Attention is all you need

Denis Volkhonskiy

Motivation

- Attention is **focusing on specific parts of the input**.
- Many animals focus on specific parts of visual inputs to compute the responses
- Let’s include such mechanism to Deep Neural models

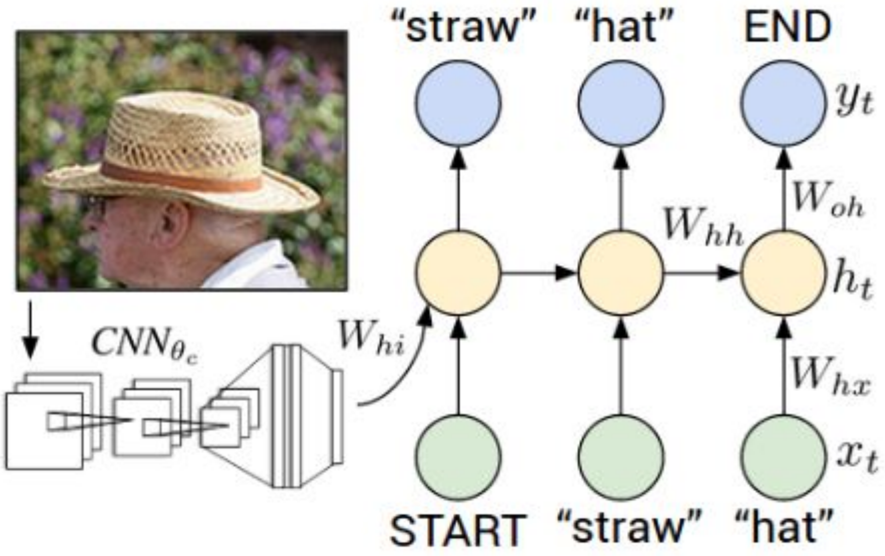
Image captioning task

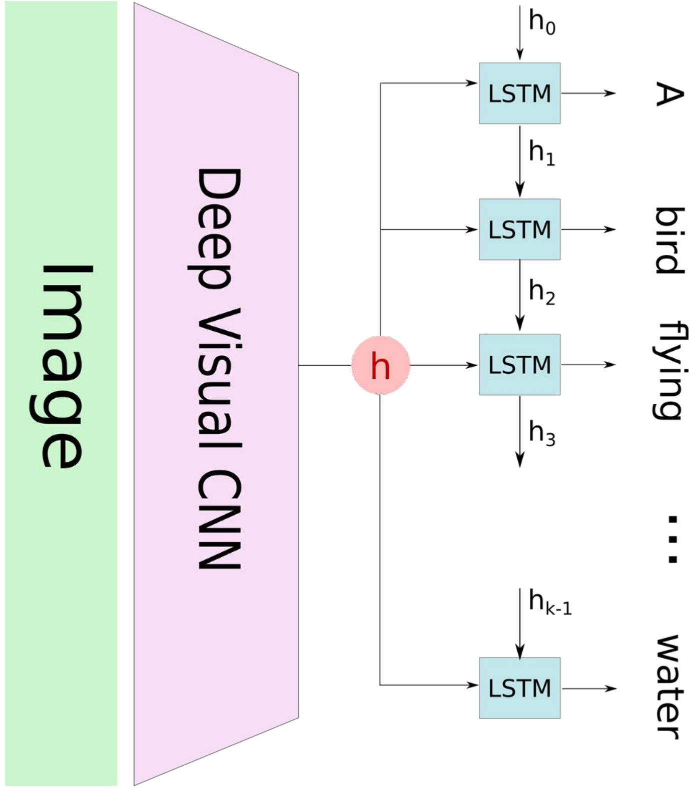
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

