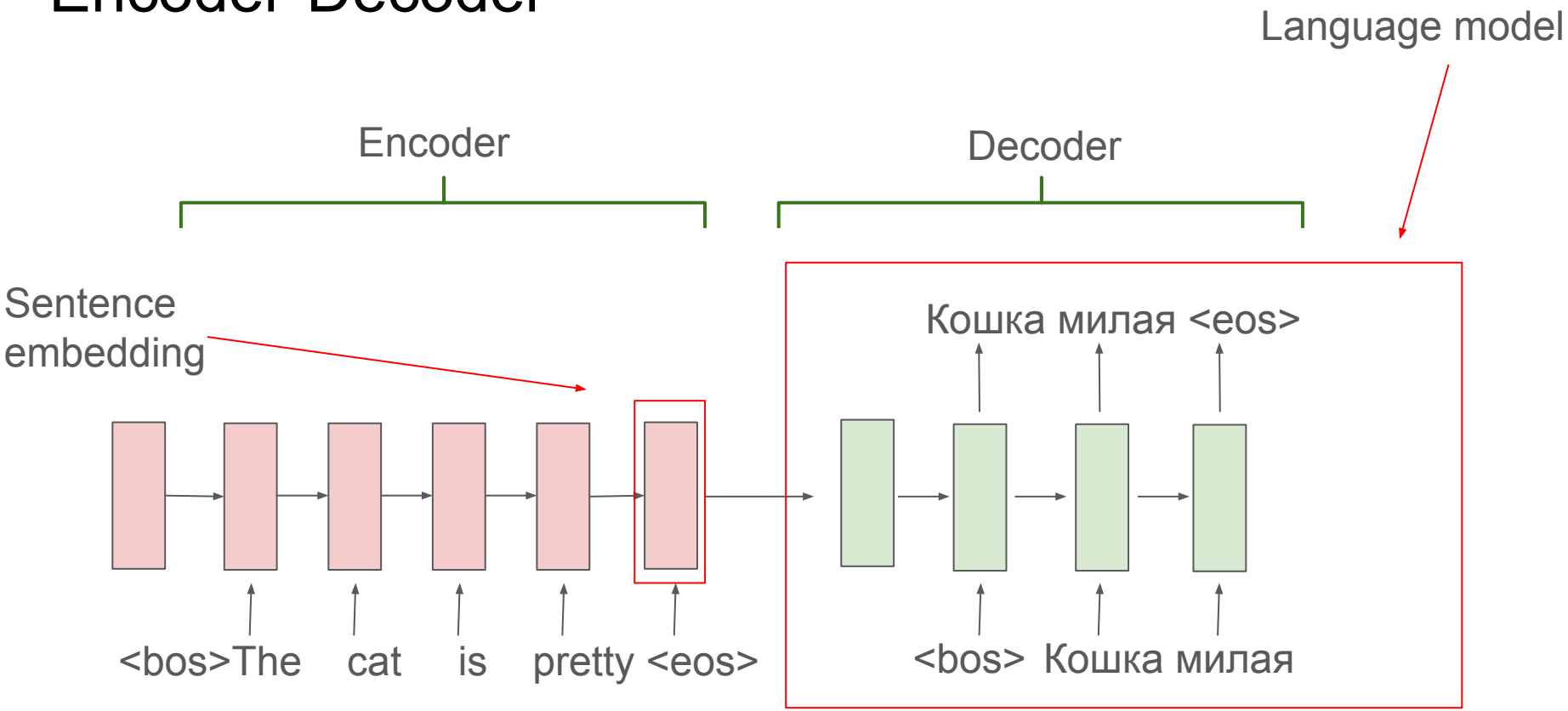
Encoder-Decoder



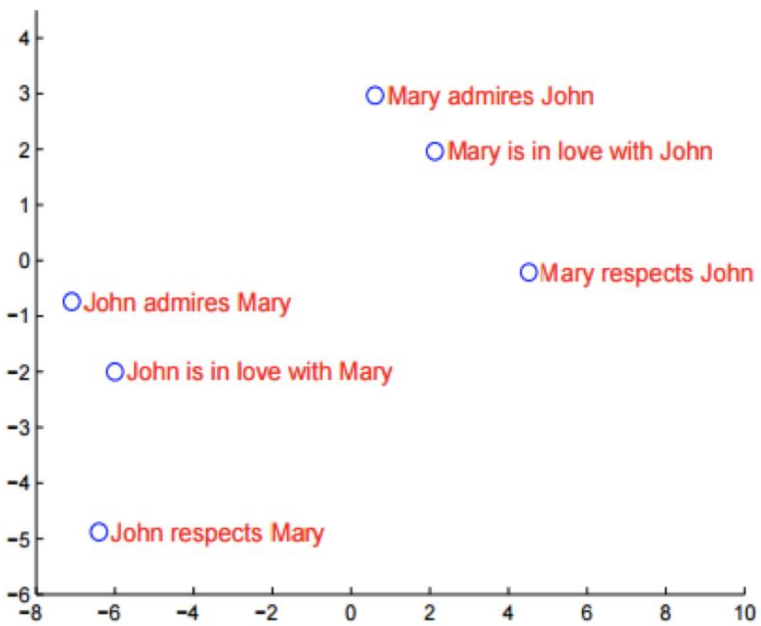
Encoder-Decoder

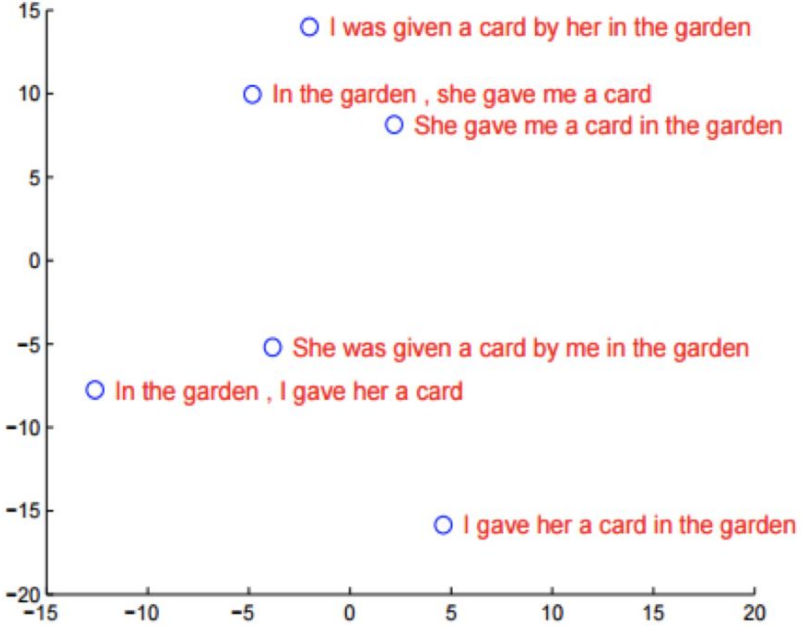


Encoder-Decoder



Sentence embedding





(Sutskever et al., 2014)

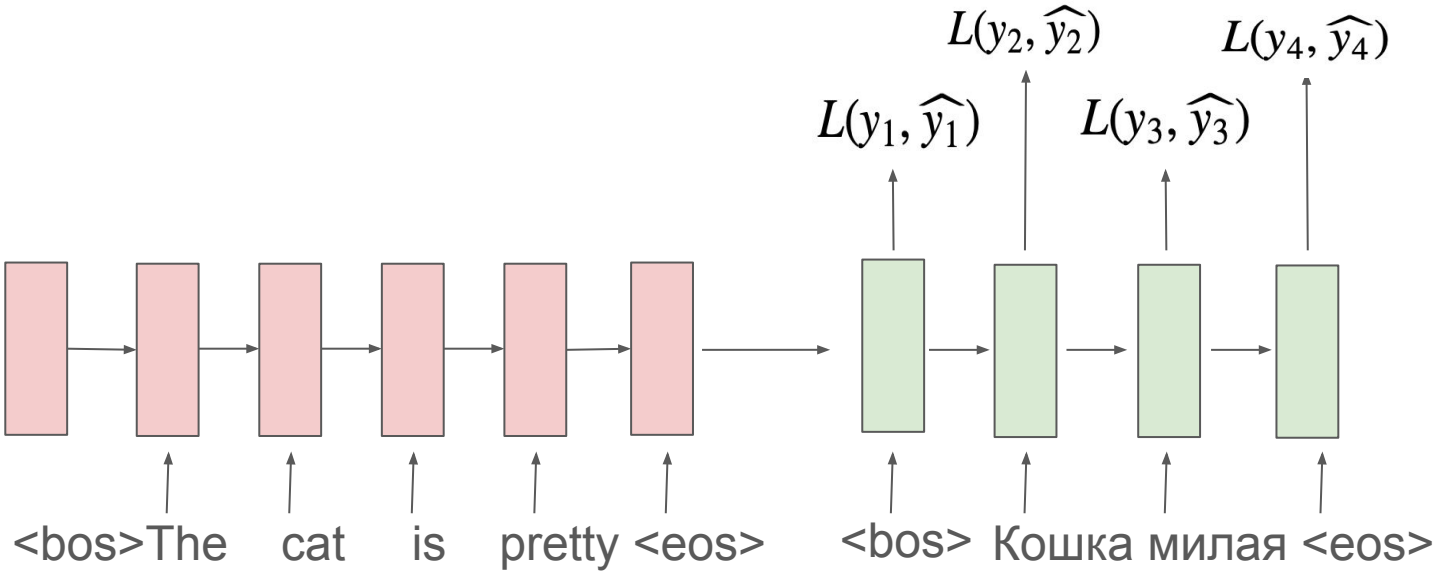
Encoder-Decoder

- consists of two parts: encoder and decoder
- encoder gathers information about source sentence

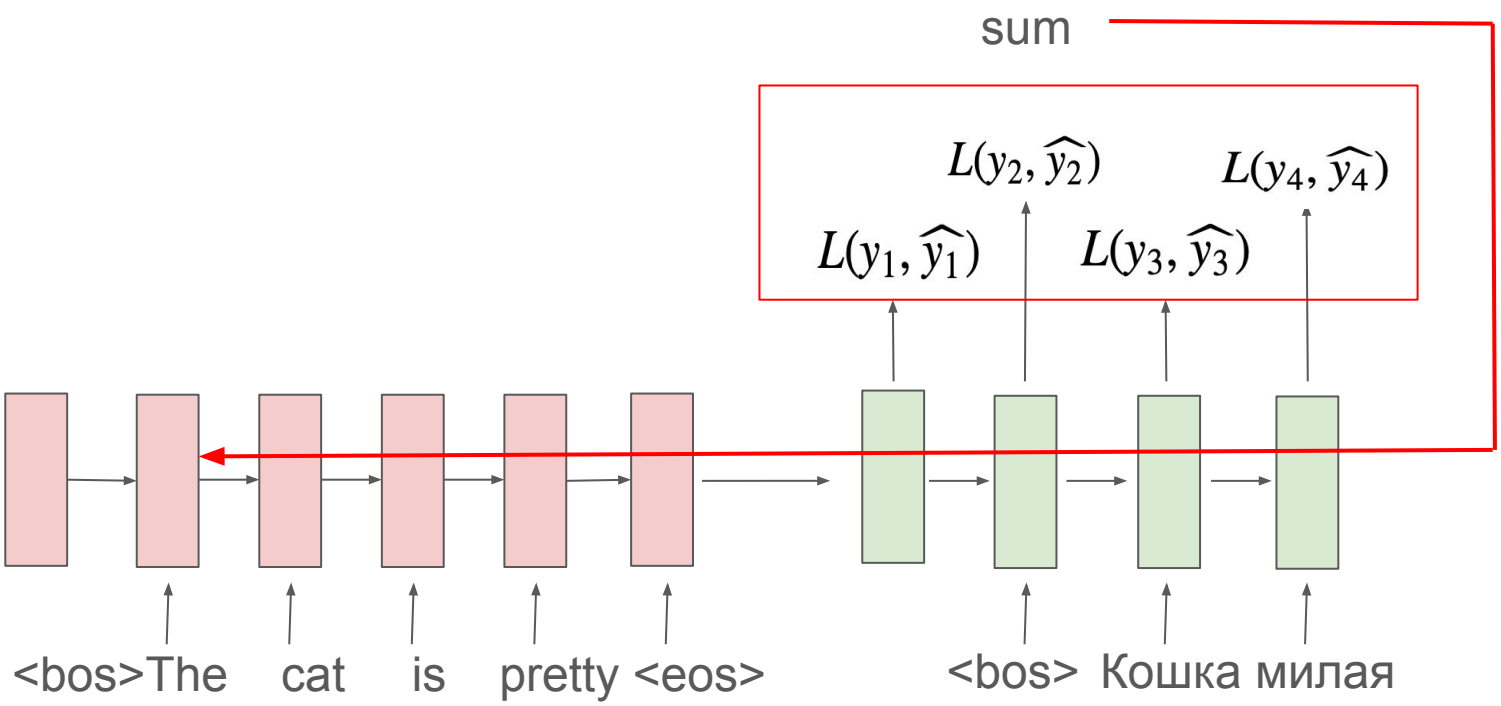
  (may be not only RNN, but anything)

- decoder uses this information to generate new sequence

Encoder-Decoder training



Encoder-Decoder training



Encoder-Decoder

- consists of two parts: encoder and decoder
- encoder gathers information about source sentence

  (may be not only RNN, but anything)

- decoder uses this information to generate new sequence

Any improvement ideas?

