12

Attention

Long Short-term Memory (LSTM)

RNNs

Forget gate:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

Input gate:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

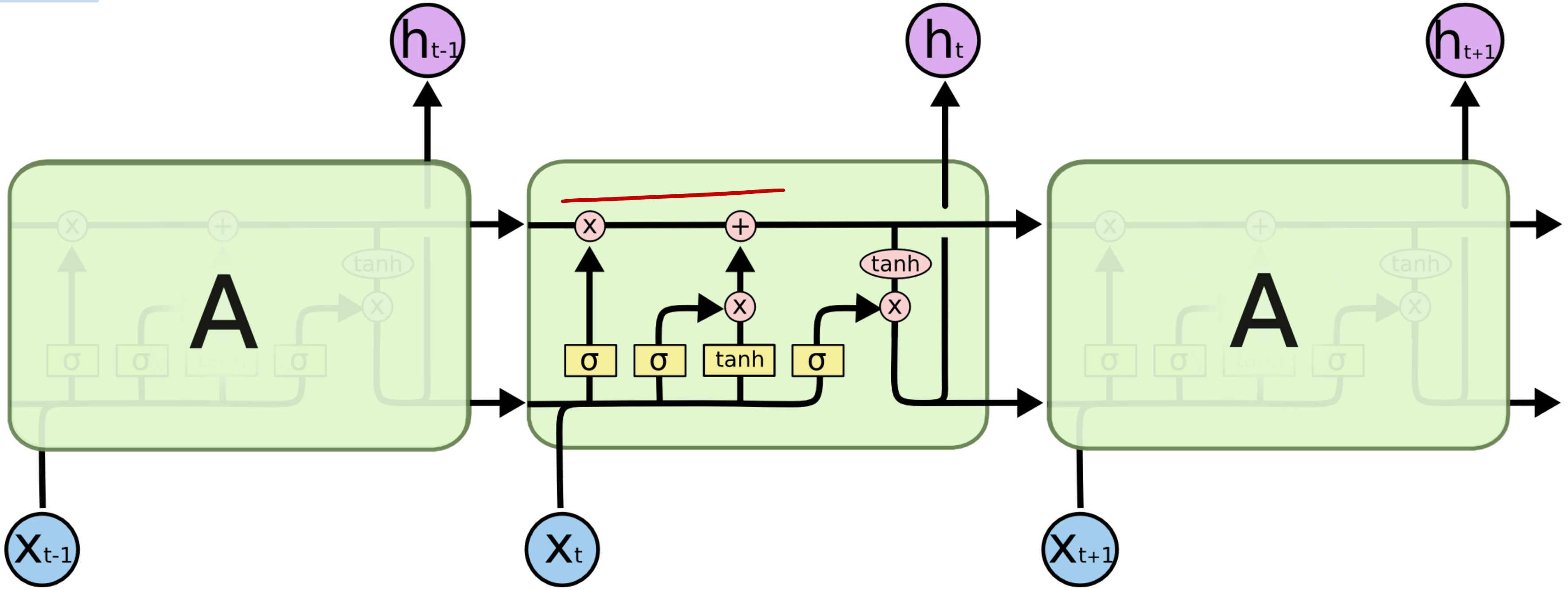
Cell update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

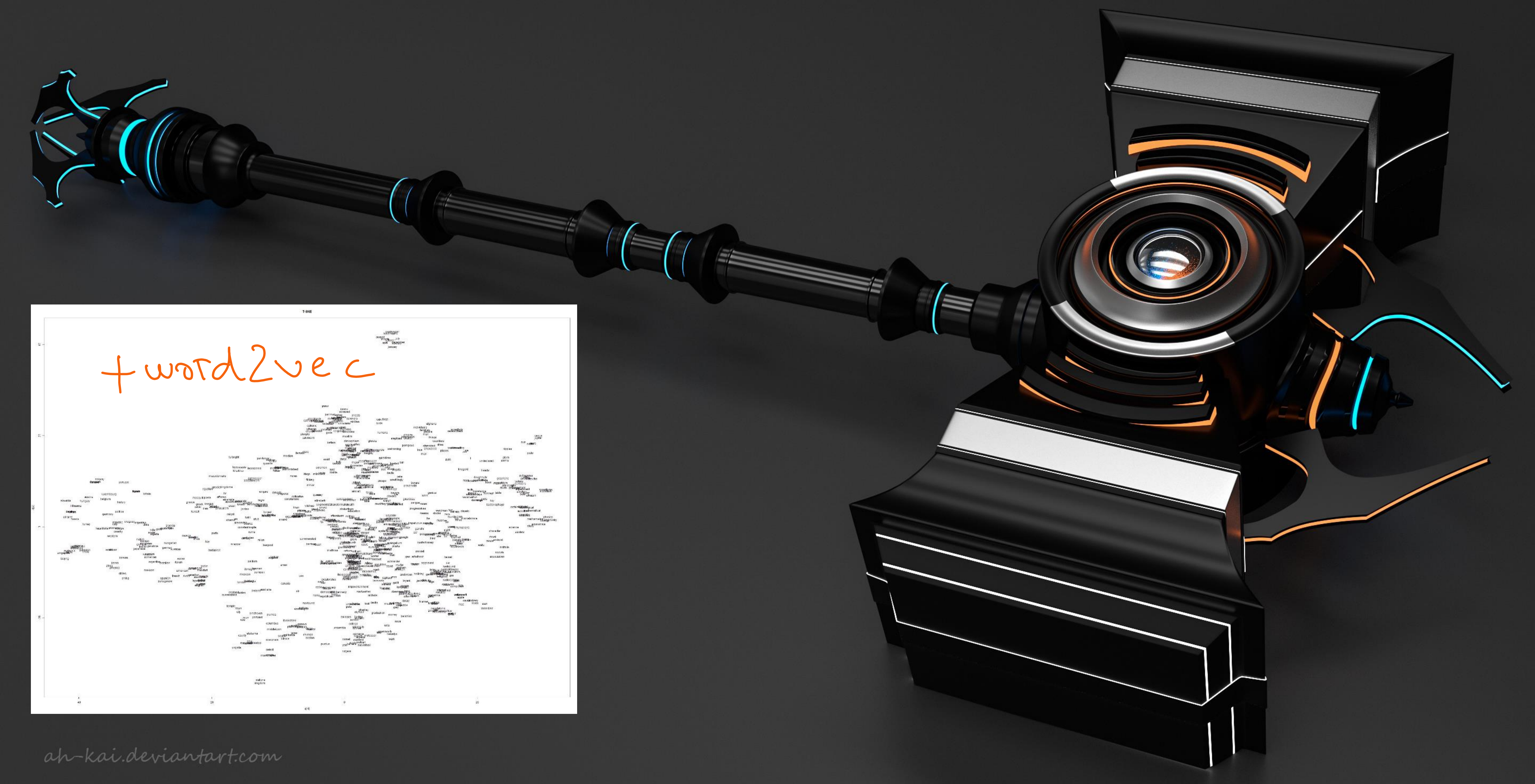
Output gate:

$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

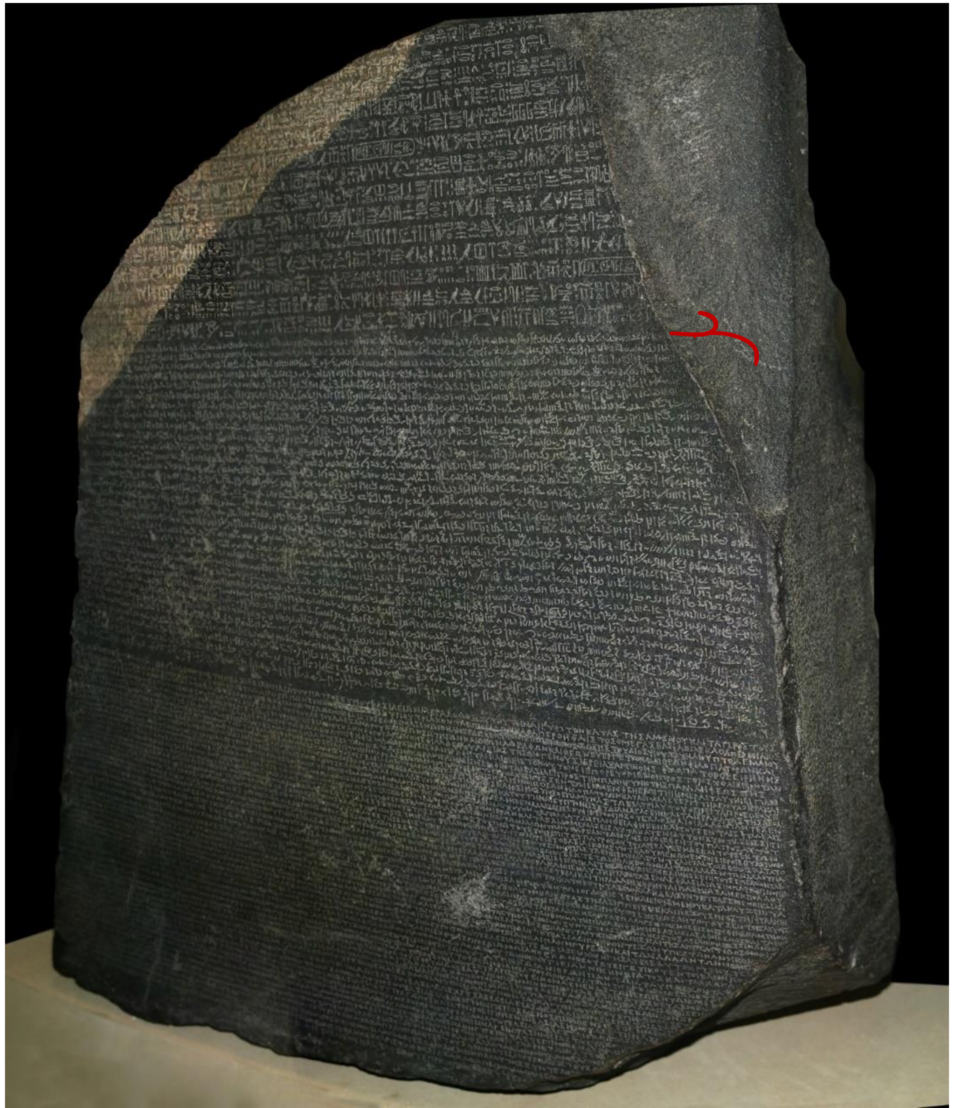


[Understanding LSTM Networks](#)


T-SNE
+word2vec
ah-kai.deviantart.com

Машинный перевод
Machine Translation

Rosetta Stone



[Wikipedia](#)

Русский → translation model → english bits → Language model

decoder

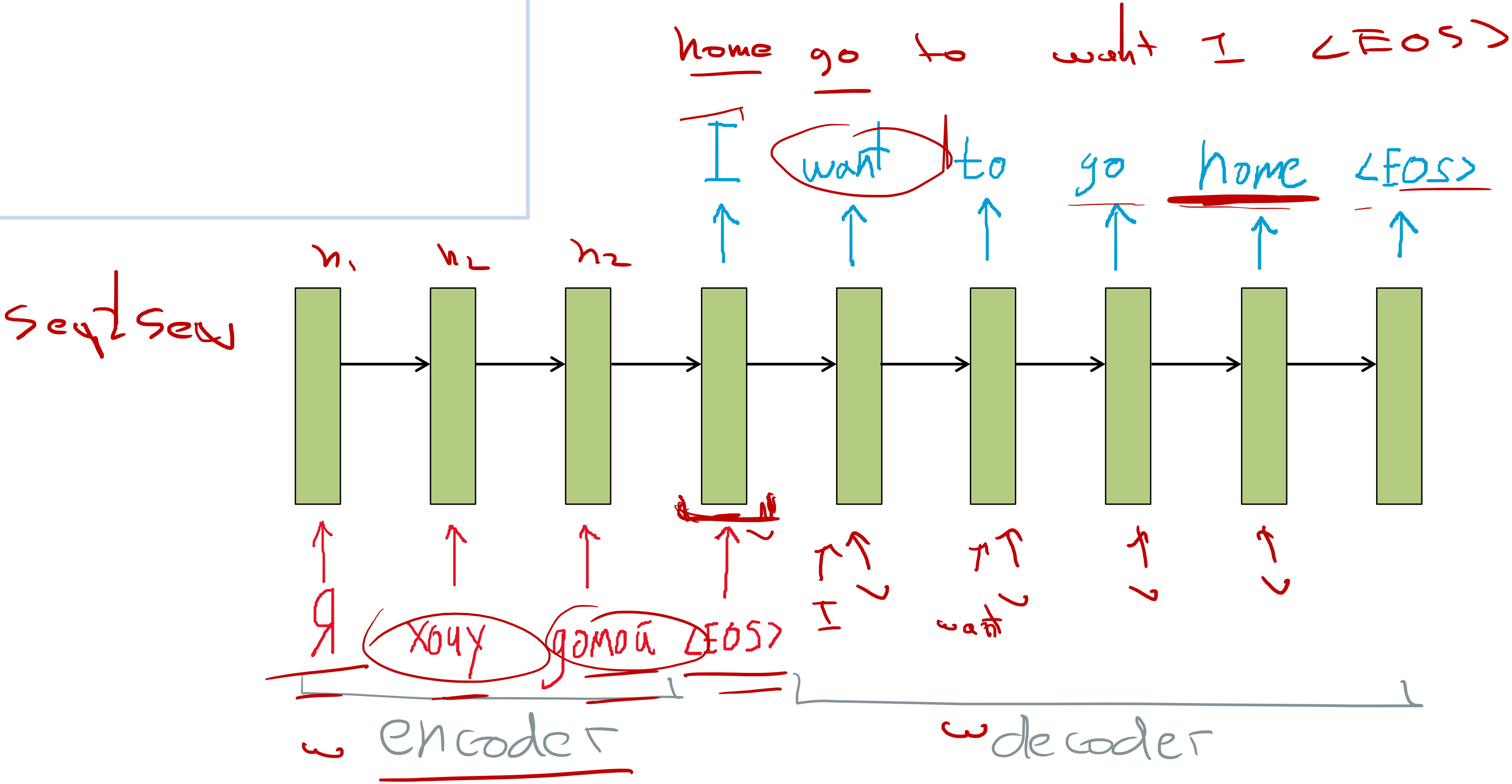
Good morning → Доброе утро

Goodbye → До свидания

Oh wow → Офигеть

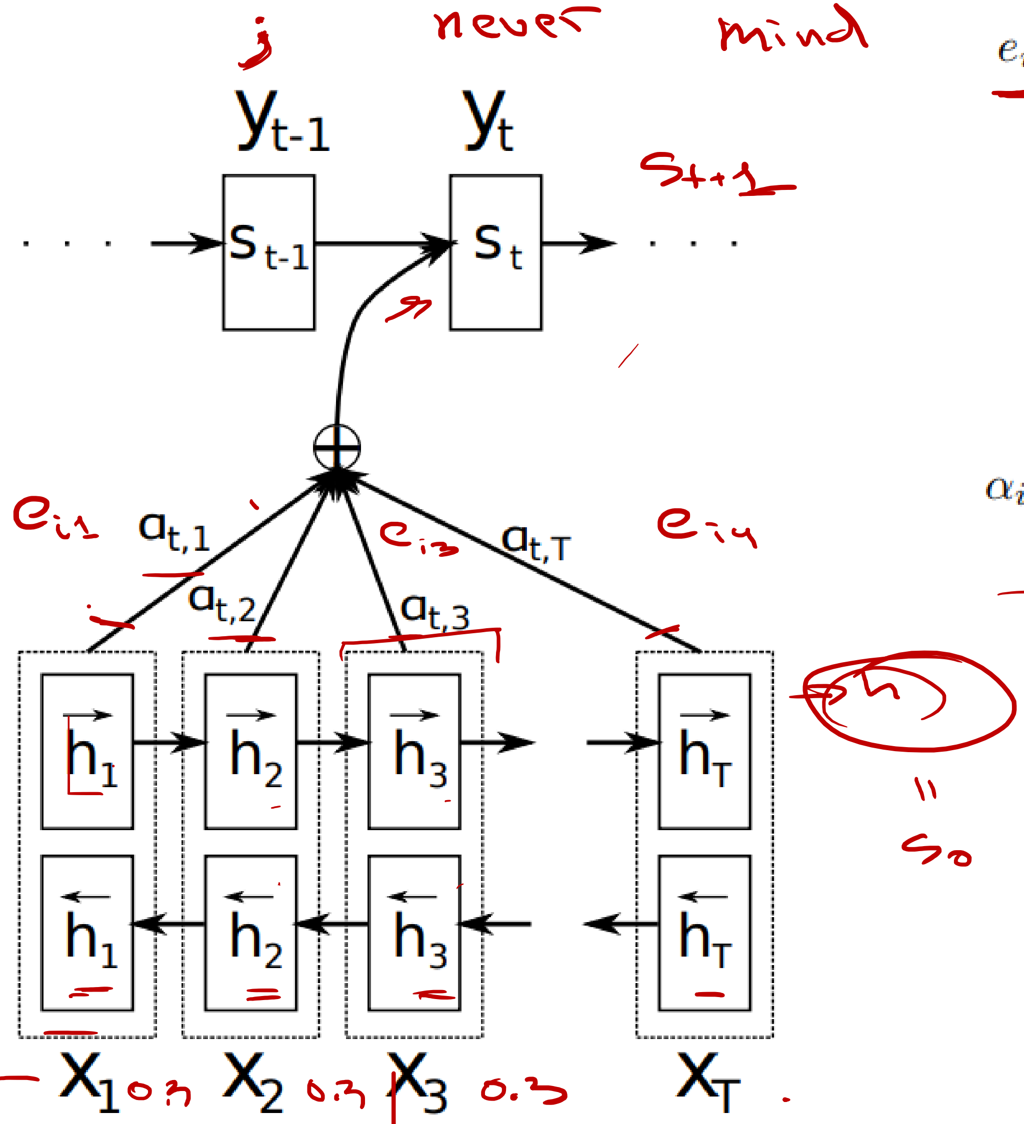
Never mind → Не обращай внимания

Sequence to Sequence



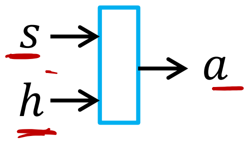
[Sequence to Sequence Learning with Neural Networks’14]