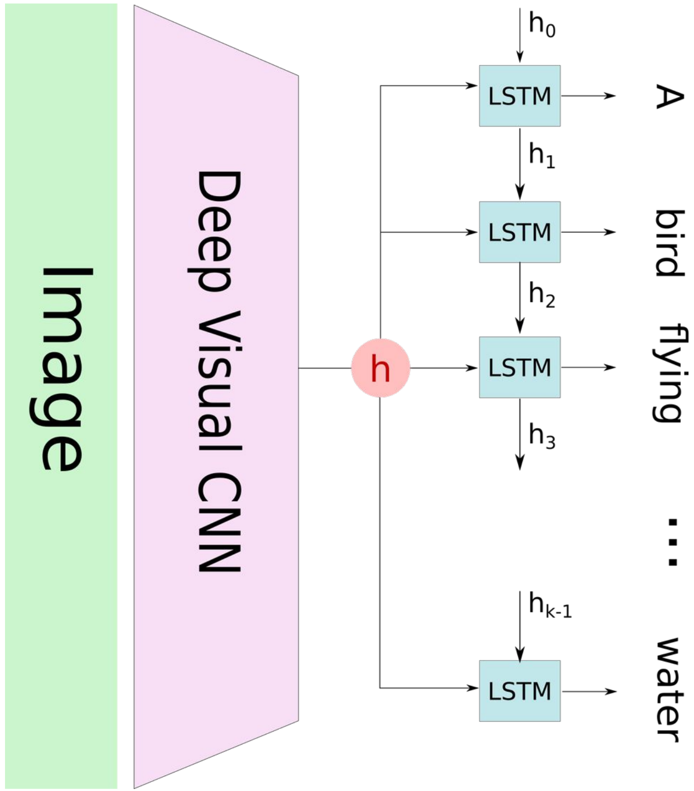
Image captioning task





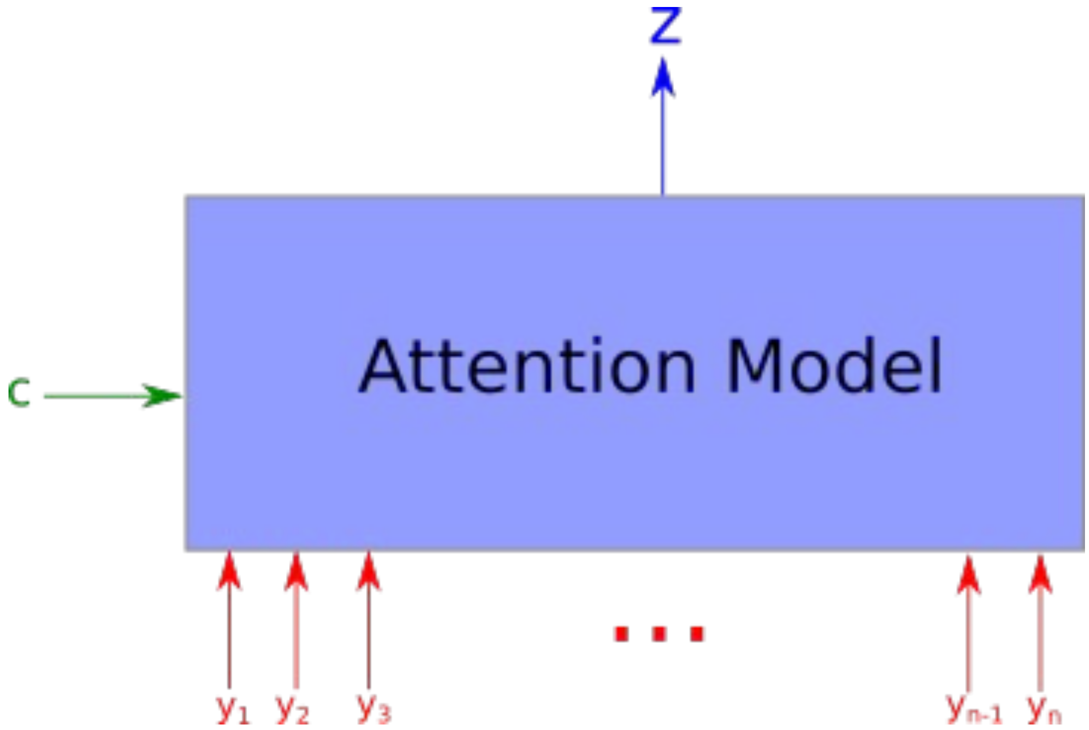
Problem

- At each iteration we generate one word
- Each word describe only a part of the image
- But we use the hole image representation h as condition for generation
- Attention will help!