Any problems seen?

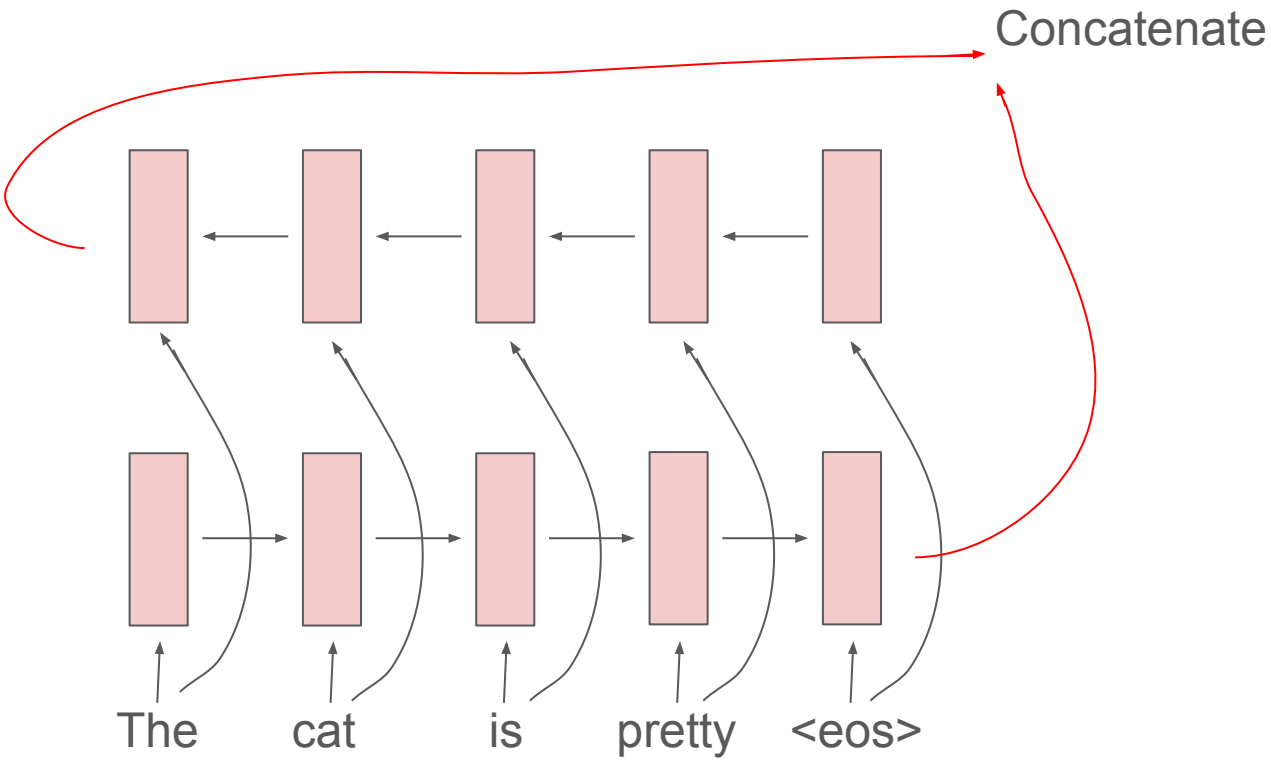
Encoder-Decoder: problems

- Encoder can forget beginning of long sentences
- Greedy decoding
- Teacher-forcing
- All the complicated human language is put into single vector!

Encoder-Decoder: problems

- **<u>Encoder can forget beginning of long sentences</u>**
- Greedy decoding
- Teacher-forcing
- All the complicated human language is put into single vector!

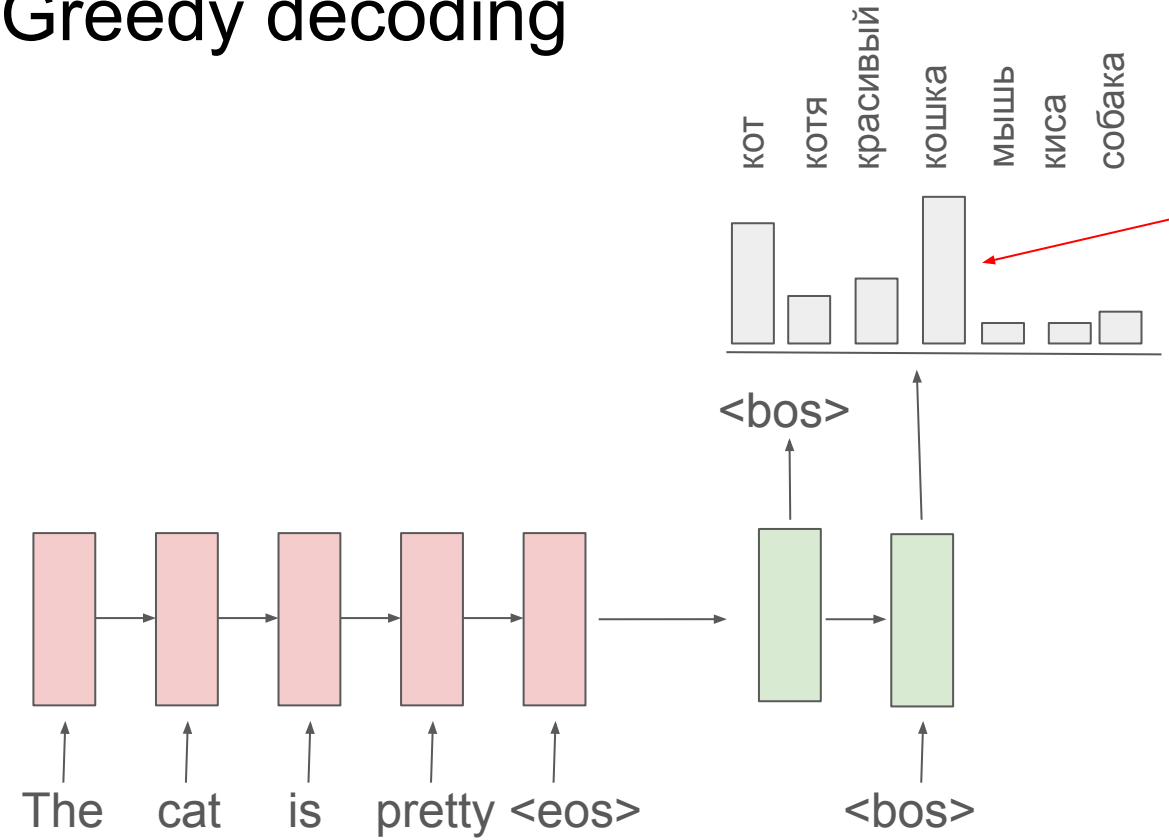
Bidirectional RNN



Encoder-Decoder: problems

- Encoder can forget beginning of long sentences
- **<u>Greedy decoding</u>**
- Teacher-forcing
- All the complicated human language is put into single vector!

Greedy decoding



Choose argmax every time

Is that a good Idea?

Greedy decoding

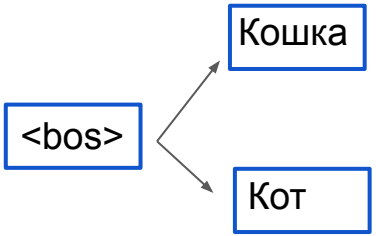
Had there been a vaccine, would he be alive.

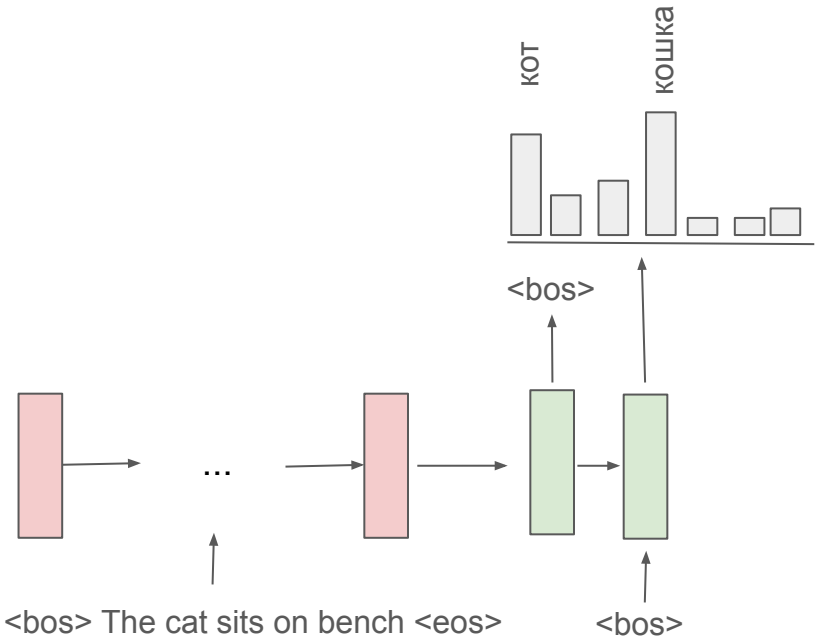
Была тут вакцина …

Если бы там была вакцина ...