Attention



encoder

$$e_{ij} = a(s_{i-1}, h_j)$$

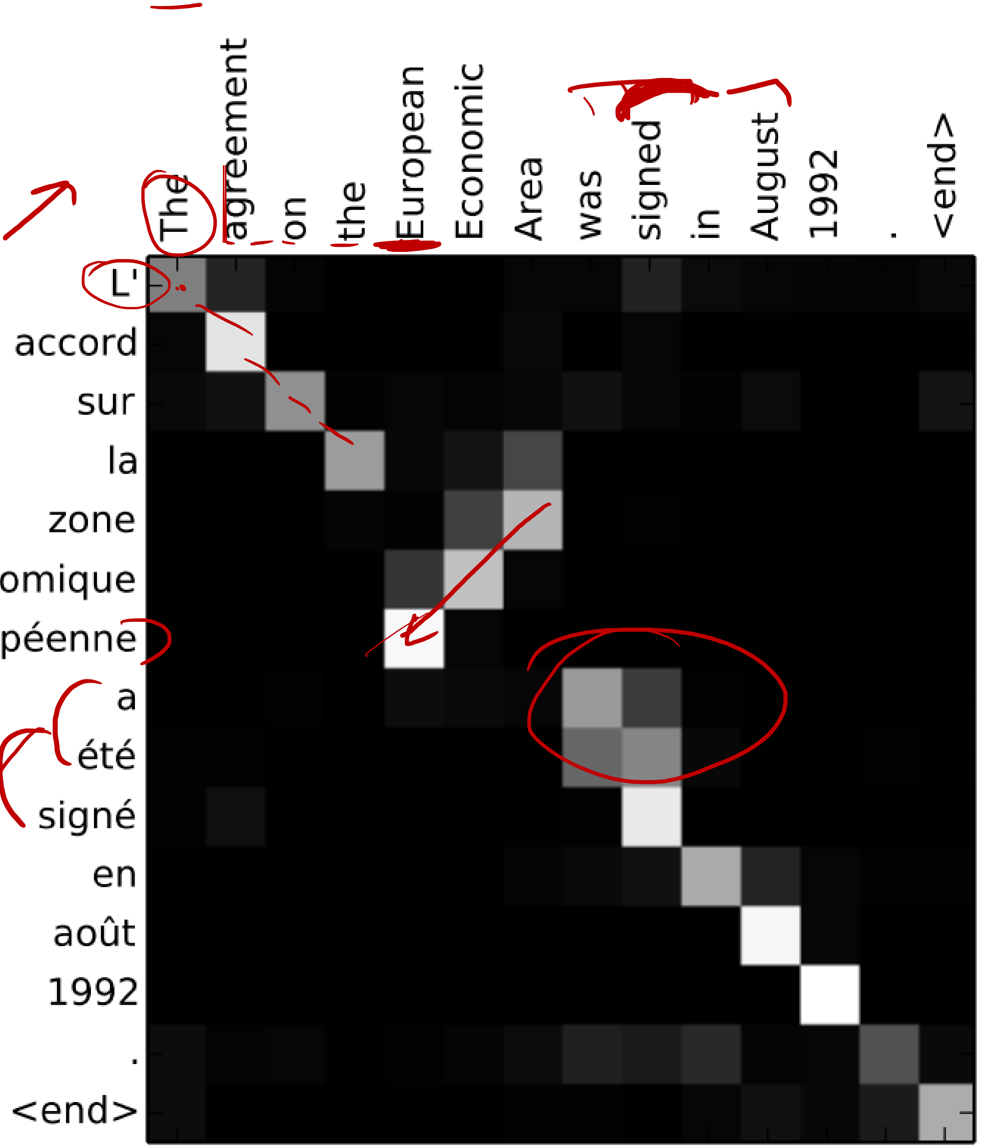


$$a(s, h) = s \cdot h$$

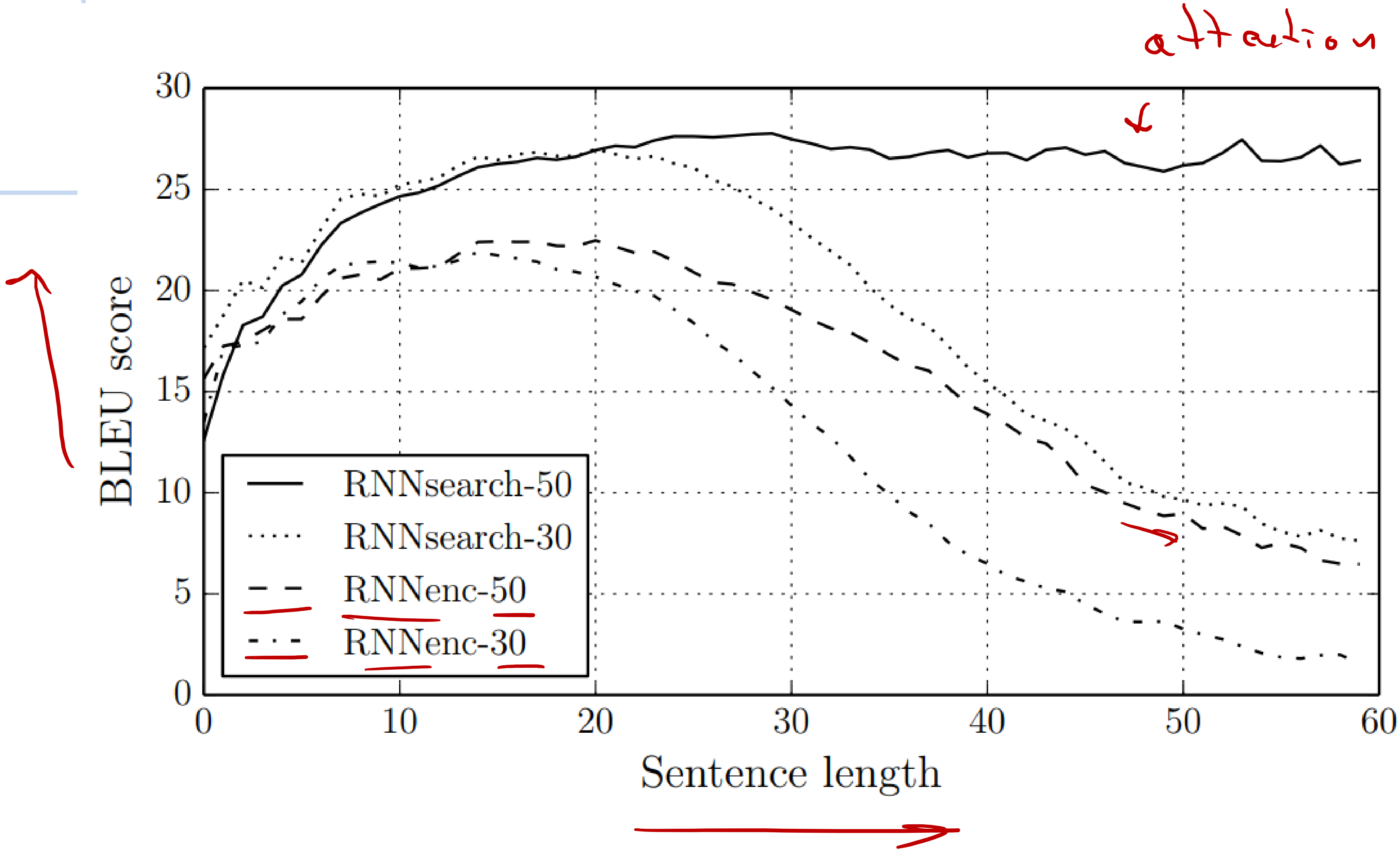
$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)},$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

Neural Machine Translation by Jointly Learning to Align and Translate'14



[Neural Machine Translation by Jointly Learning to Align and Translate’14]