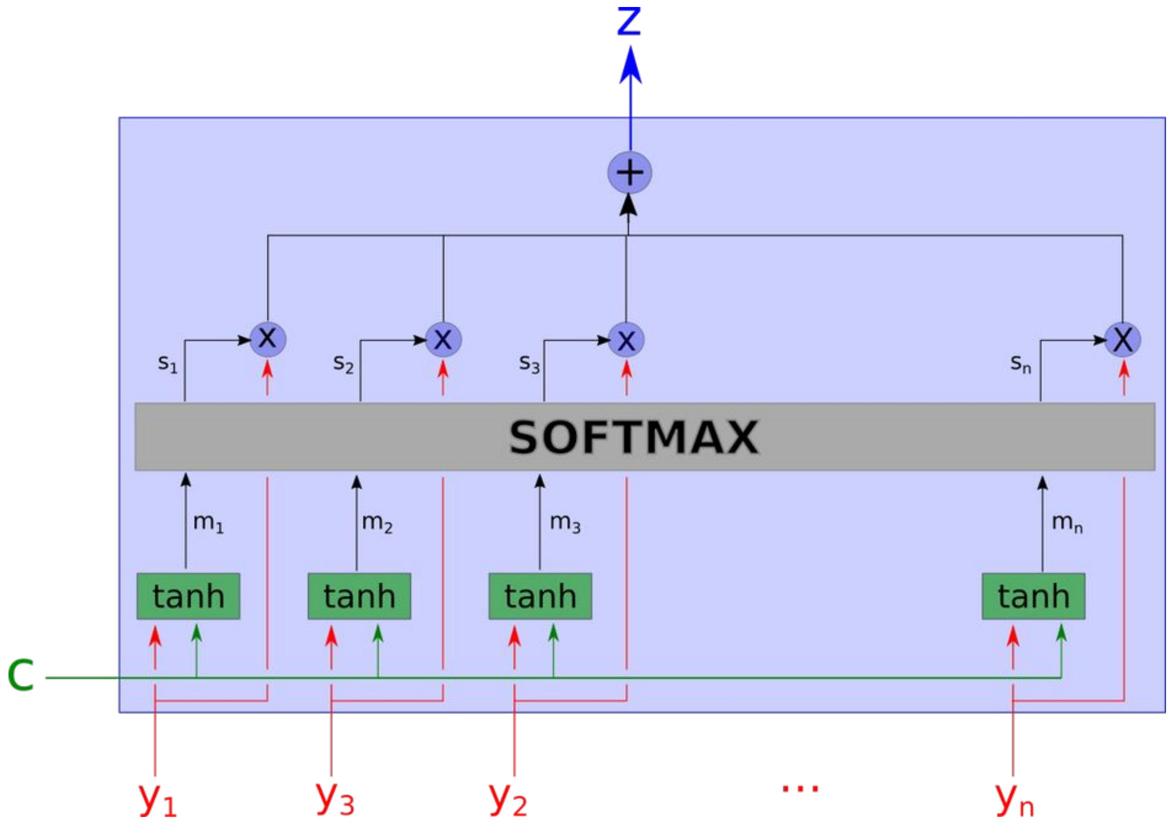


Attention layer

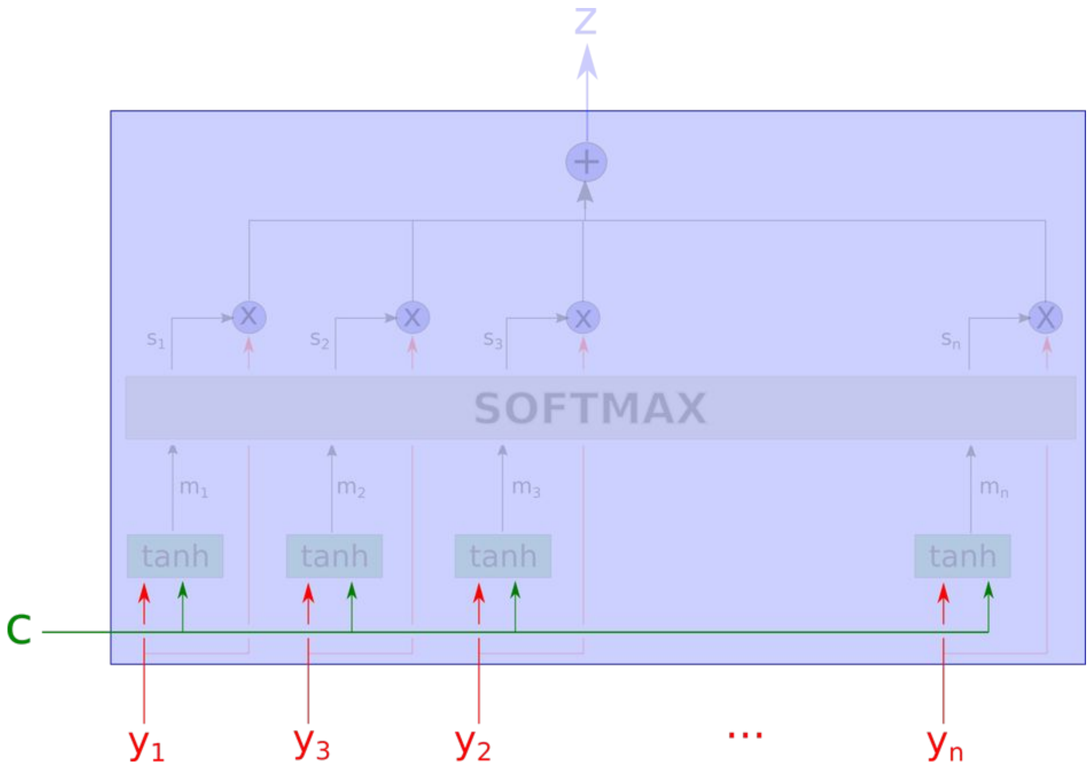
- n input arguments y_i
- Context c
- Output z is summary of y_i focusing on the context c