Beam search

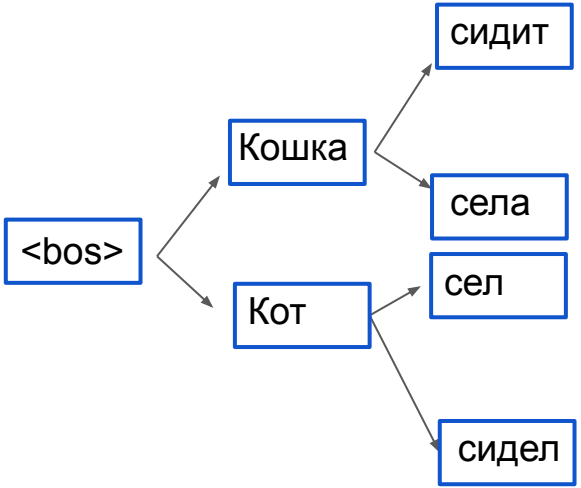
Maintain fixed number of hypotheses, extend them and choose most probable ones

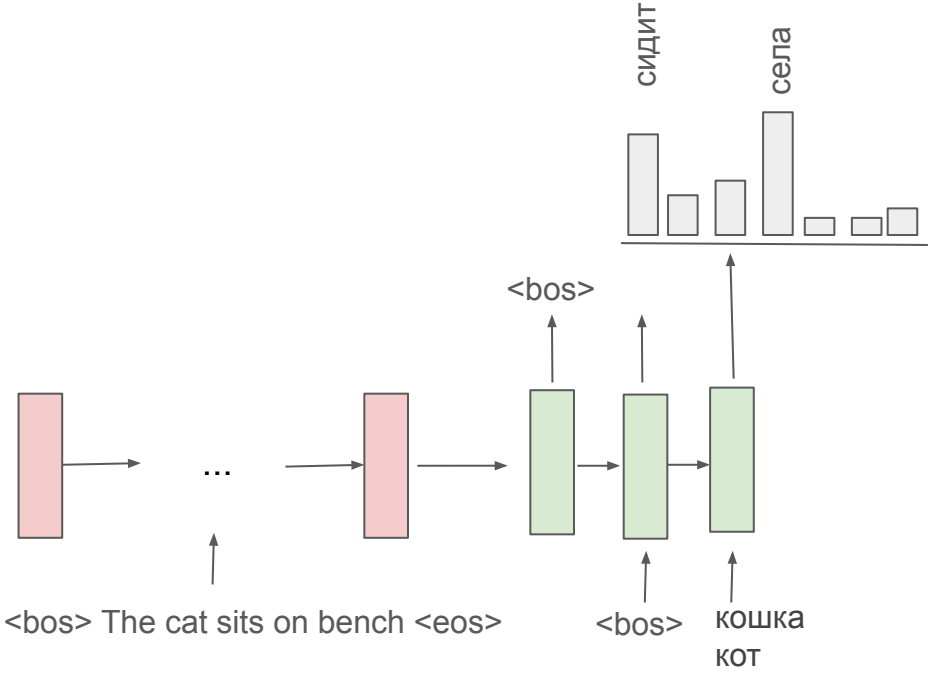




Beam search

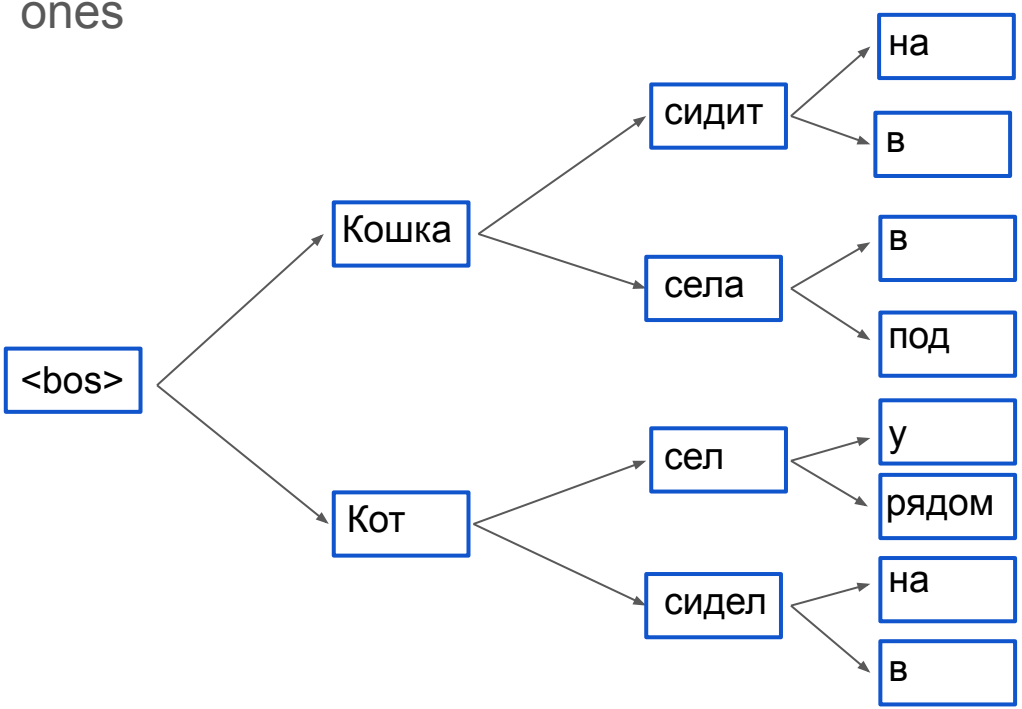
Maintain fixed number of hypotheses, extend them and choose most probable ones





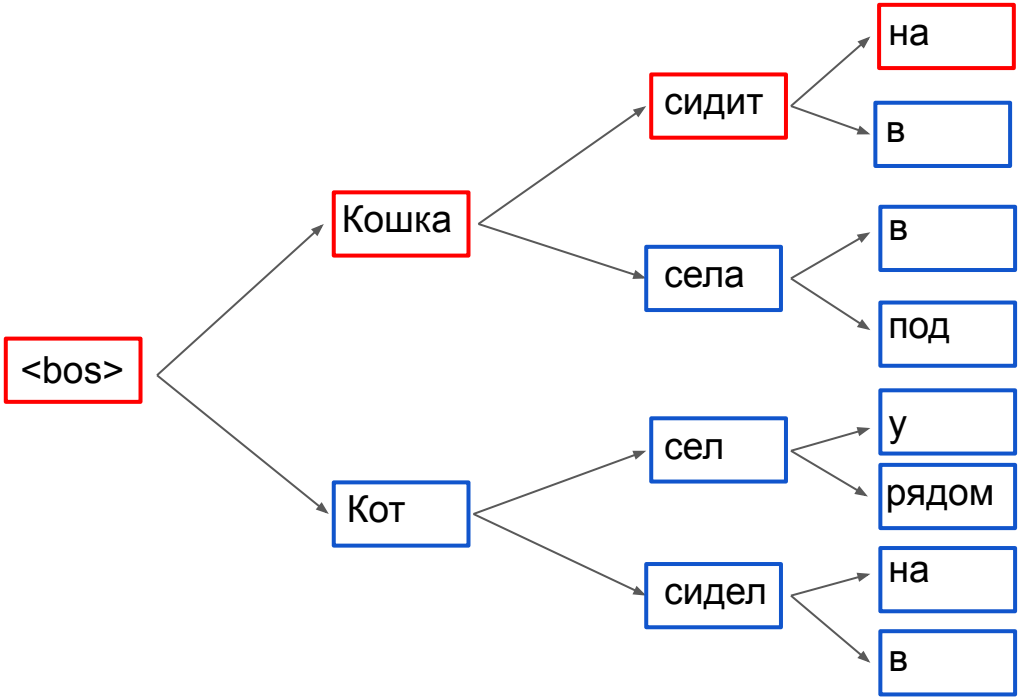
Beam search

Maintain fixed number of hypotheses, extend them and choose most probable ones



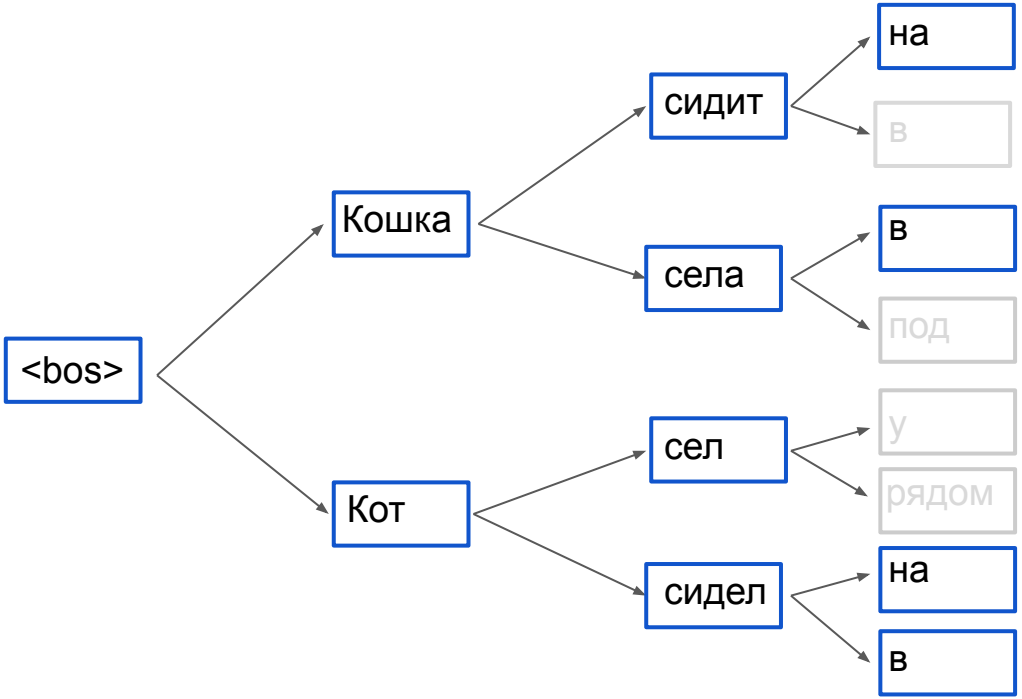
Beam search

P(Кошка сидит на) = P(на | кошка сидит) * P(сидит | кошка) * P(кошка)