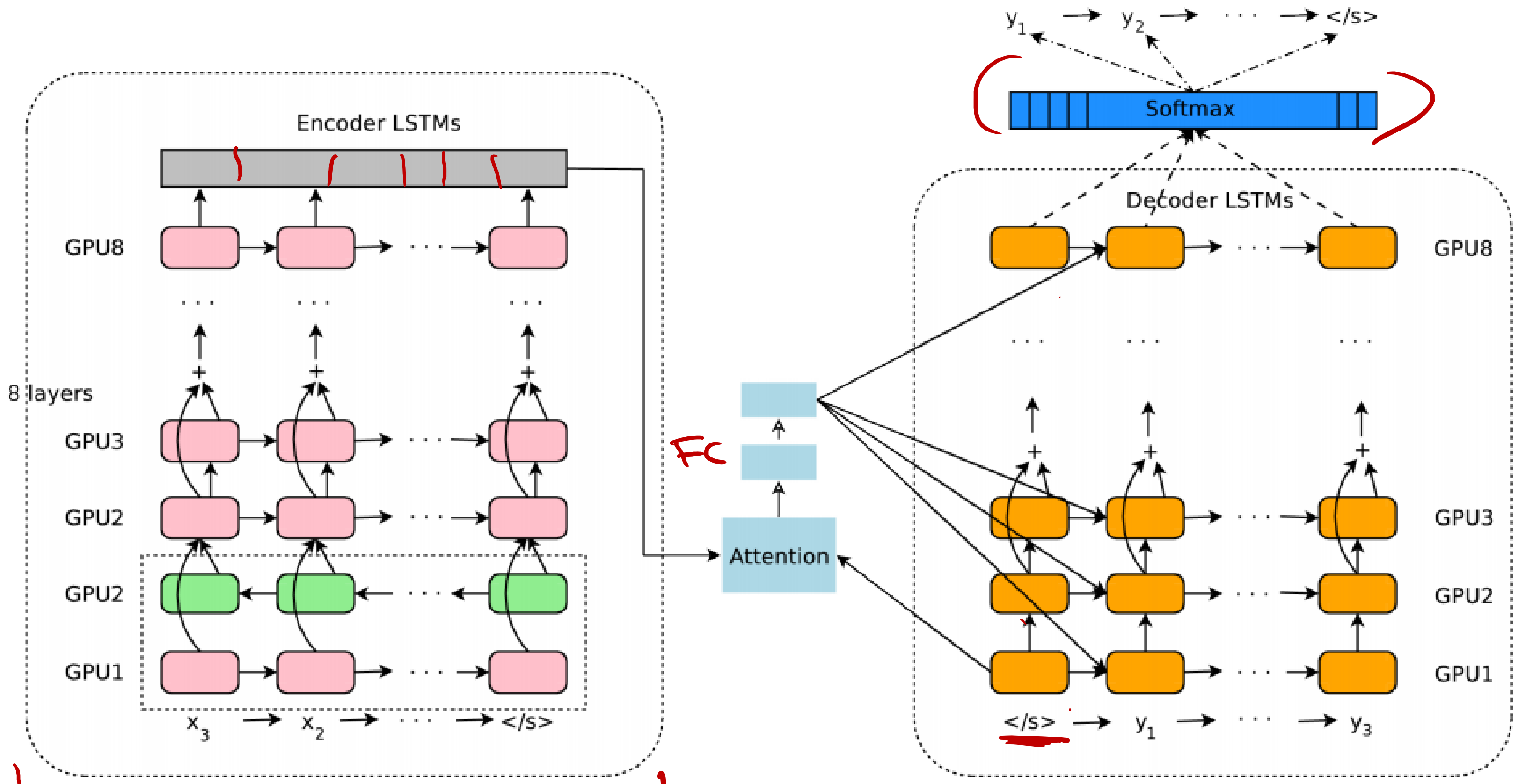


[Neural Machine Translation by Jointly Learning to Align and Translate'14](#)

more awesome stuff at ThumbPress.com

Google Translate

GOV



enc

- **Word**: Jet makers feud over seat width with big orders at stake
- **wordpieces**: _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

[Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'16]()

до

Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

после

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

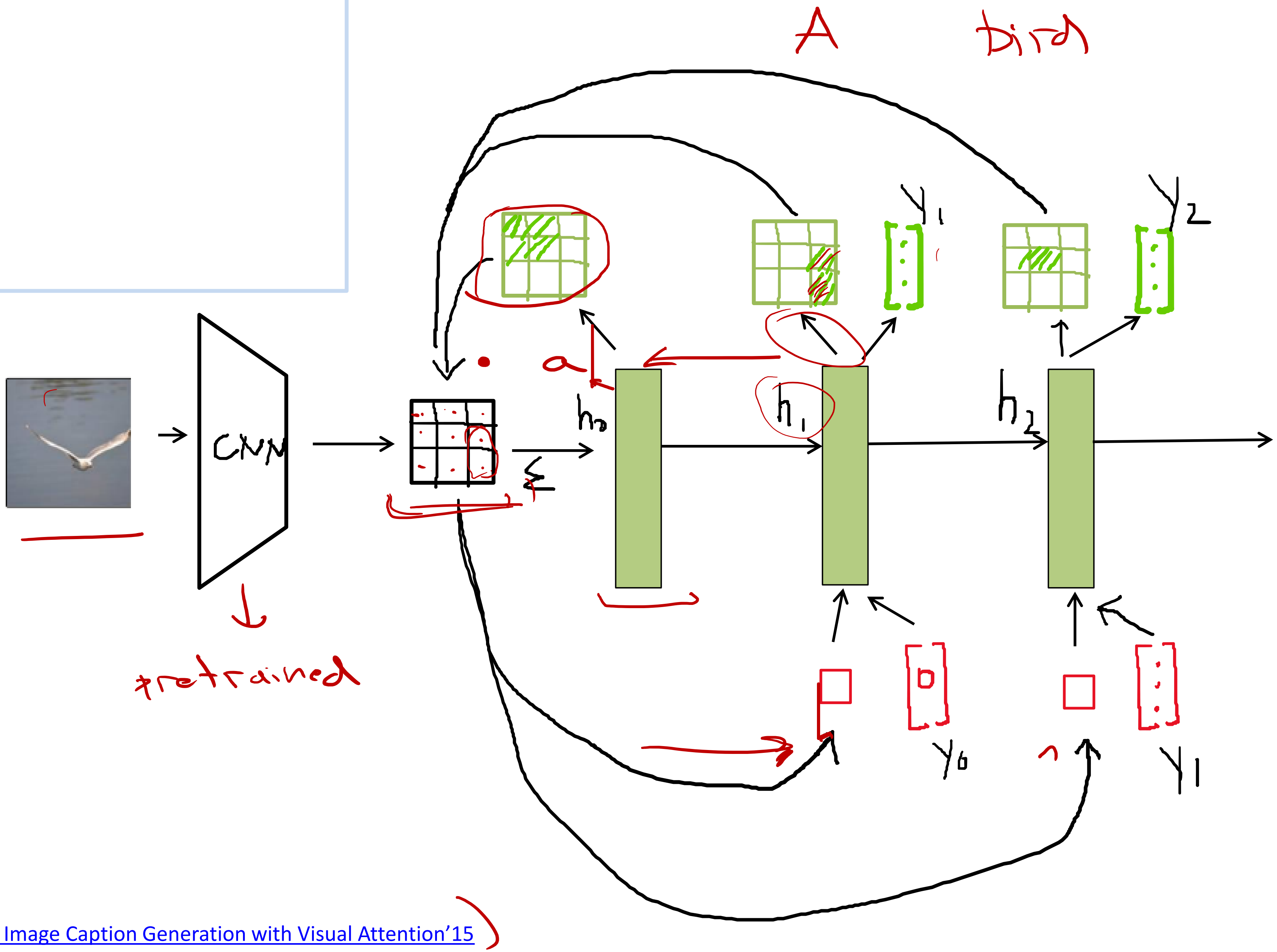
https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Image captioning

A bird is flying over a body of water.

A woman is throwing a frisbee in a park.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention'15



[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention'15]

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of
(top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