Attention layer

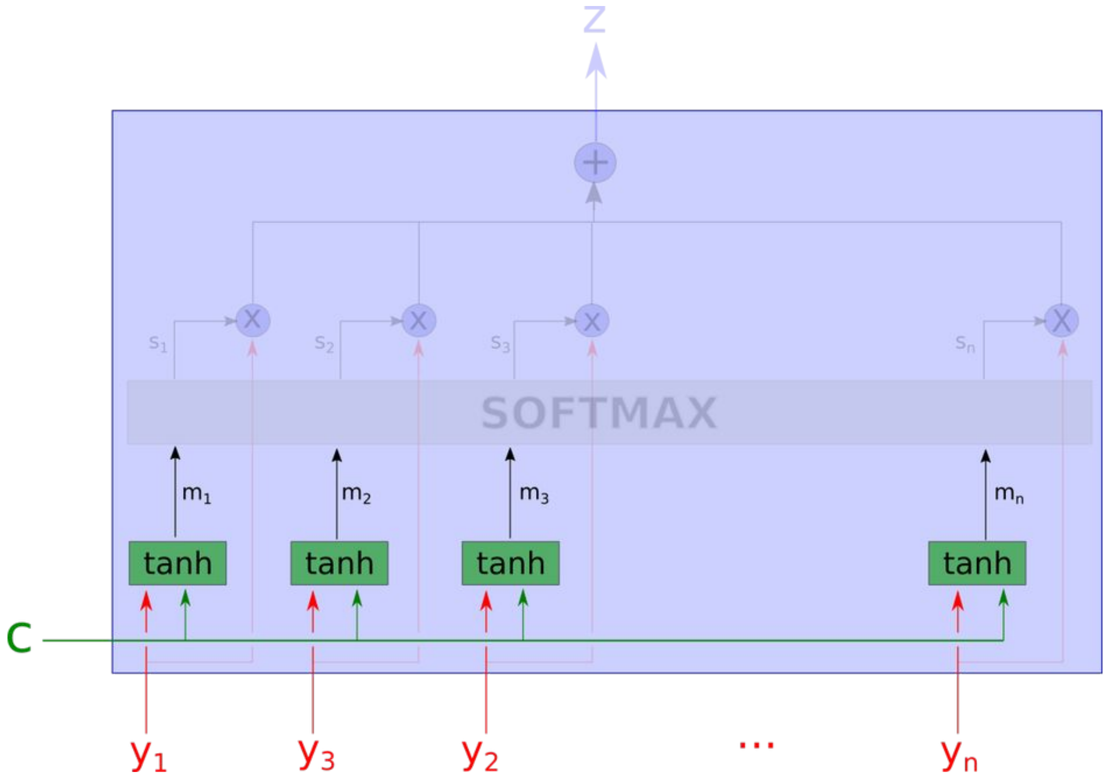


Step 1



Step 2

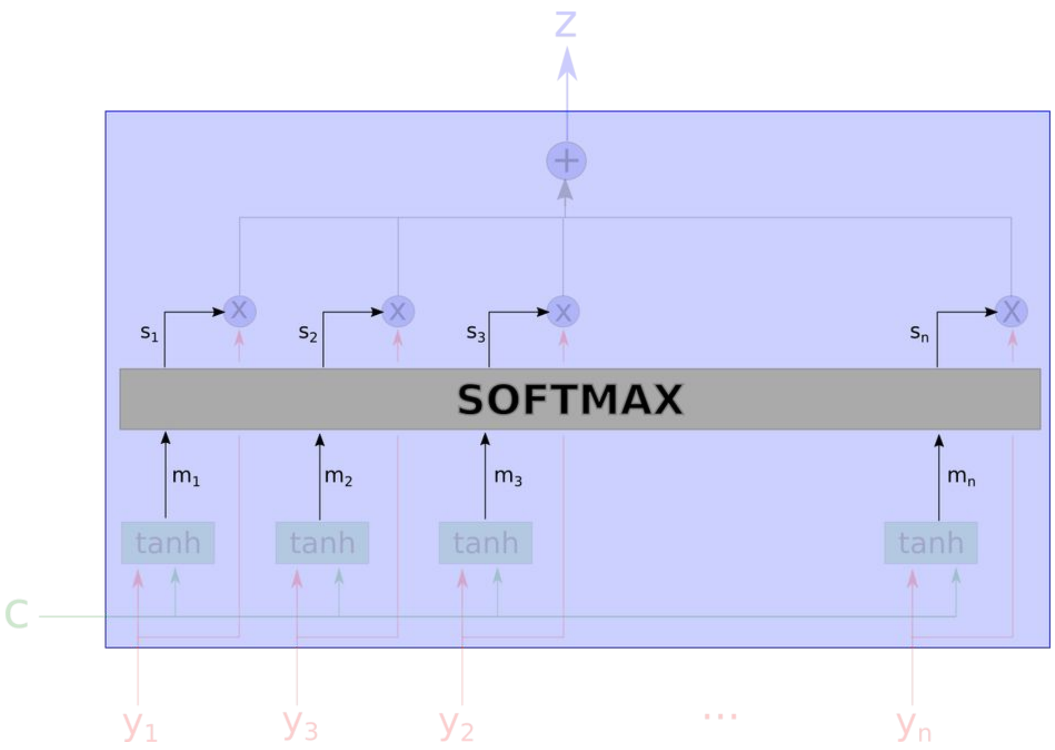
$$m_i = \tanh\left(W_{cm}c + W_{ym}y_i\right)$$