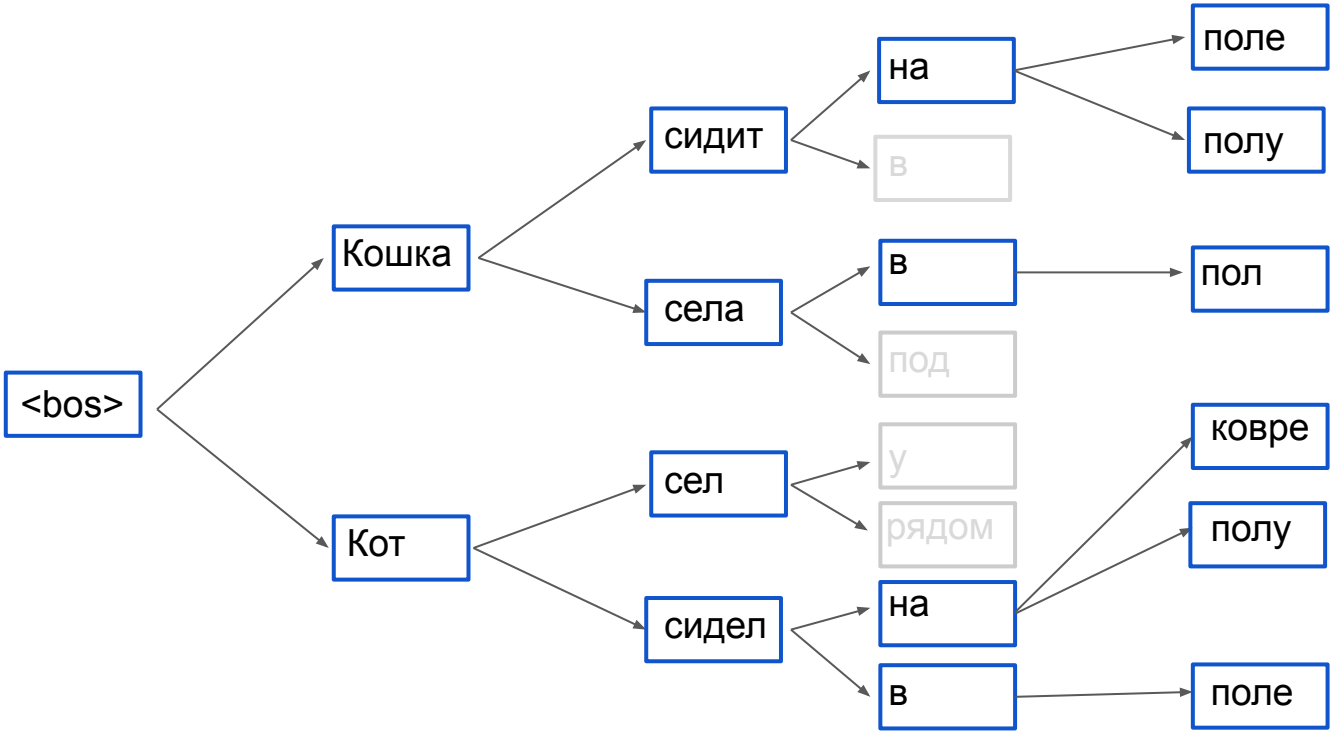
Beam search

Leave only fixed number of hypotheses with best scores

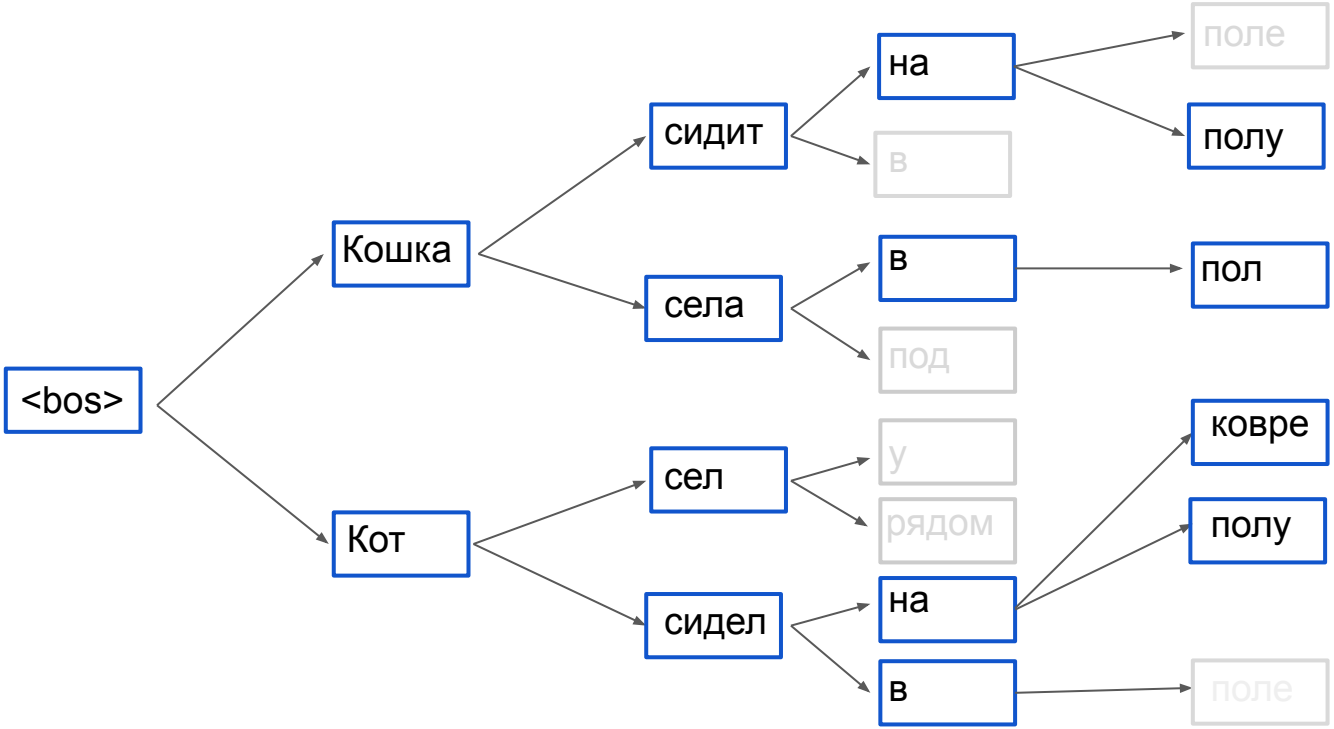


Beam search

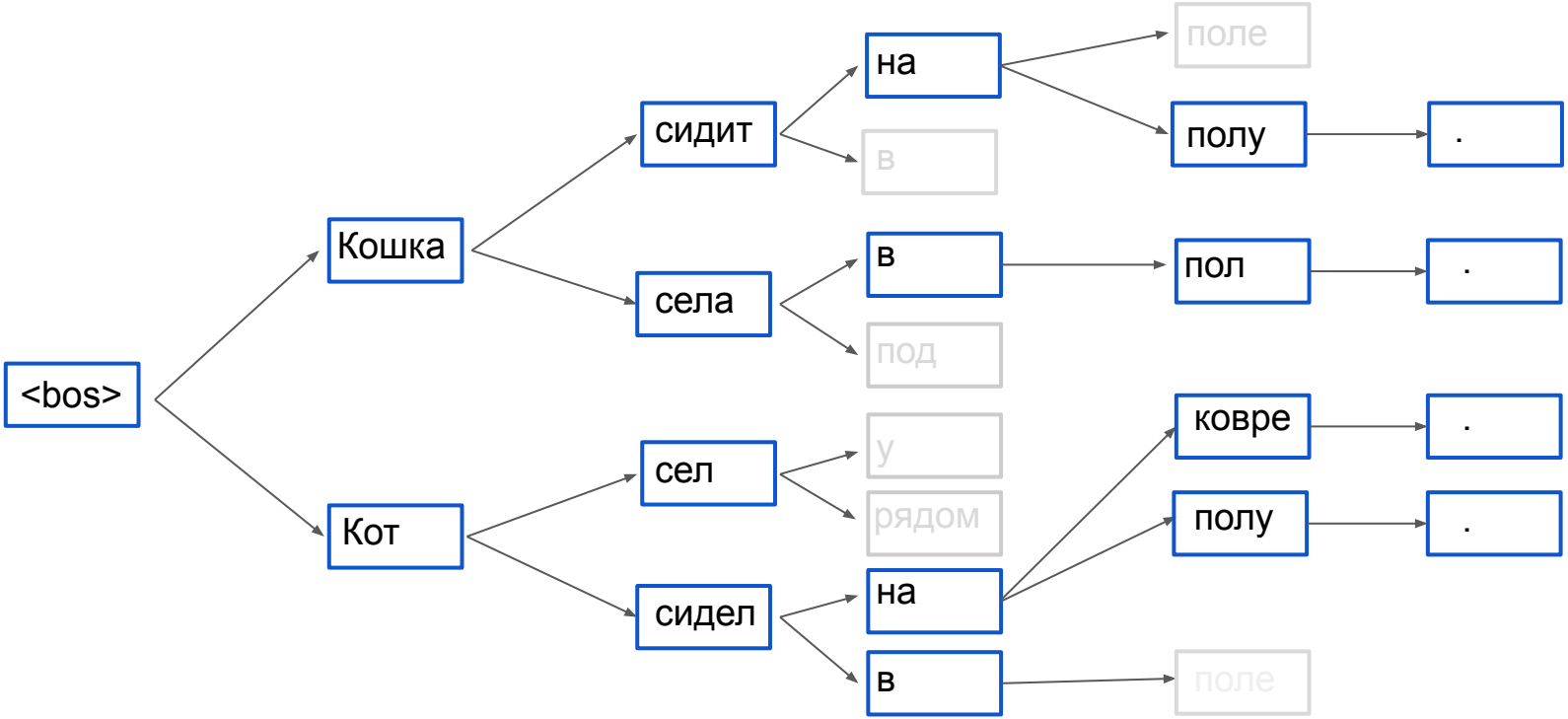
Expand best hypotheses further



Beam search

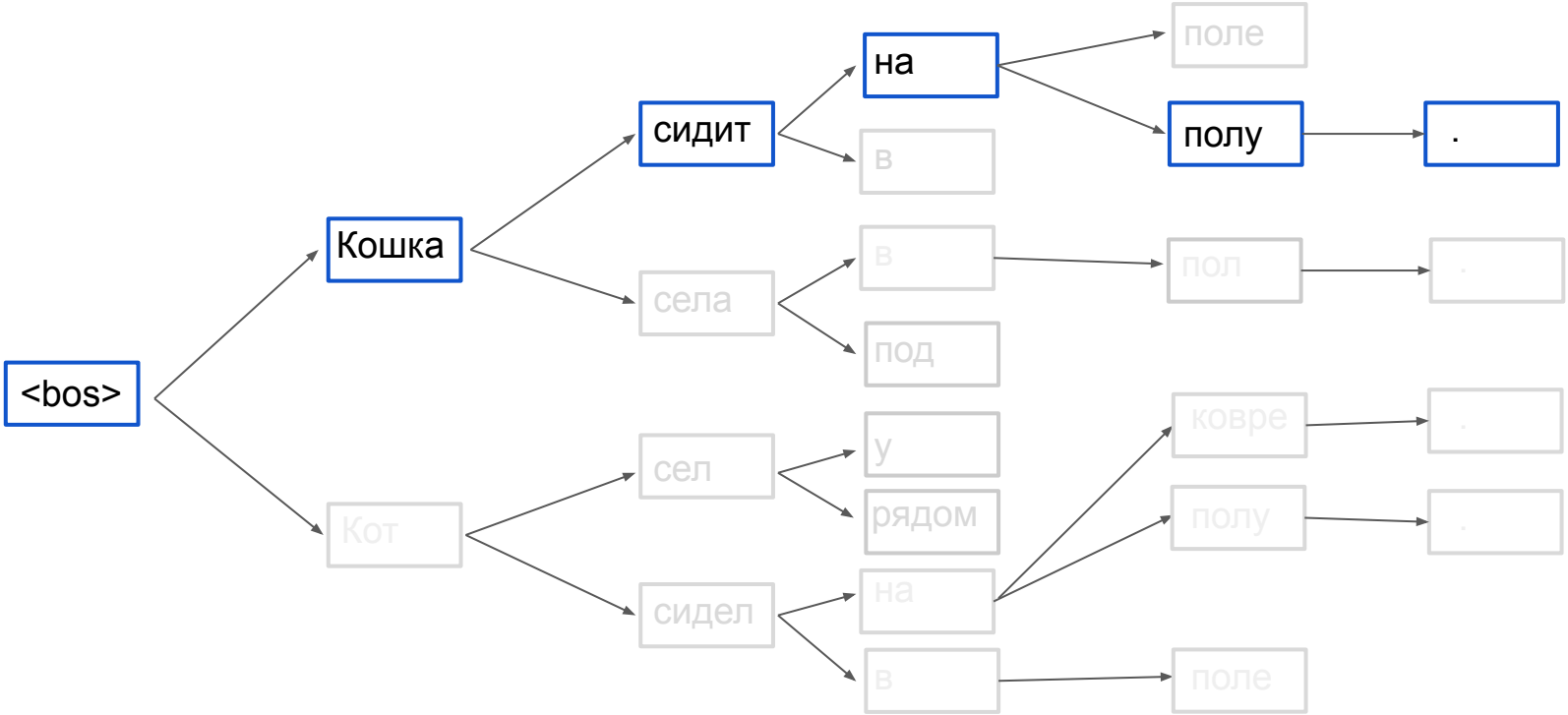

<bos>
Кошка
Кот
сидит
села
сел
сидел
на
в
в
под
у
рядом
на
в
поле
полу
пол
ковре
полу
поле

Beam search


<bos>
Кошка
Кот
сидит
села
сел
сидел
на
в
в
под
у
рядом
на
в
поле
полу
пол
ковре
полу
поле
.
.
.
.

Beam search

Choose the best overall hypothesis.



Beam search

Maintain fixed number of hypotheses, extend them and choose most probable ones

Optimal beam size ~ 4-8

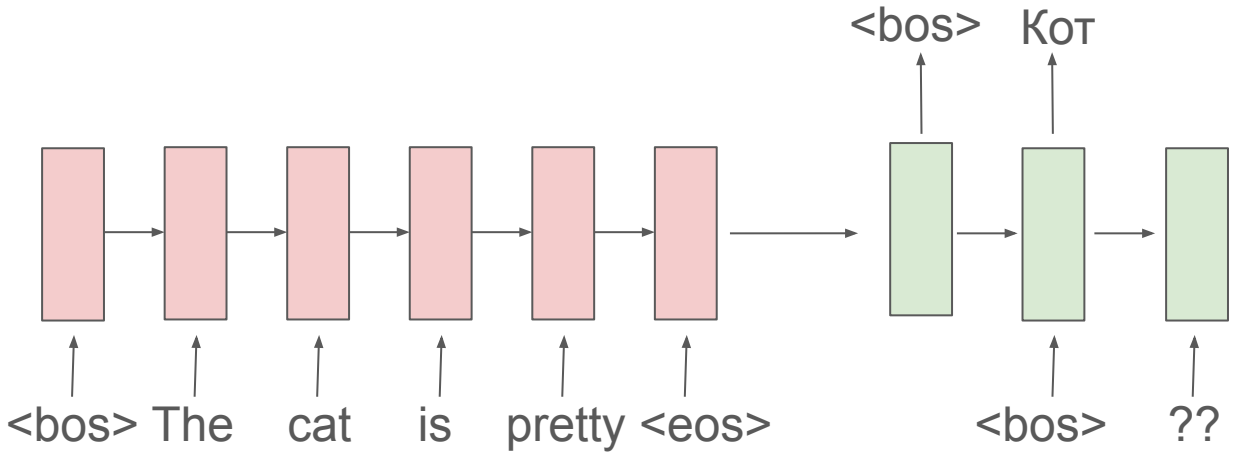
If beam size is too big, the translation model behaves badly (quality decreases)

Encoder-Decoder: problems

- Encoder can forget beginning of long sentences
- Greedy decoding
- **Teacher-forcing**
- All the complicated human language is put into single vector!