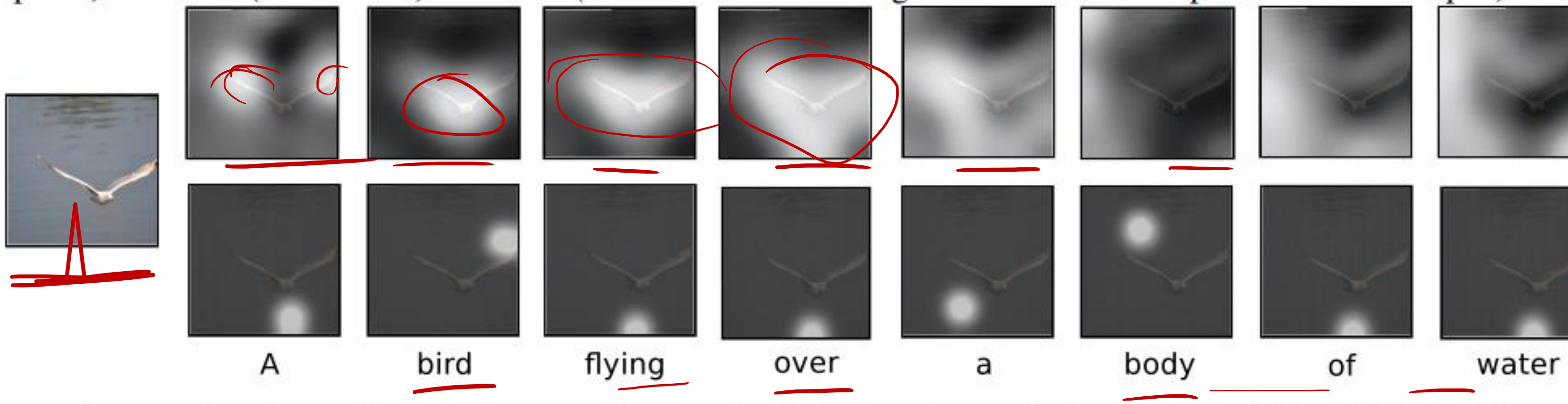
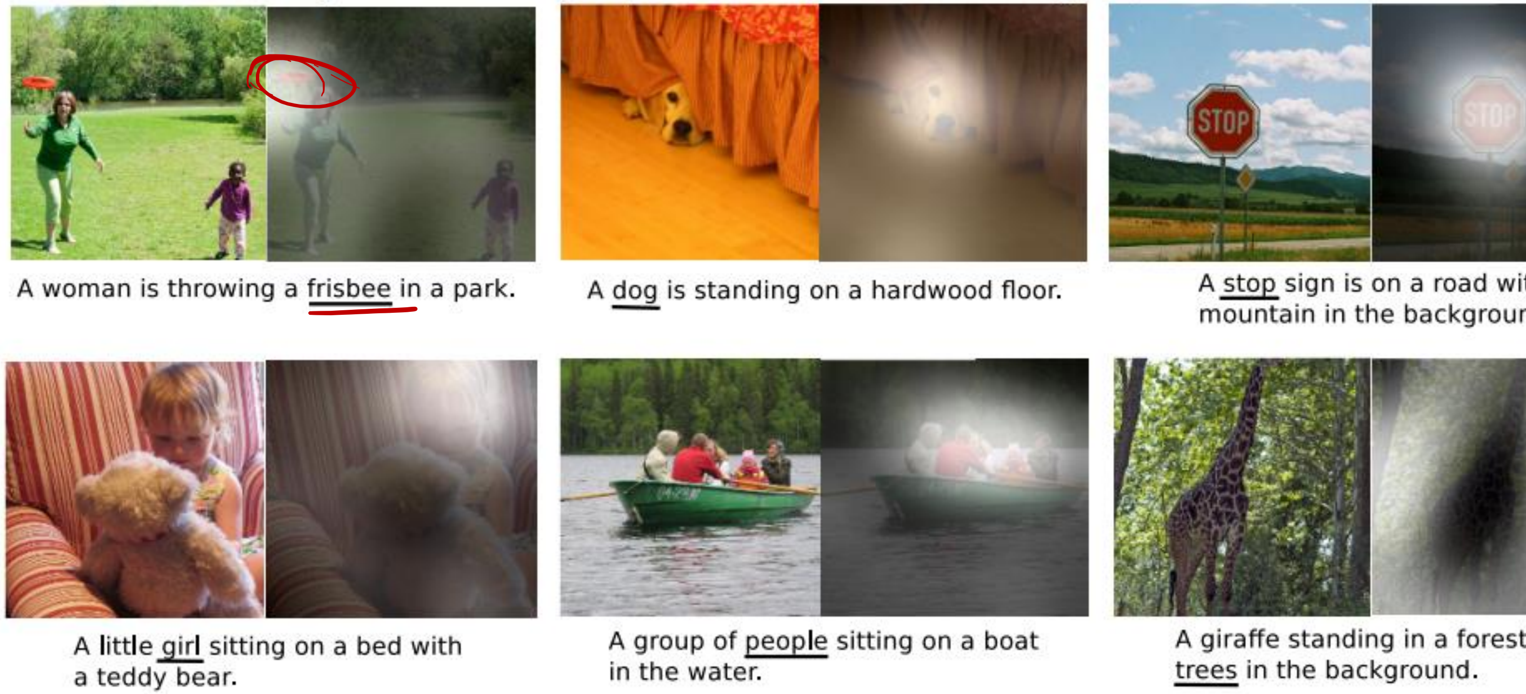


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the co



[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention'15]


THIS IS GREAT!
NOT ONLY AM I NOT LEARNIN',

Attention is All You Need

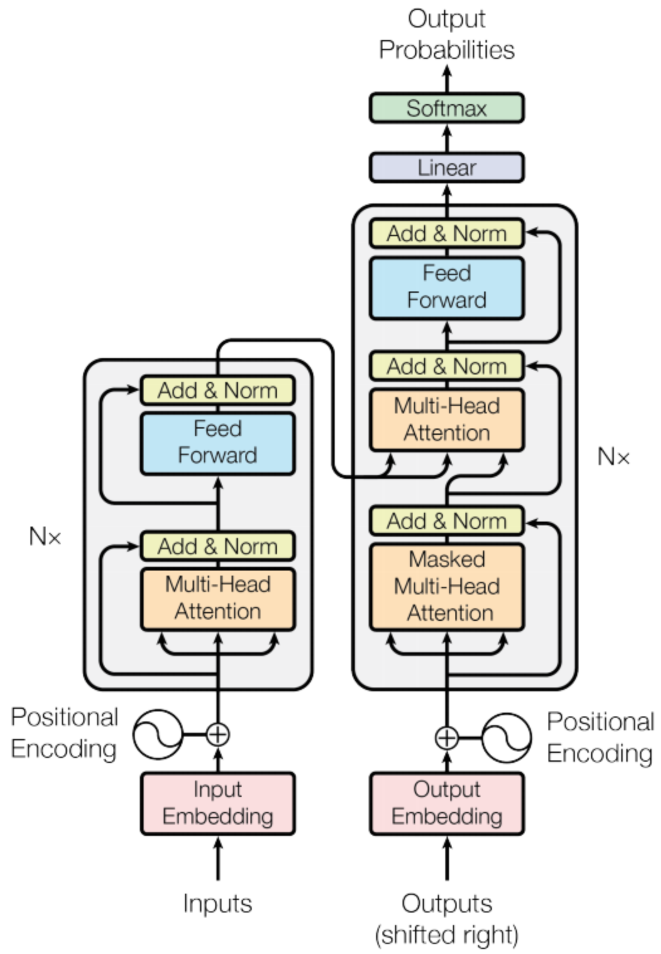


Figure 1: The Transformer - model architecture.

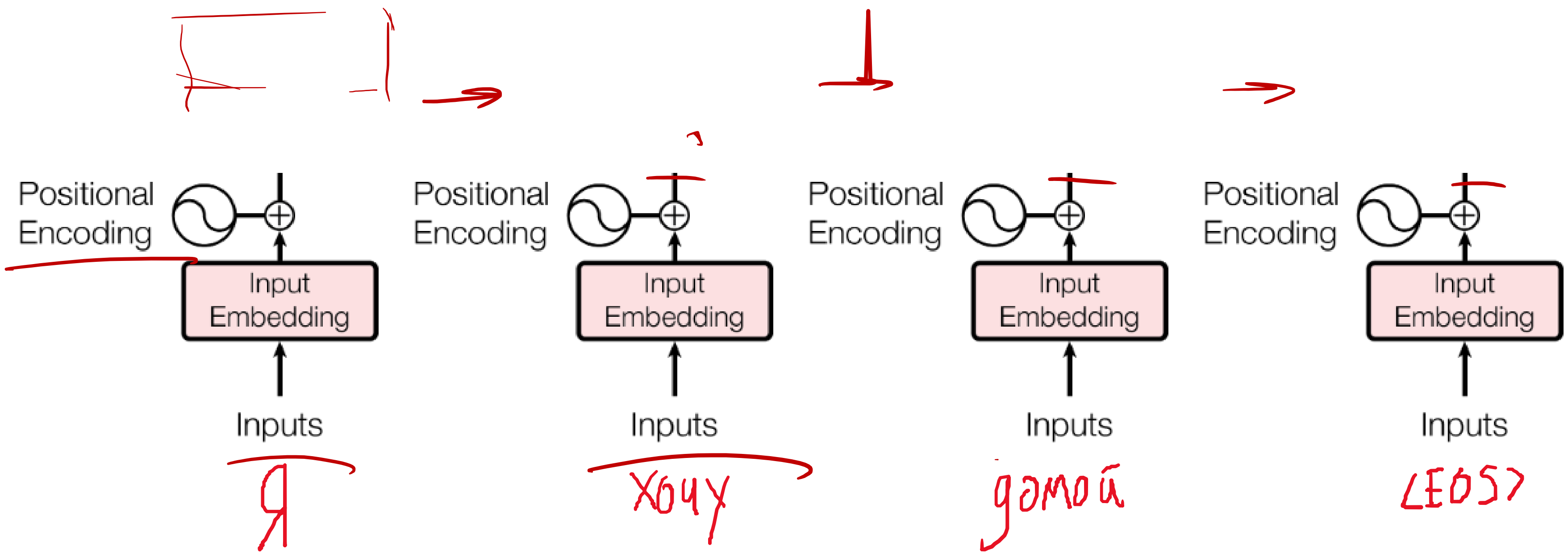
Вот тут я про него рассказываю

https://habrahabr.ru/post/341240/

[Attention Is All You Need’17]()