Step 3

$$s_i \propto \exp\left(\langle w_m, m_i \rangle\right)$$

$$\sum_i s_i = 1$$



Step 4

$$z = \sum_i s_i y_i$$

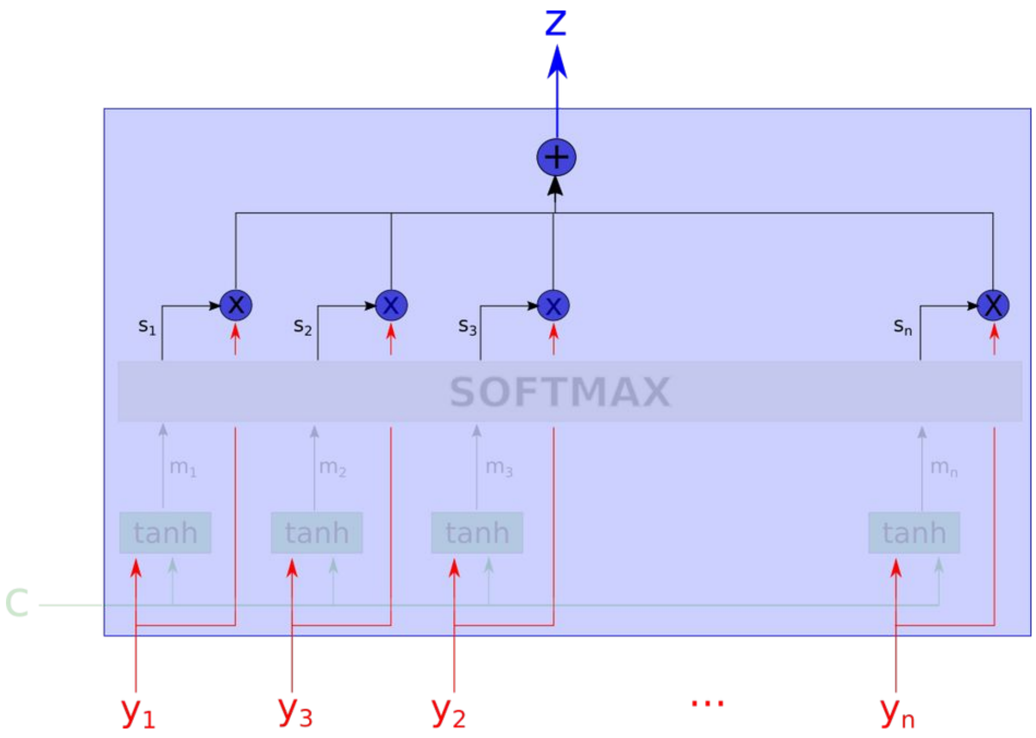
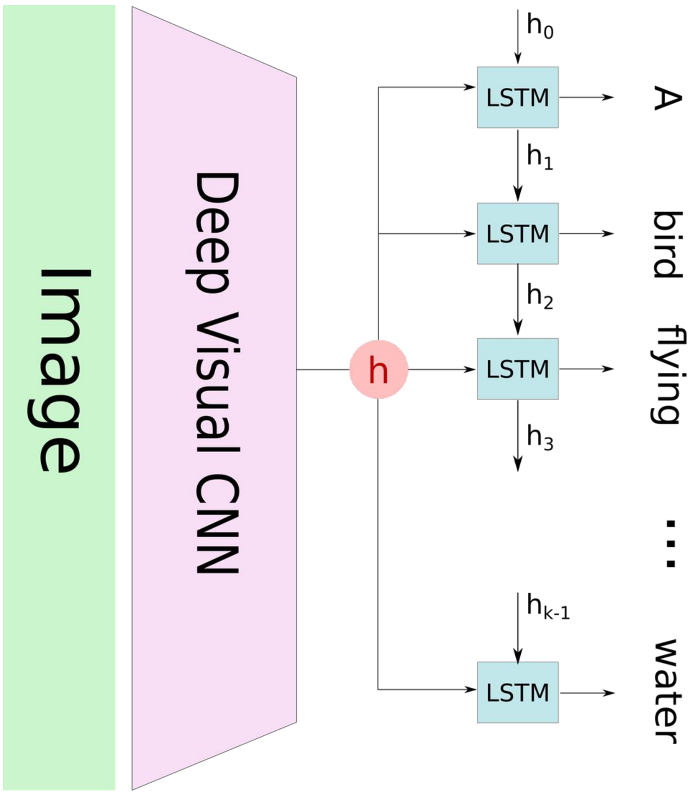