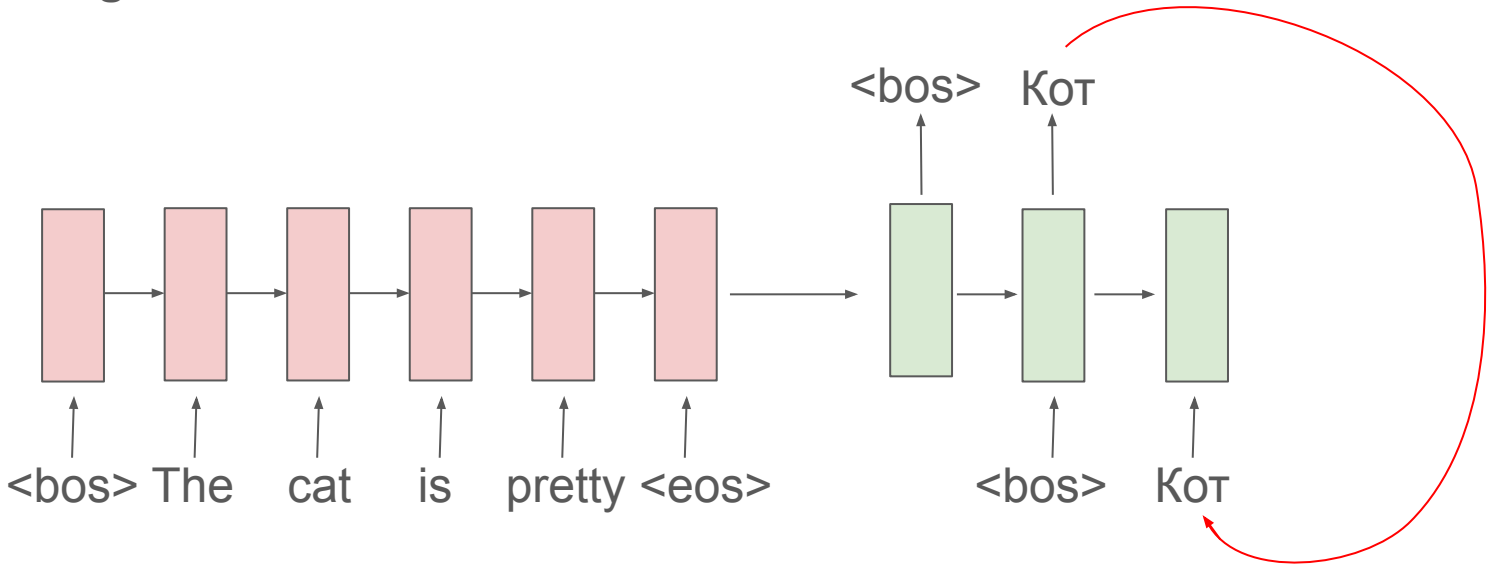
Teacher-forcing

Source: The cat is pretty
Target: Кошка милая



Teacher-forcing

Source: The cat is pretty
Target: Кошка милая

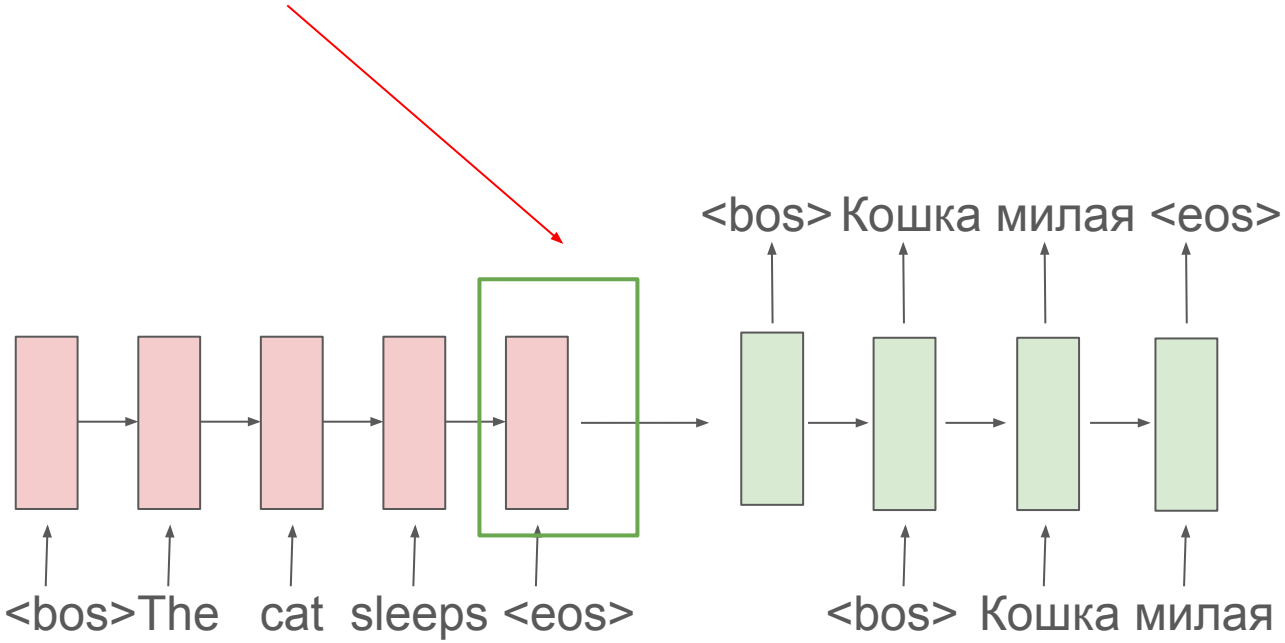


Encoder-Decoder: problems

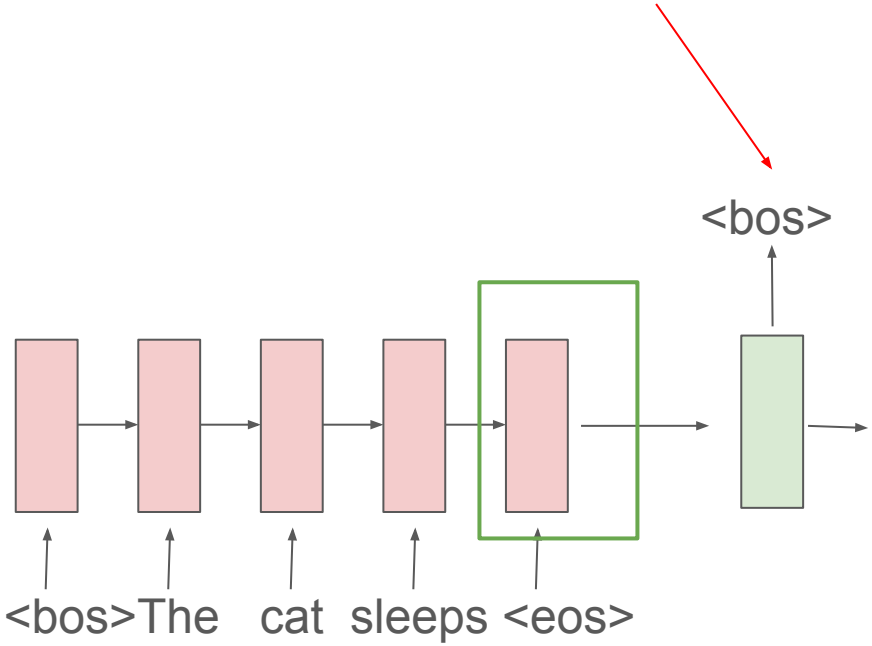
- Encoder can forget beginning of long sentences
- Greedy decoding
- Teacher-forcing
- **All the complicated human language is put into single vector!**

Attention

A poor single vector responsible for whole sentence



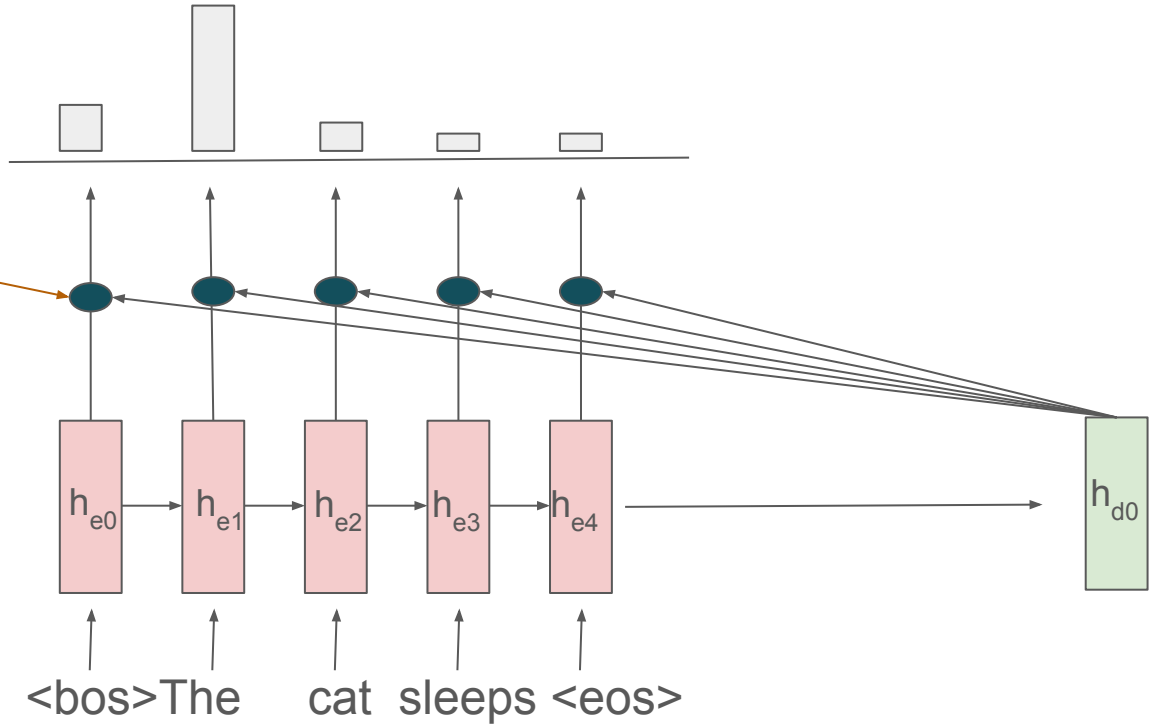
Get prediction based only on a hidden state of RNN & context (embedding)