Transformer

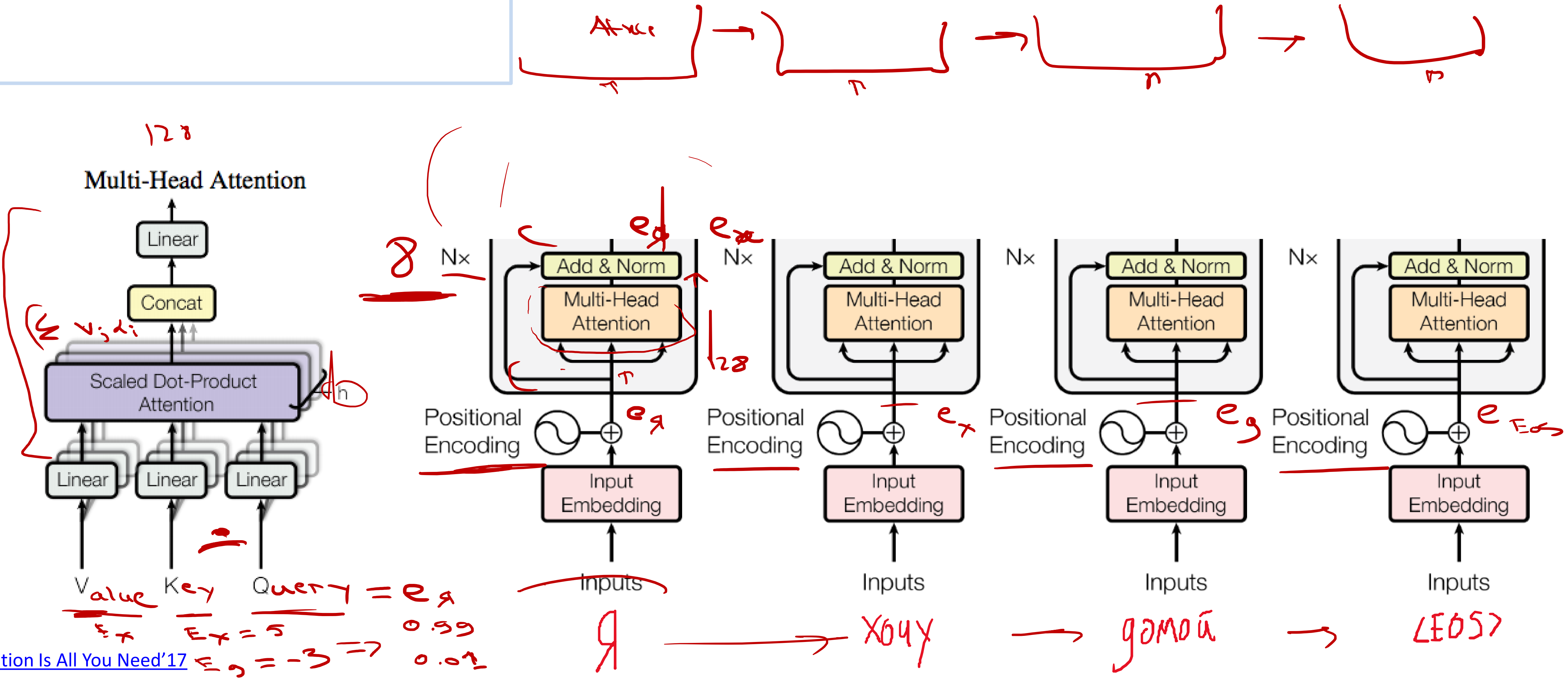
Encoder



[Attention Is All You Need'17](#)