Image captioning



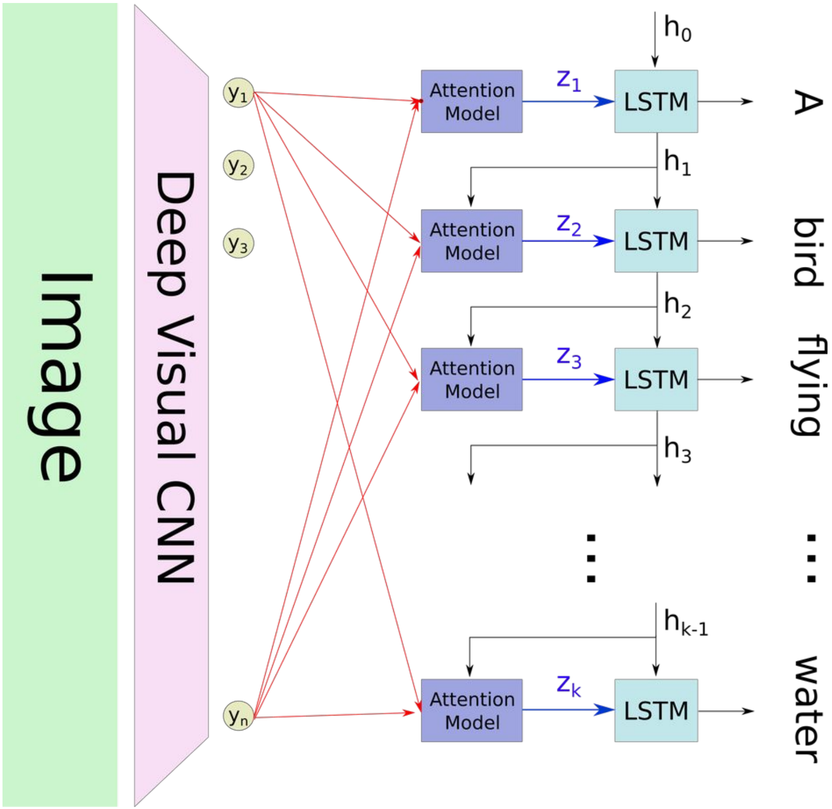
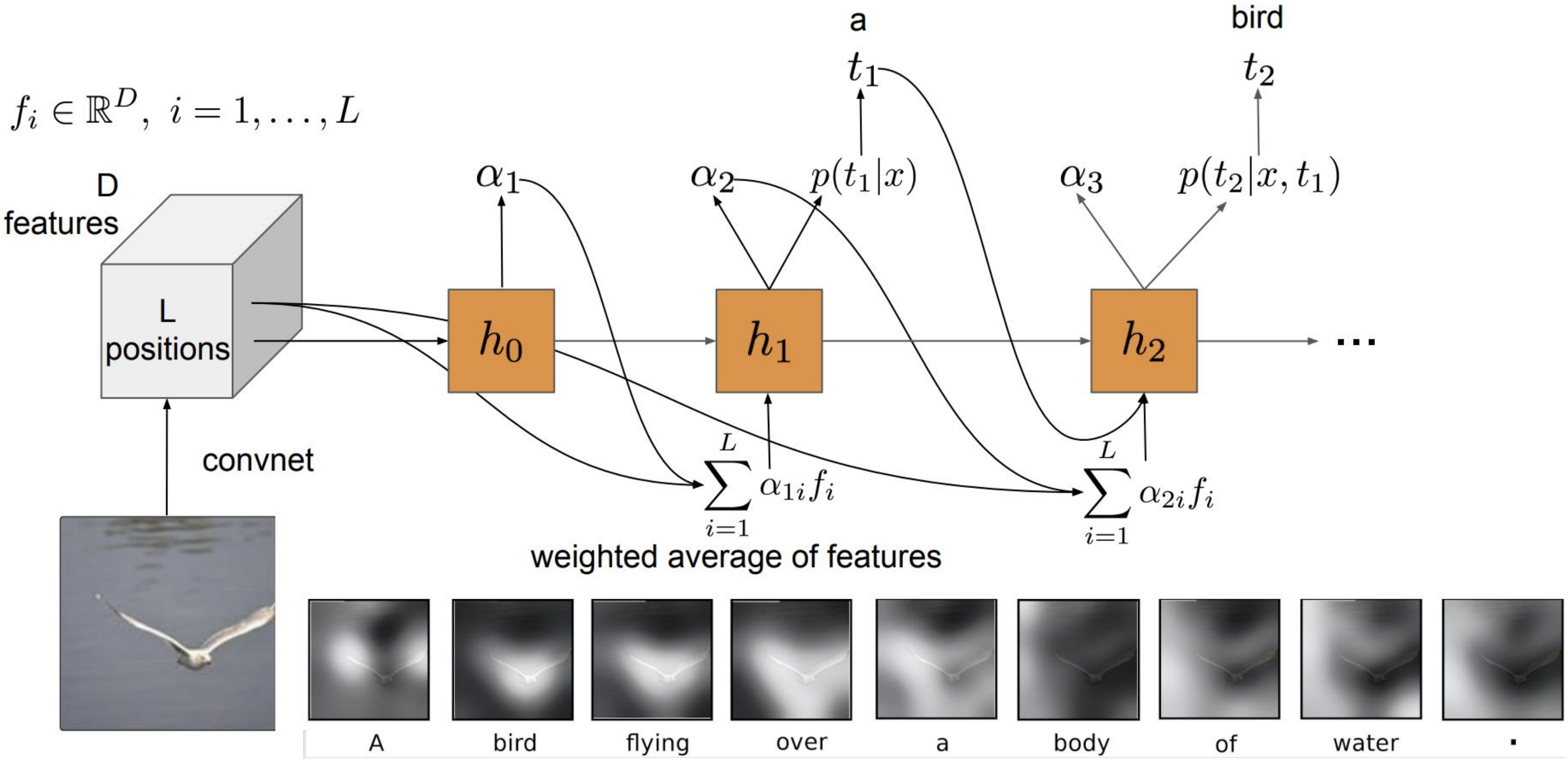