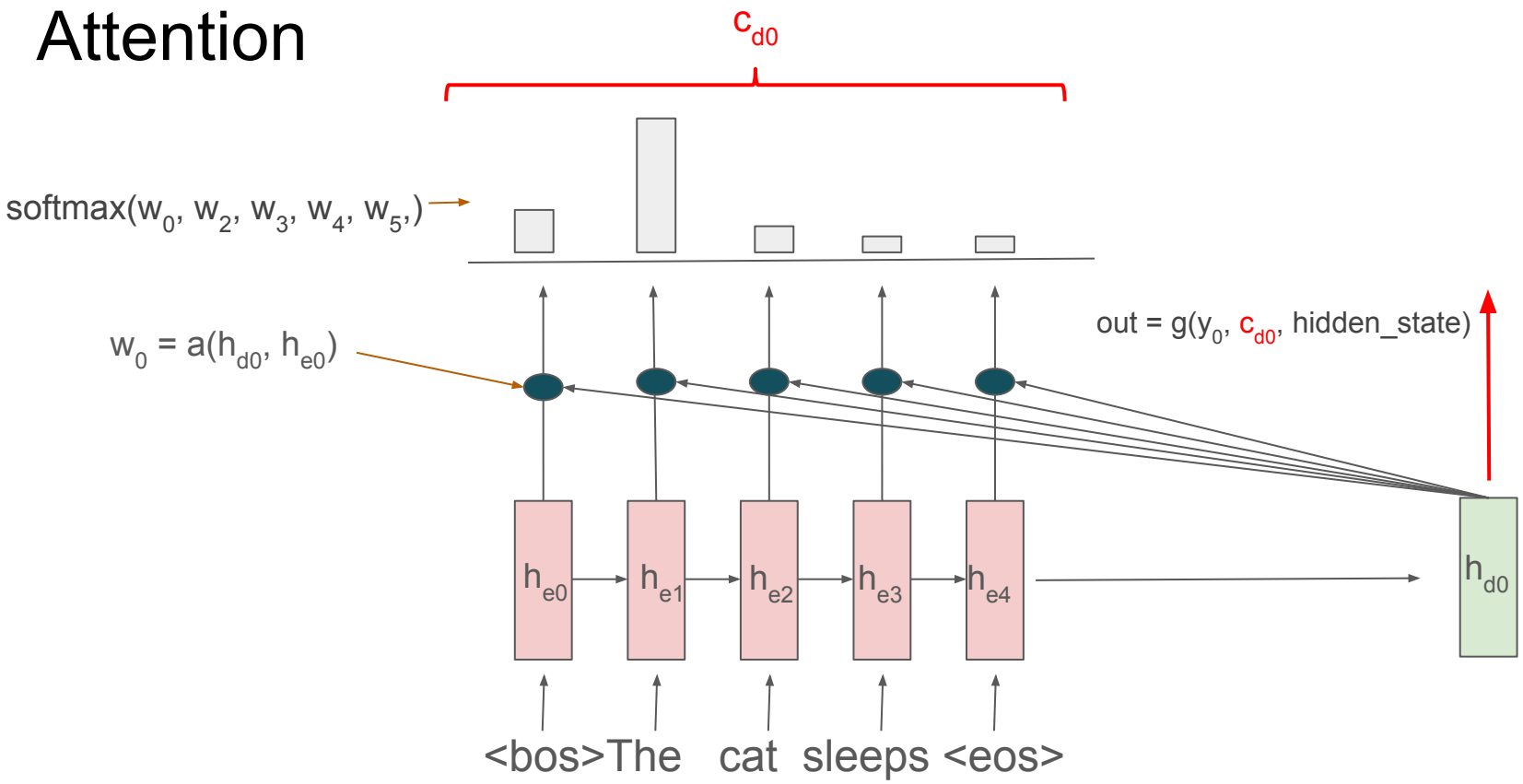
Attention

Get weight of word “the”:

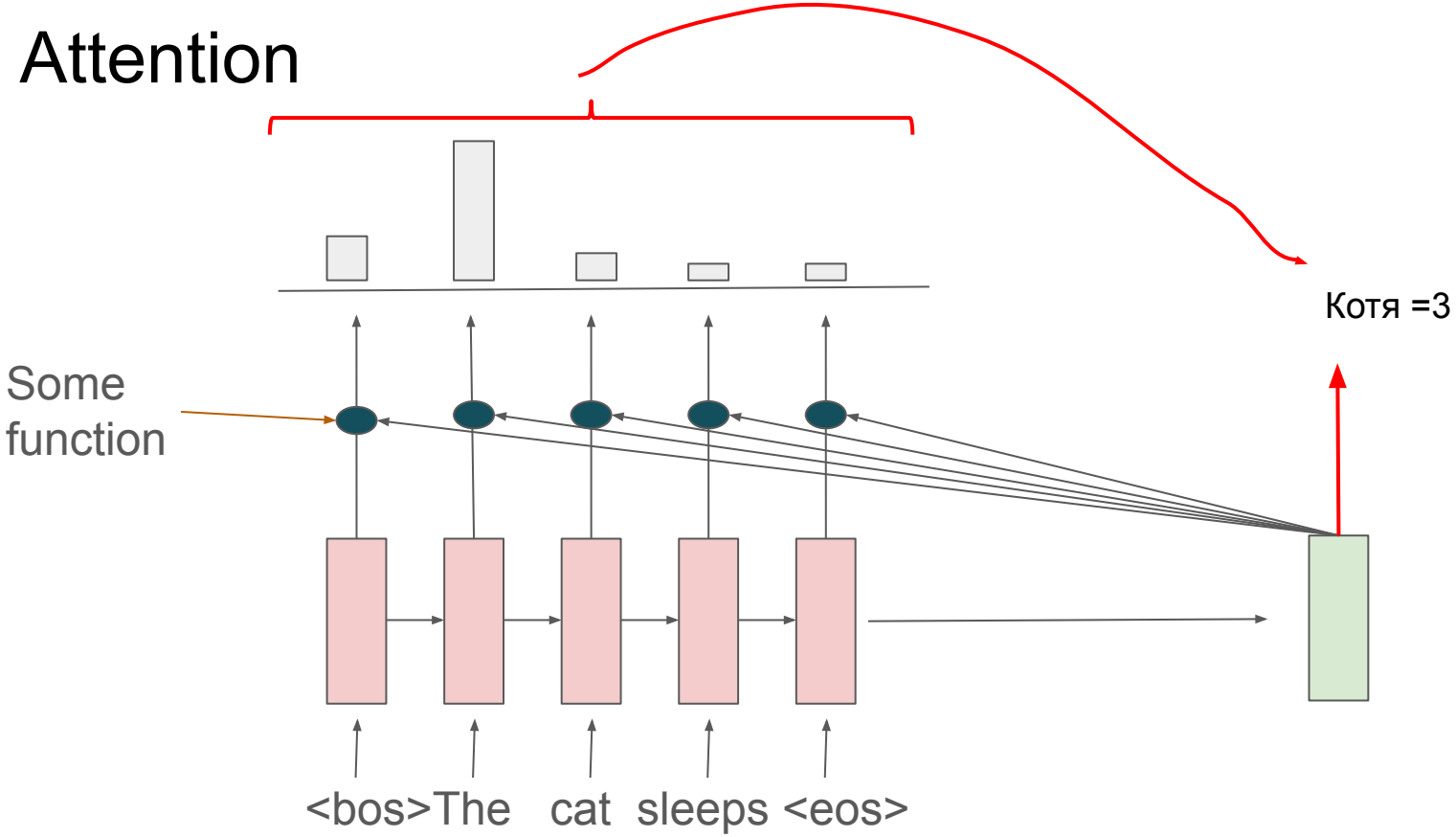
$w_0 = a(h_{d0}, h_{e0})$



Attention


c_{d0}
softmax(w_0, w_2, w_3, w_4, w_5,)
w_0 = a(h_{d0}, h_{e0})
out = g(y_0, c_{d0}, hidden_state)
h_{e0}
h_{e1}
h_{e2}
h_{e3}
h_{e4}
h_{d0}
<bos>The cat sleeps <eos>

Attention



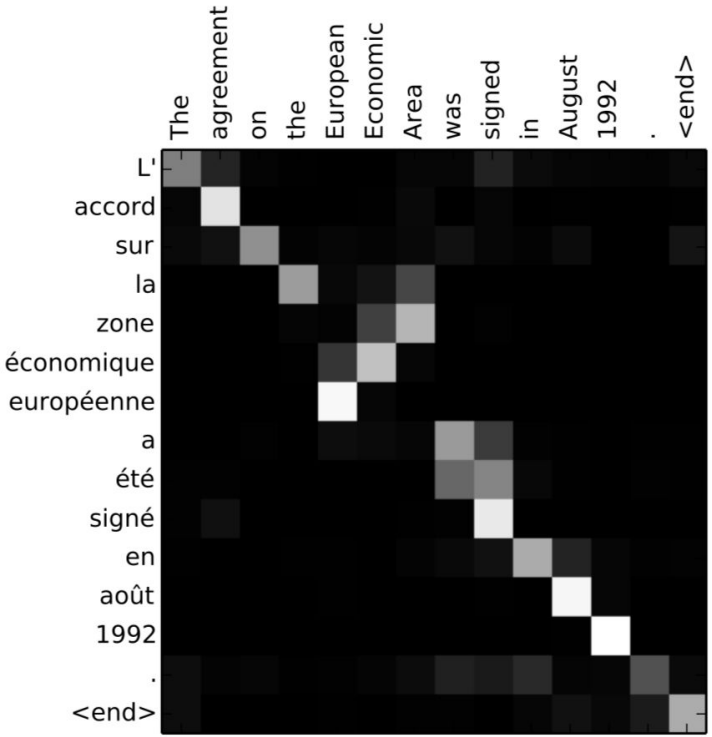
Attention

Advantages:

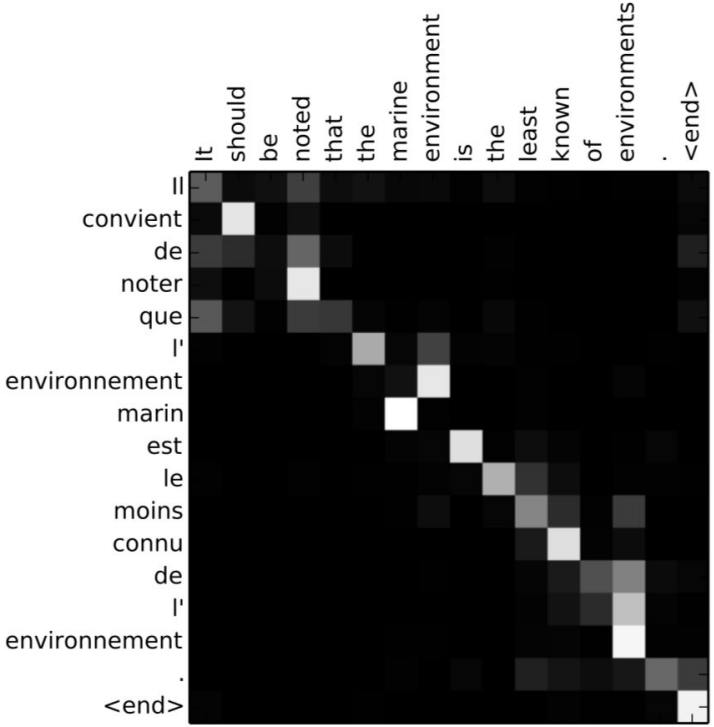
- behaves like a human
- we don’t need to remember whole sentence

  (don’t need to put the whole sentence to one fixed-size vector)

Attention



(a)