Transformer

Encoder



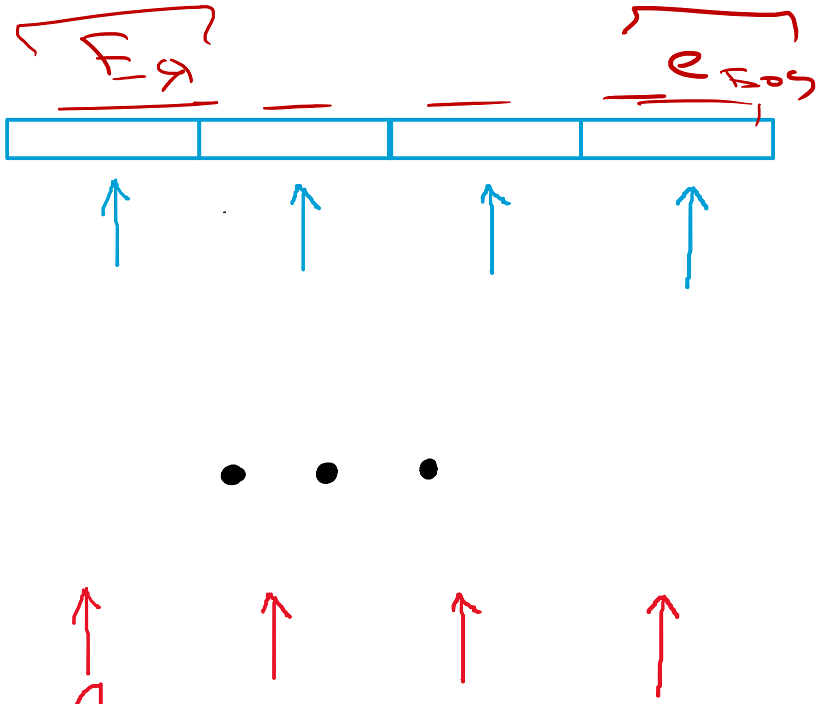
Attention Is All You Need’17

$n = 1$

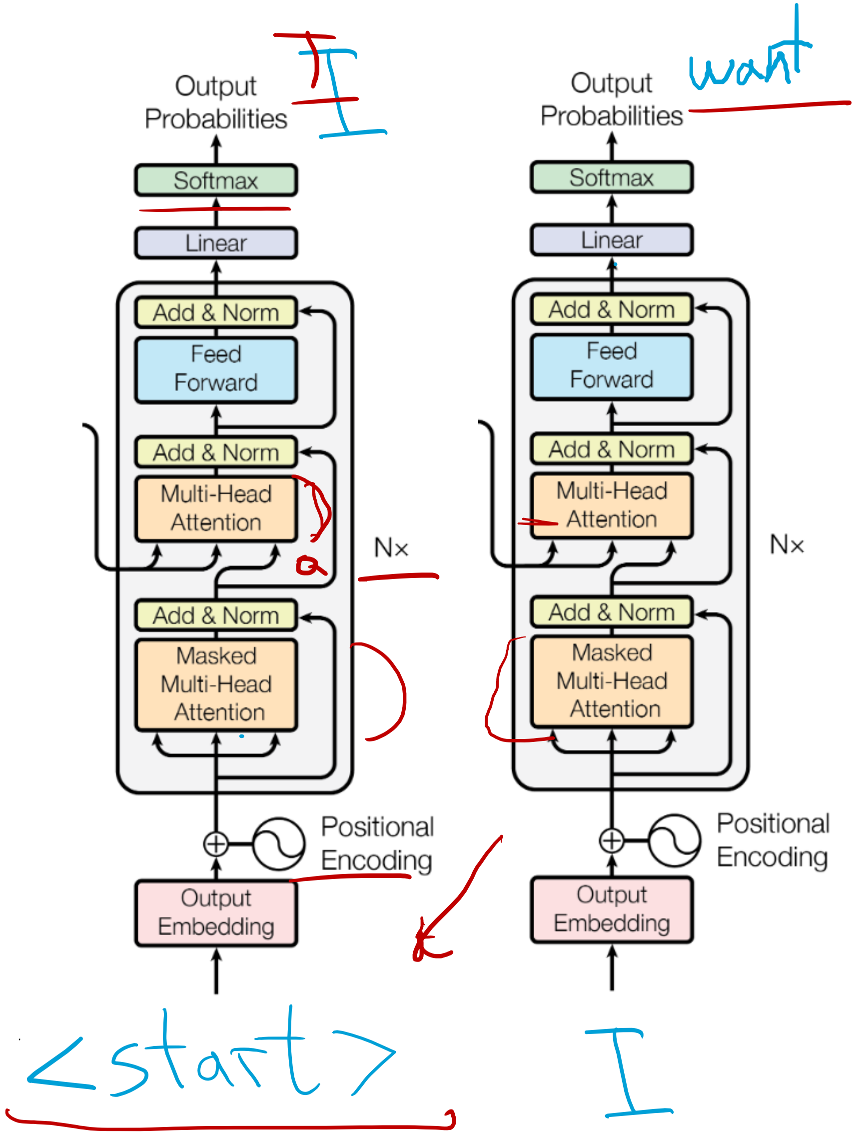
$n = 2$

Transformer

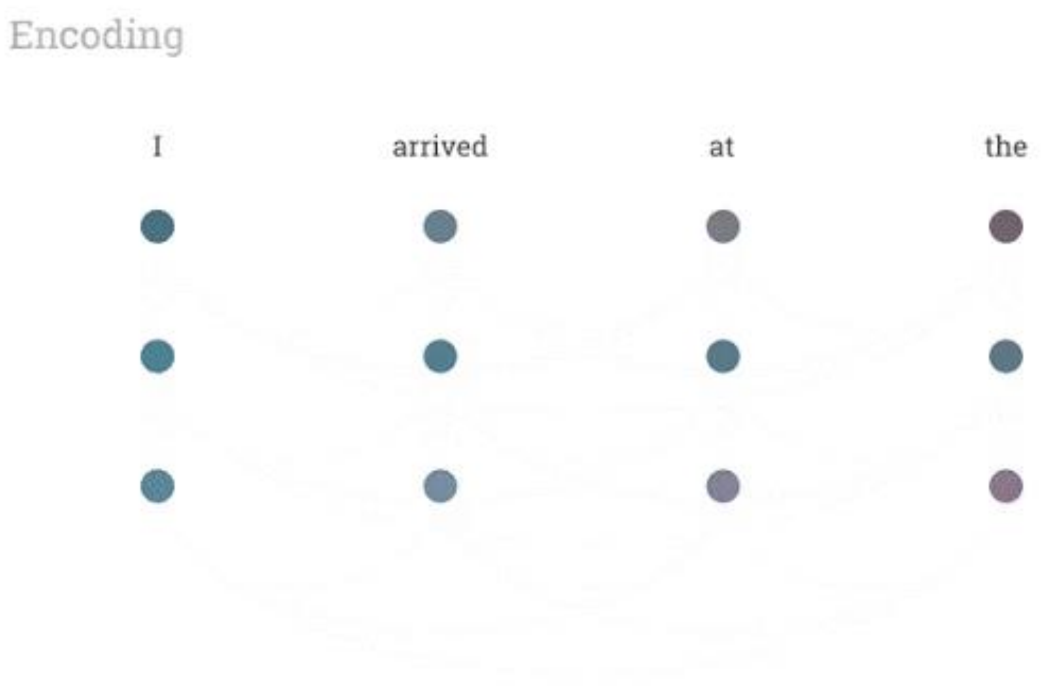
Decoder



Я хочу домой <EOS>



[Attention Is All You Need’17]

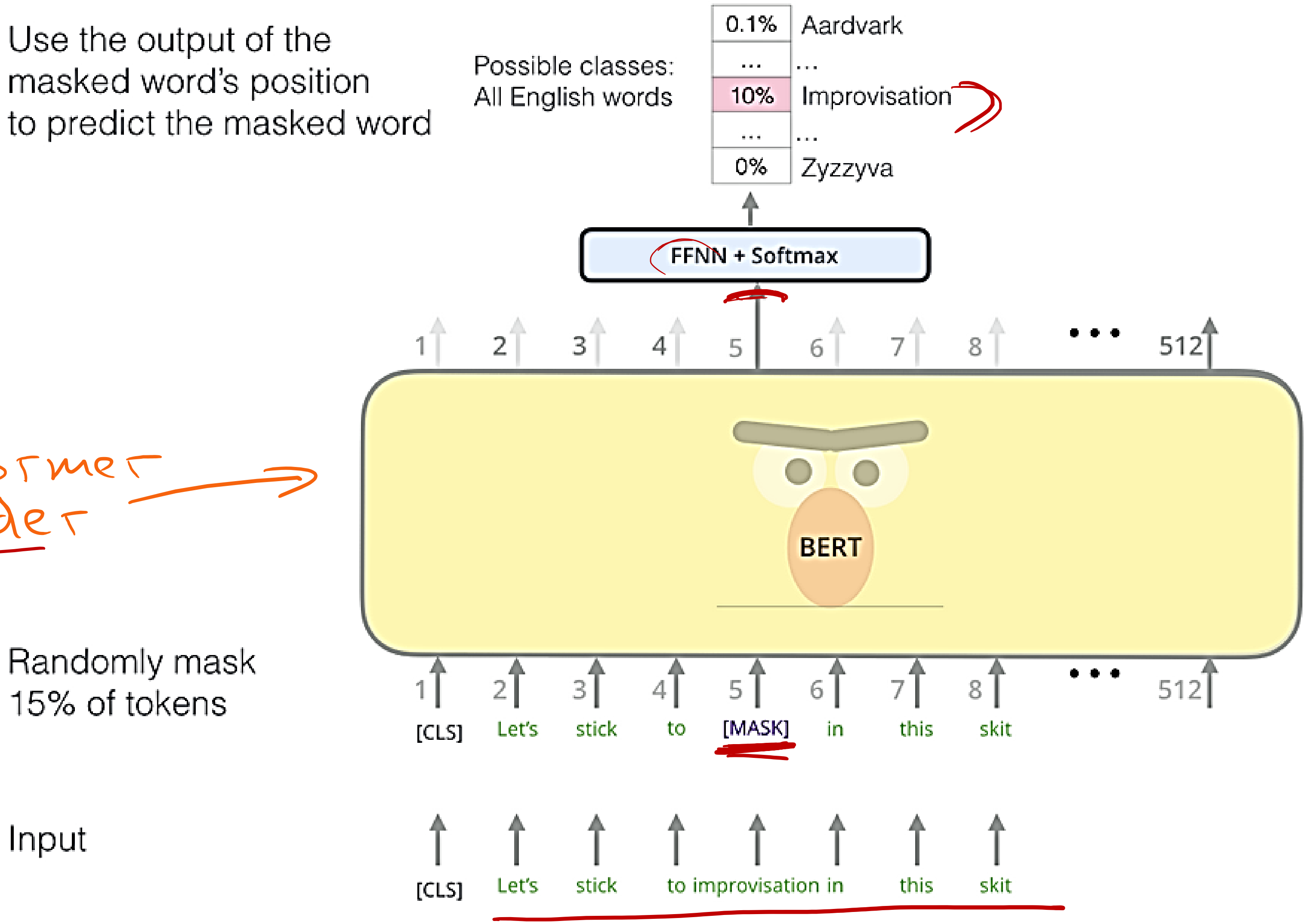


[Google Research Blog]


АРХИТЕКТУРА
ПРОСТО ТЕРМОЯД
imgflip.com

transformer encoder

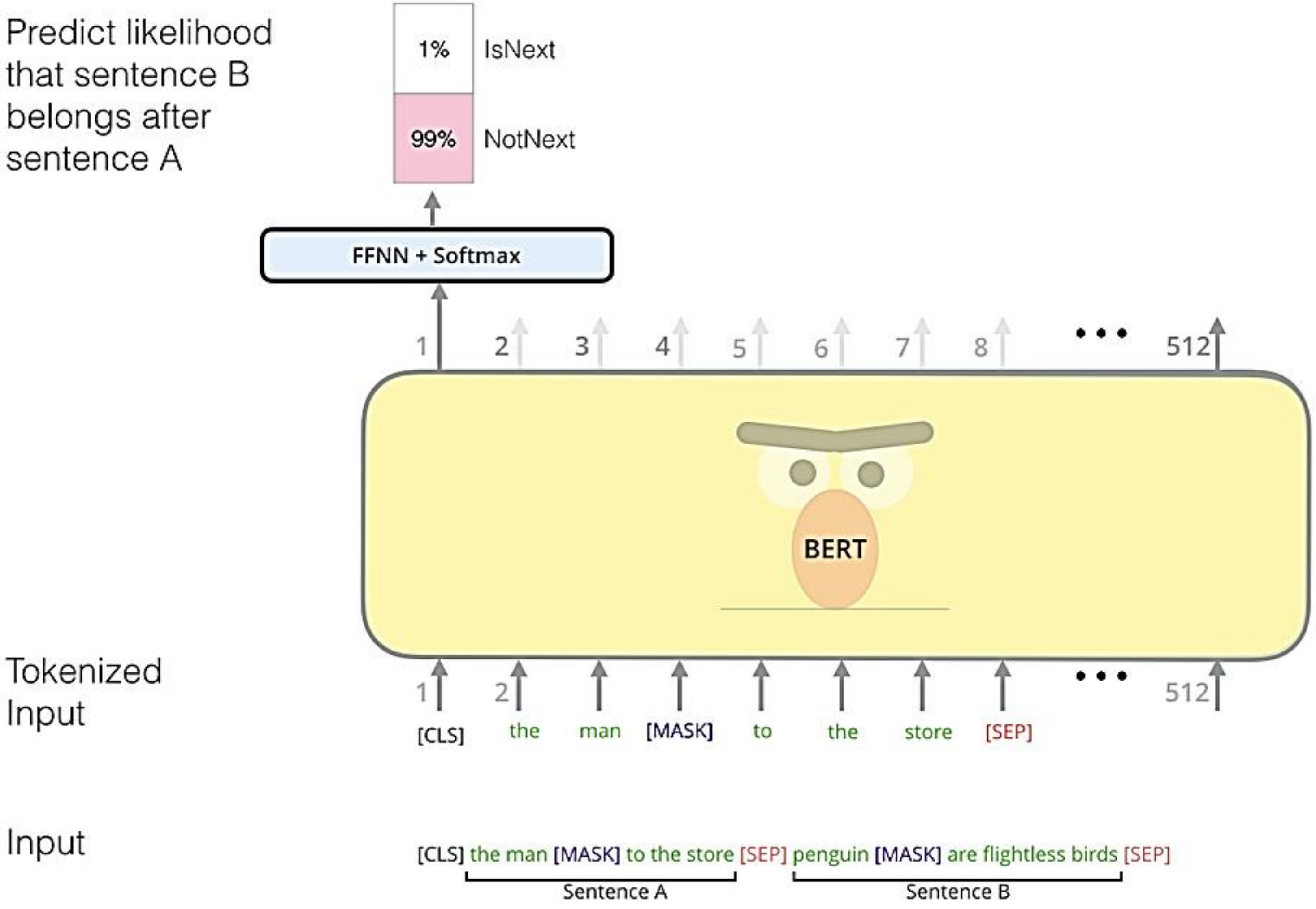
Encoder
320M parameters
24 transformer blocks



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Image source

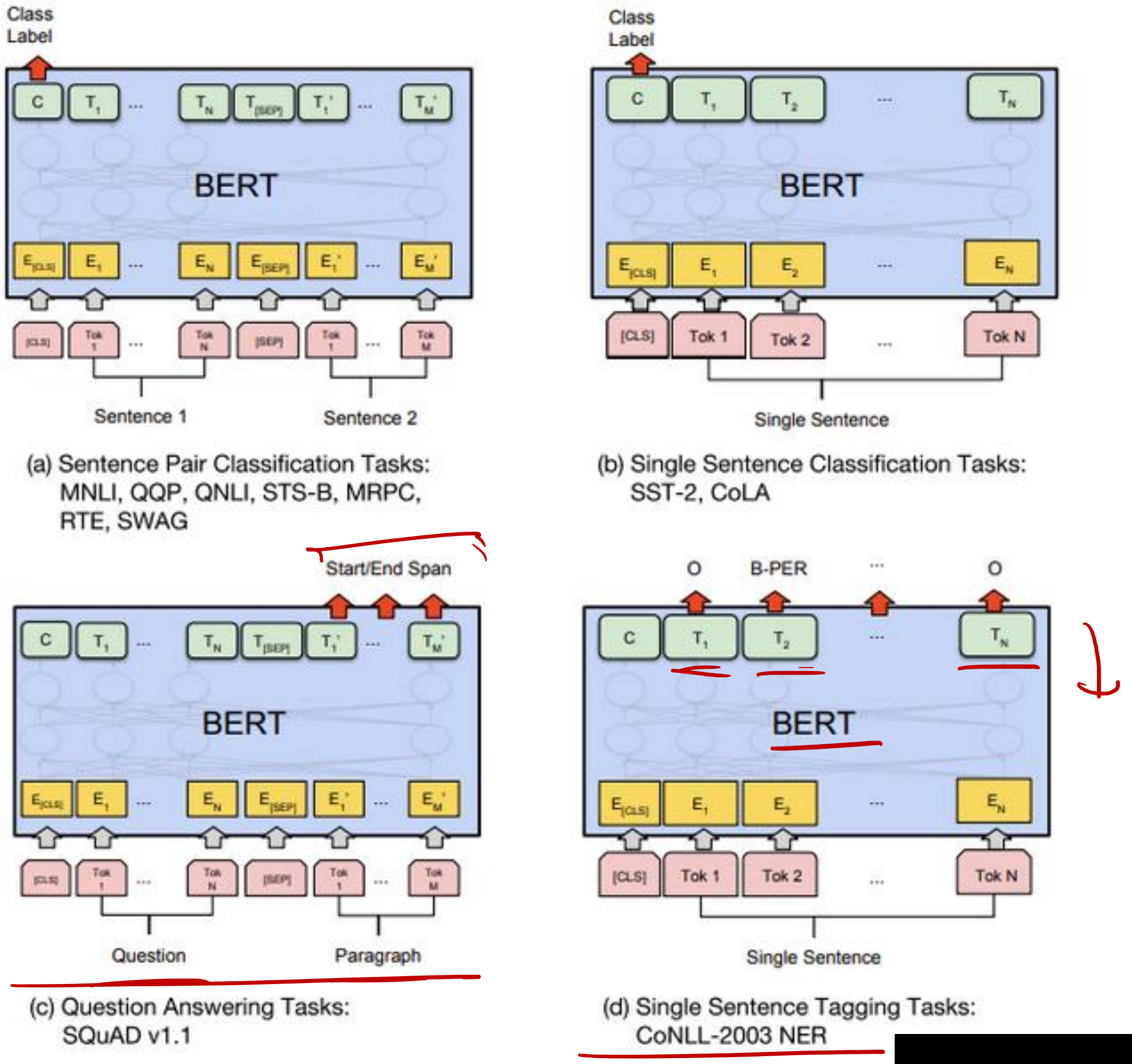
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'18]



The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

[Image source](#)

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’18](#)



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

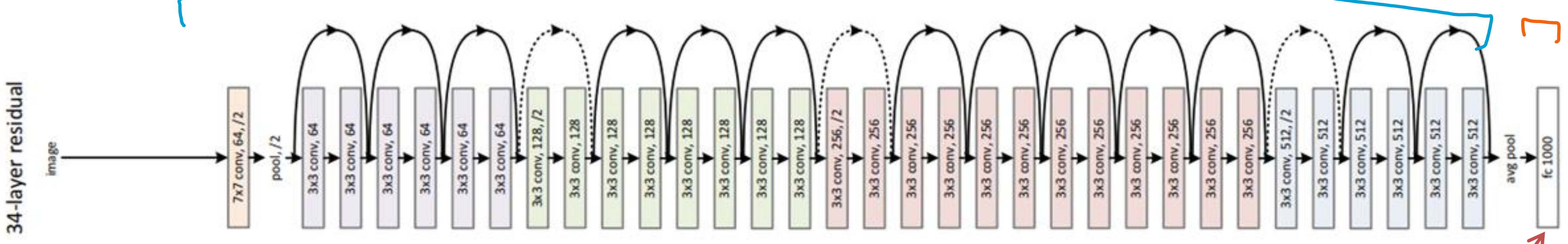
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'18]

Transfer learning в NLP!

freeze

train



Линейный классификатор

$BERT_{LARGE}$

Question Answering

Because of the complexity of medications including specific indications, effectiveness of treatment regimens, safety of medications (i.e., drug interactions) and patient compliance issues (in the hospital and at home) many pharmacists practicing in hospitals gain more education and training after pharmacy school through a pharmacy practice residency and sometimes followed by another residency in a specific area. Those pharmacists are often referred to as clinical pharmacists and they often specialize in various disciplines of pharmacy. For example, there are pharmacists who specialize in hematology/oncology, HIV/AIDS, infectious disease, critical care, emergency medicine, toxicology, nuclear pharmacy, pain management, psychiatry, anti-coagulation clinics, herbal medicine, neurology/epilepsy management, pediatrics, neonatal pharmacists and more.

Where do pharmacists acquire more preparation following pharmacy school?
Ground Truth Answers: a pharmacy practice residency | pharmacy practice residency | pharmacy practice residency

What do clinical pharmacists specialize in?
Ground Truth Answers: various disciplines of pharmacy | various disciplines of pharmacy | various disciplines of pharmacy

What is one issue that adds to the complexity of a pharmacist's job?
Ground Truth Answers: effectiveness of treatment regimens | effectiveness of treatment regimens | effectiveness of treatment regimens

Which pharmacists are likely to seek additional education following pharmacy school?
Ground Truth Answers: pharmacists practicing in hospitals | pharmacists practicing in hospitals | clinical pharmacists

Where do pharmacists not go following pharmacy school?

[SQUAD 2.0]

Question Answering

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122

SQUAD 2.0

Language Modeling

1B Words / Google Billion Word benchmark

The One-Billion Word benchmark is a large dataset derived from a news-commentary site. The dataset consists of 829,250,940 tokens over a vocabulary of 793,471 words. Importantly, sentences in this model are shuffled and hence context is limited.

Model	Test perplexity	Number of params	Paper / Source	Code
Transformer-XL Large (Dai et al., 2018) ***under review***	21.8	0.8B	Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context	Official
Transformer-XL Base (Dai et al., 2018) ***under review***	23.5	0.46B	Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context	Official
Transformer with shared adaptive embeddings - Very large (Baevski and Auli, 2018)	23.7	0.8B	Adaptive Input Representations for Neural Language Modeling	Link
10 LSTM+CNN inputs + SNM10-SKIP (Jozefowicz et al., 2016) ***ensemble***	23.7	43B?	Exploring the Limits of Language Modeling	Official

nlpprogress.com

OpenAI GPT-2'19

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

BACK TO THE PRESENT