(b)

https://arxiv.org/pdf/1409.0473.pdf

Attention

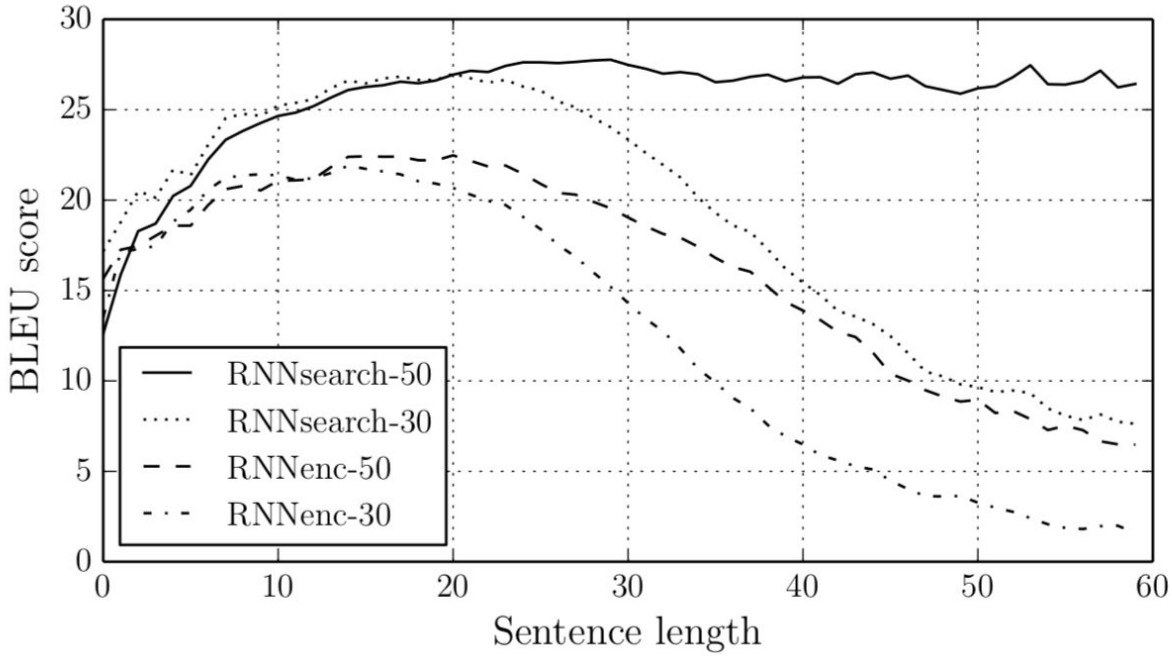
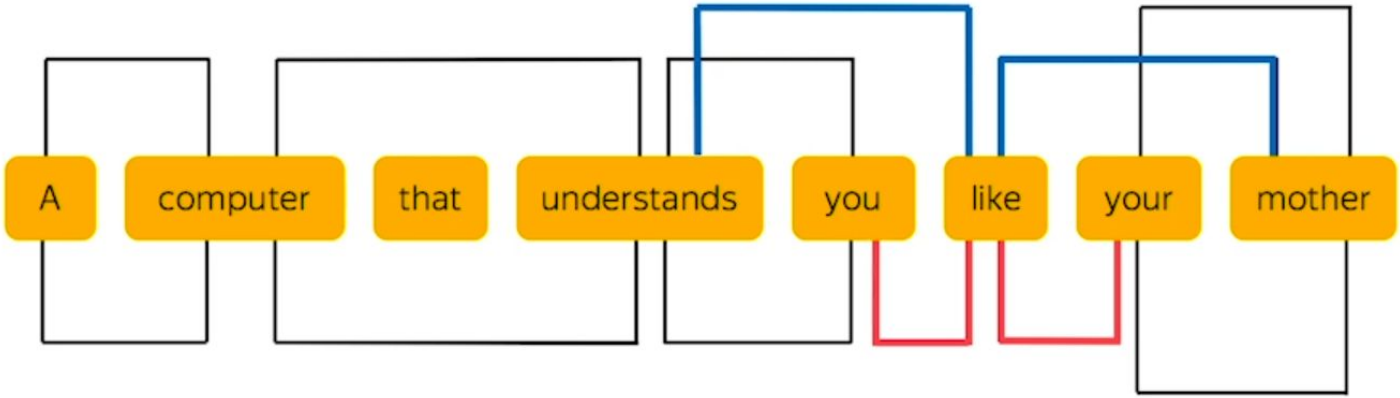


Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

https://arxiv.org/pdf/1409.0473.pdf

Self-attention



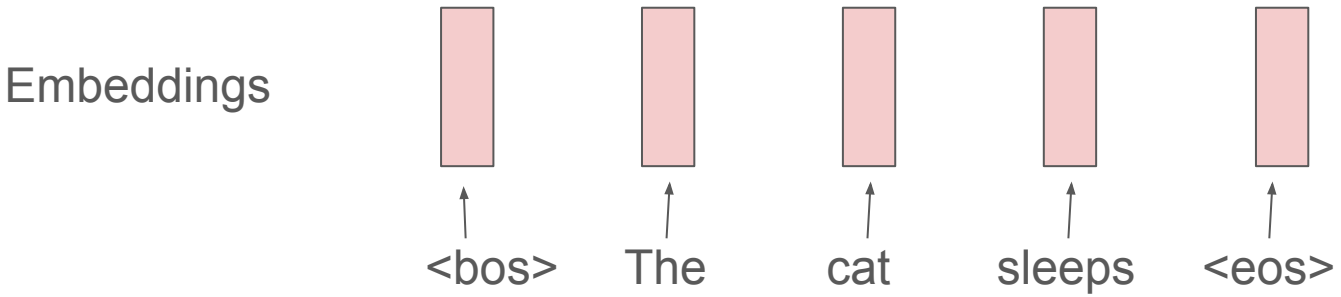
Example from David Talbot, Yandex MT, YSDA

Self-attention

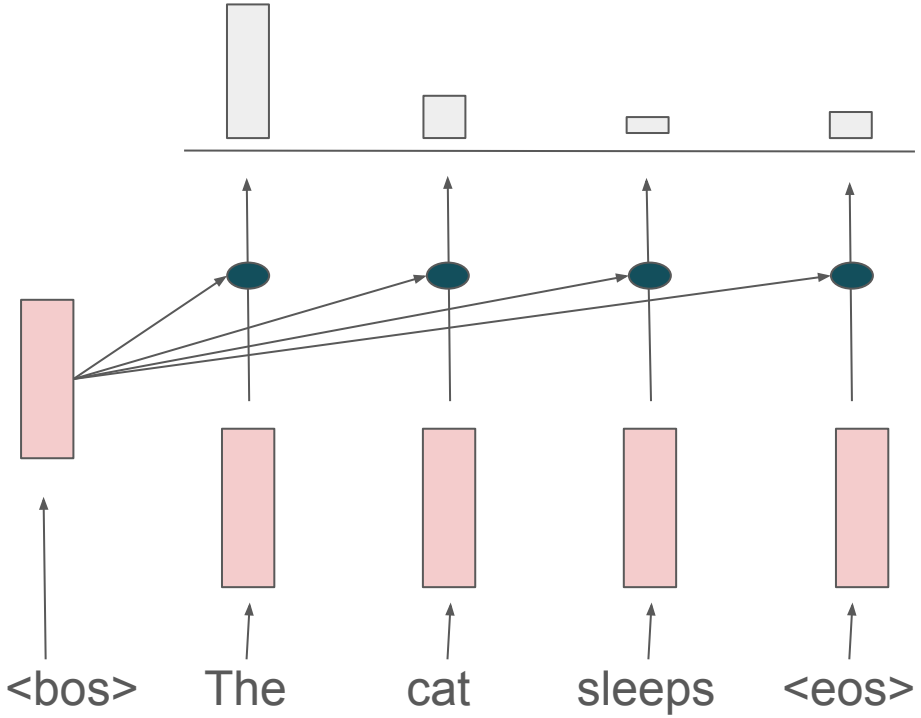
Она ест вкусную шаурму

- Case agreement
- Gender agreement

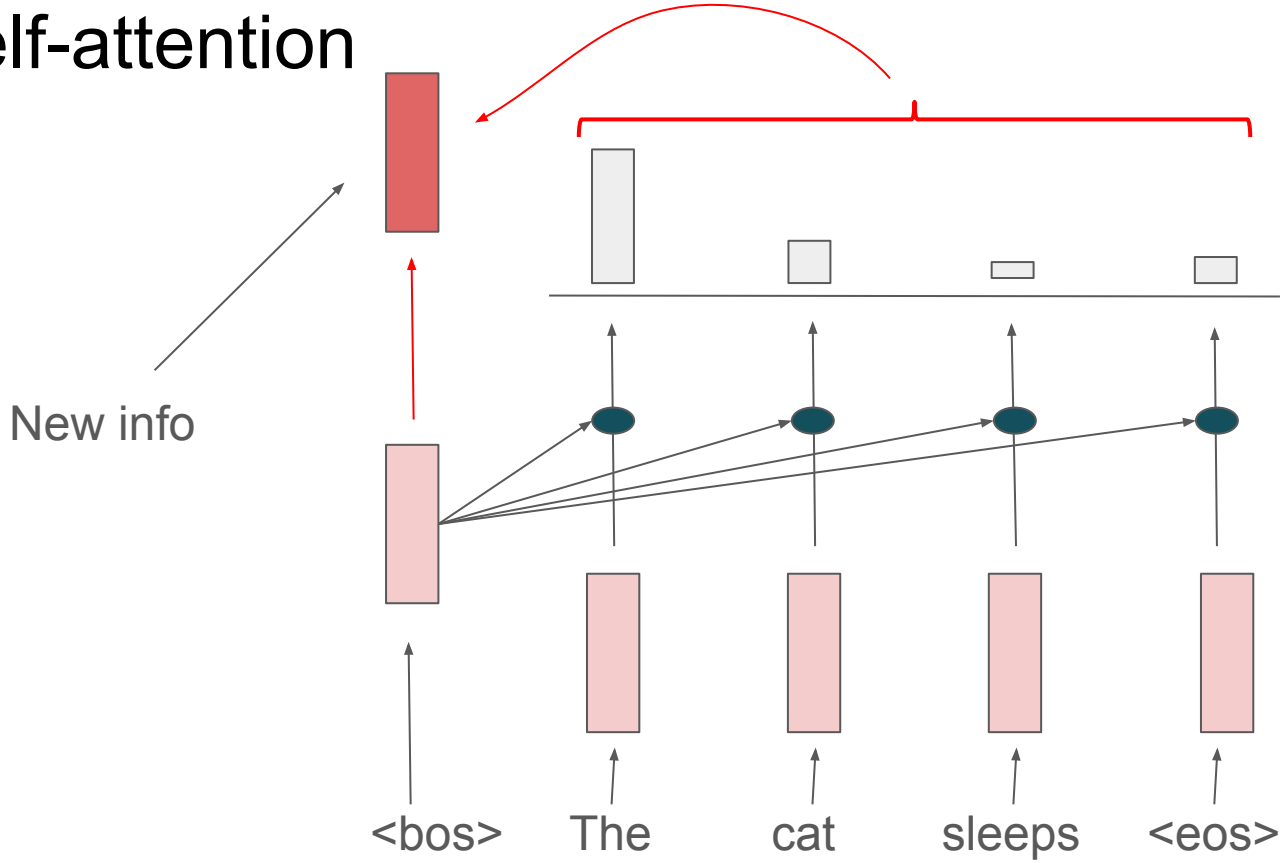
Self-attention



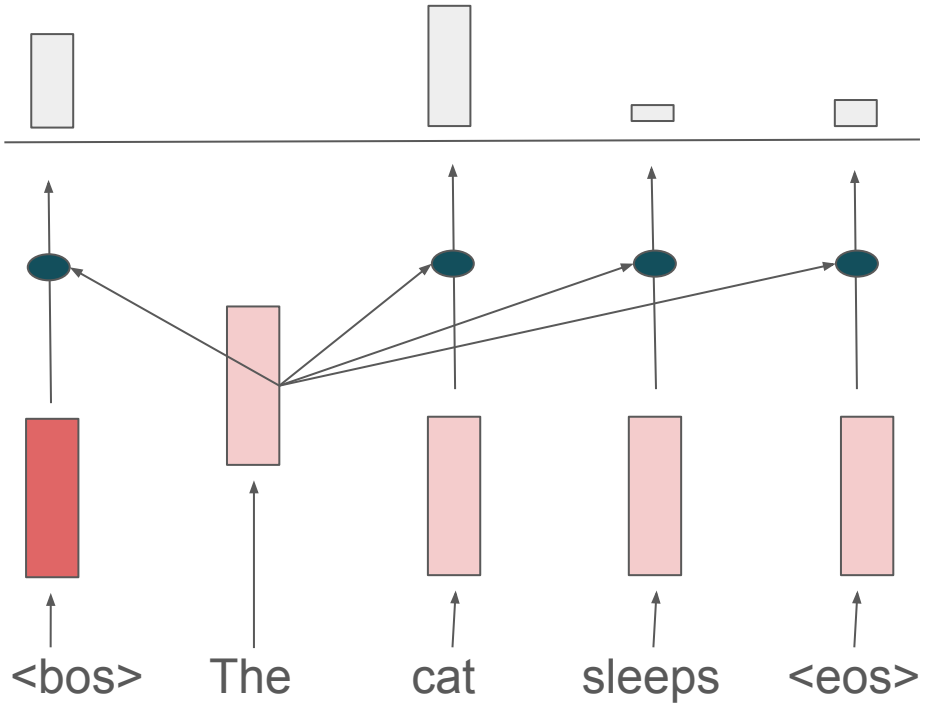
Self-attention



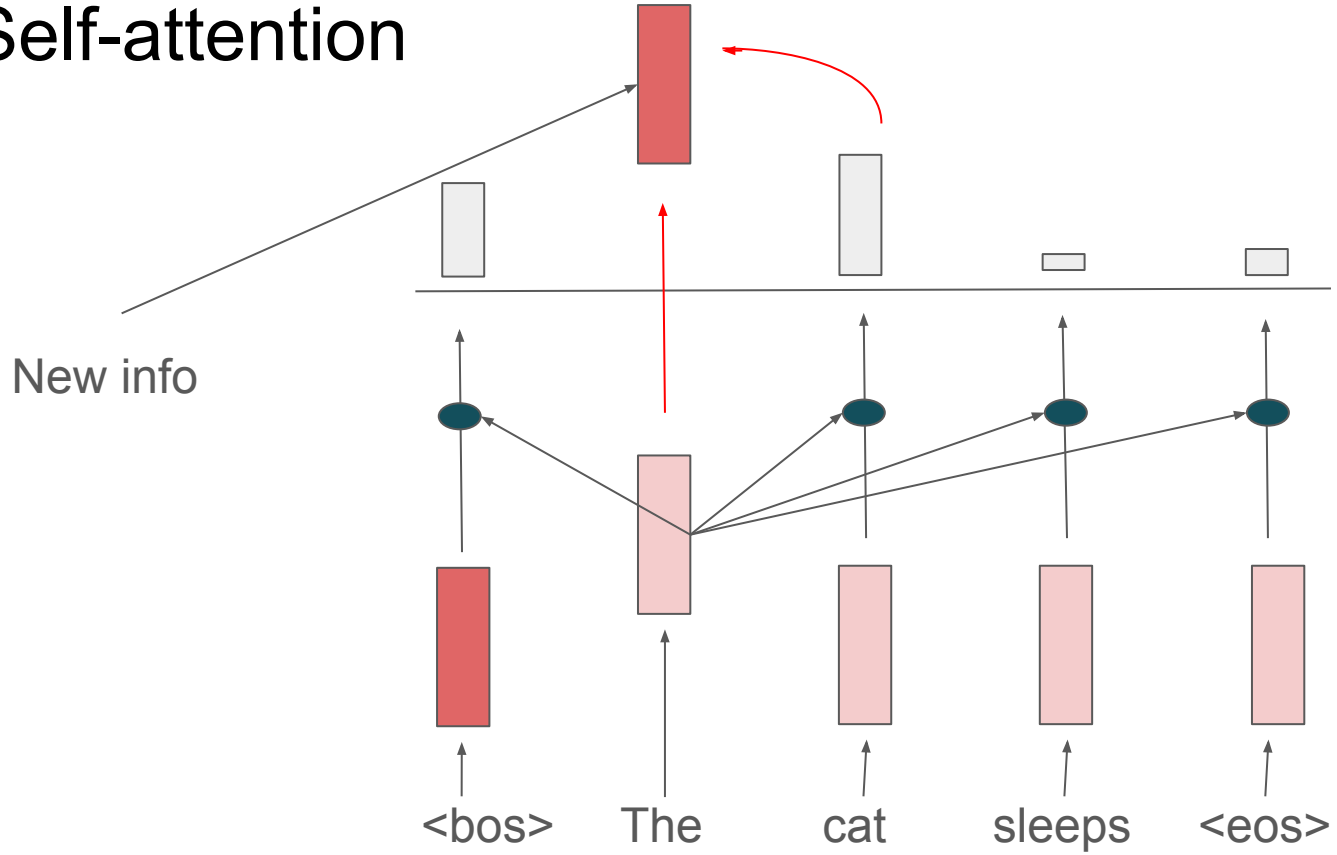
Self-attention



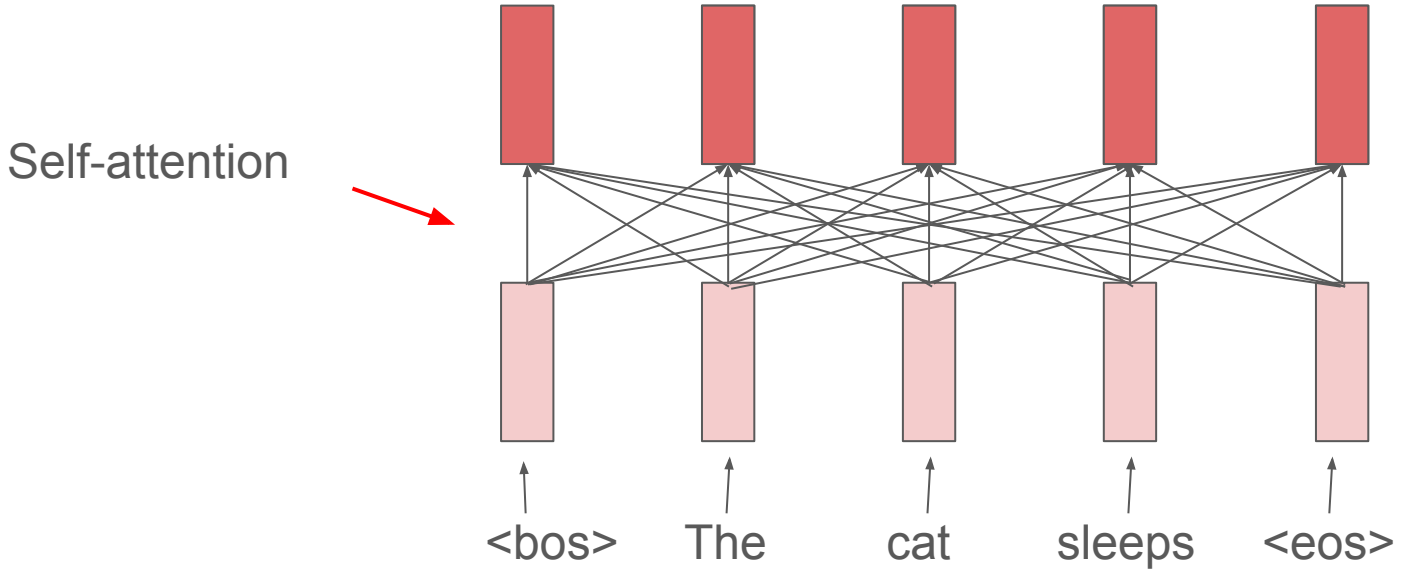
Self-attention



Self-attention



Self-attention



Multi-head attention

Она ест вкусную шаурму

- Case agreement
- Gender agreement

$$Multi-head\ attn = Concat(attn_1, attn_2, \ldots, attn_8)W$$

Attention (~ Transformer)

- No recurrent -> parallel encoding

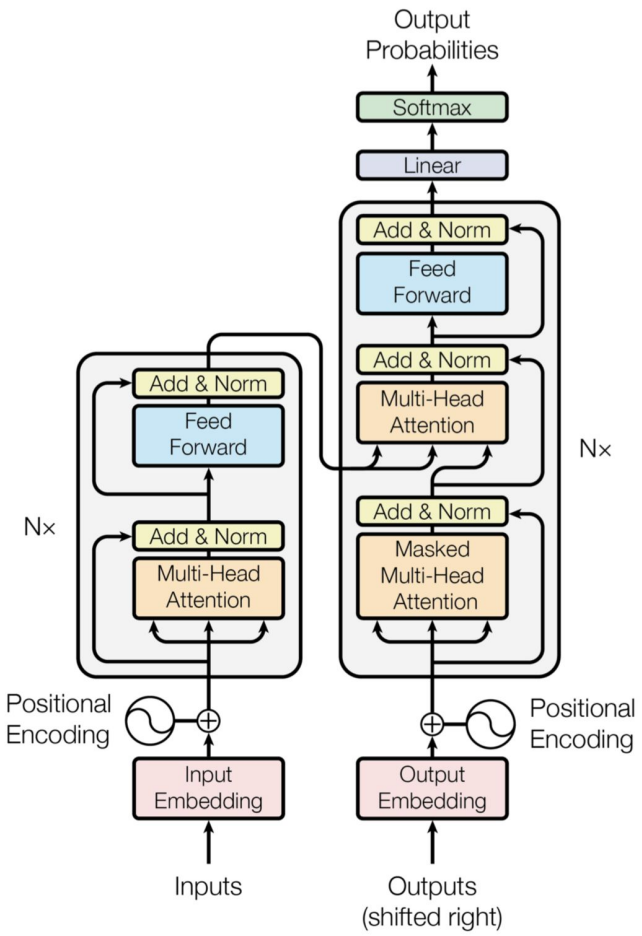
  -> faster

- Many attentions -> models does

  not have to remember much

- Multi-head attention -> able to

  pay attention to different aspects



https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Transformer

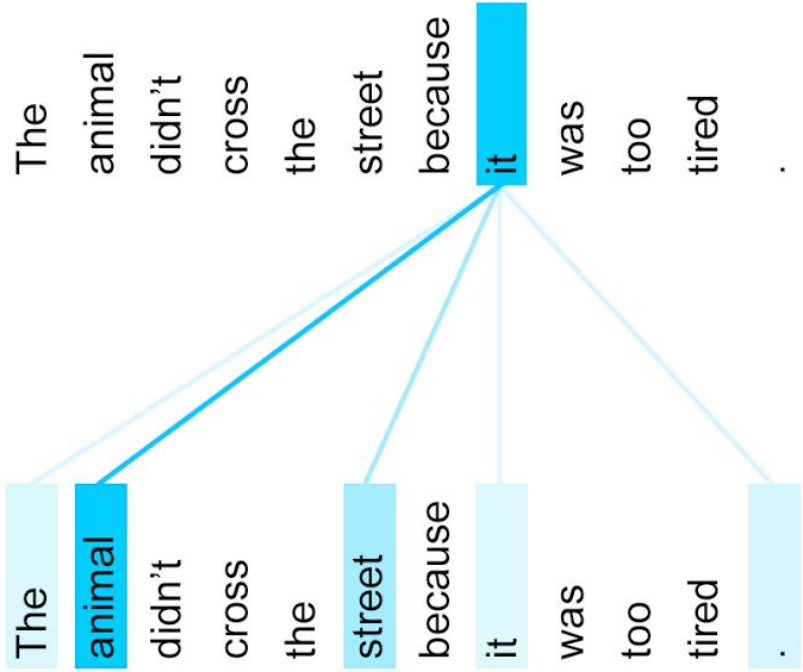
Transformer

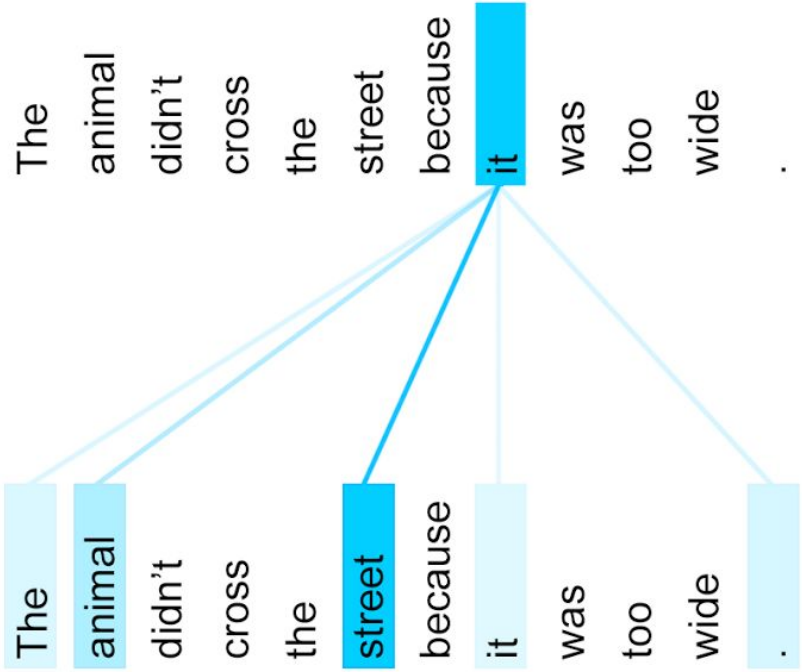
The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer





https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html