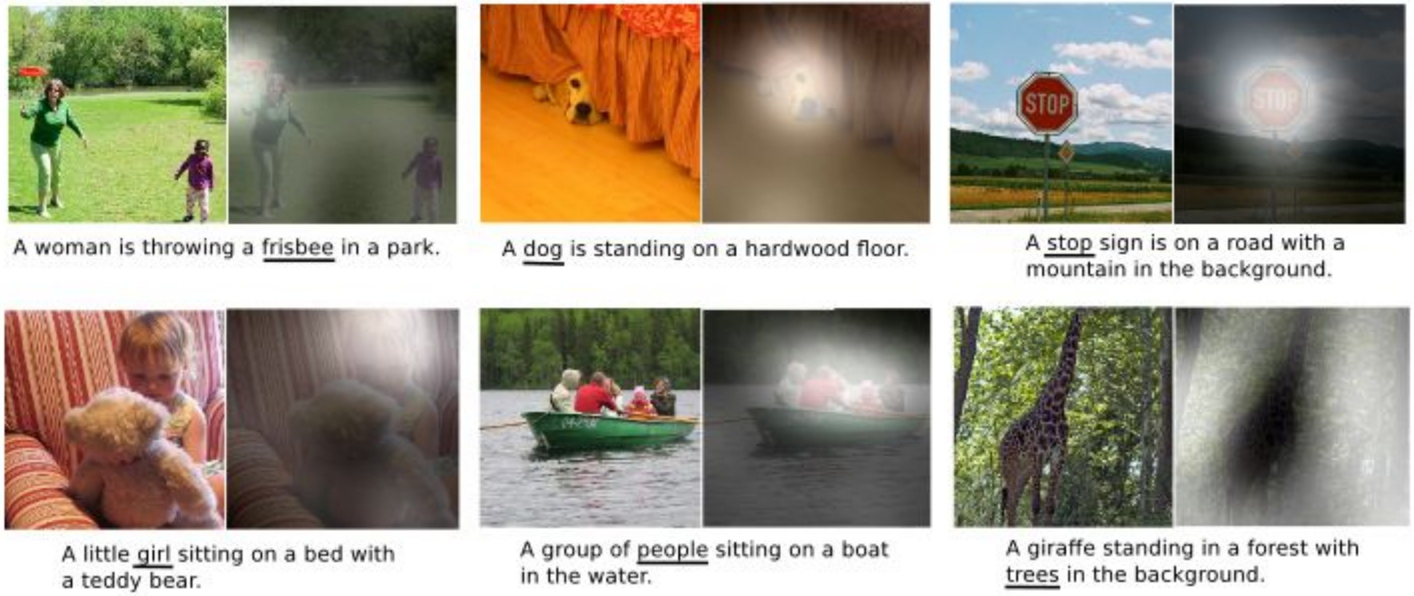


Image Captioning with attention



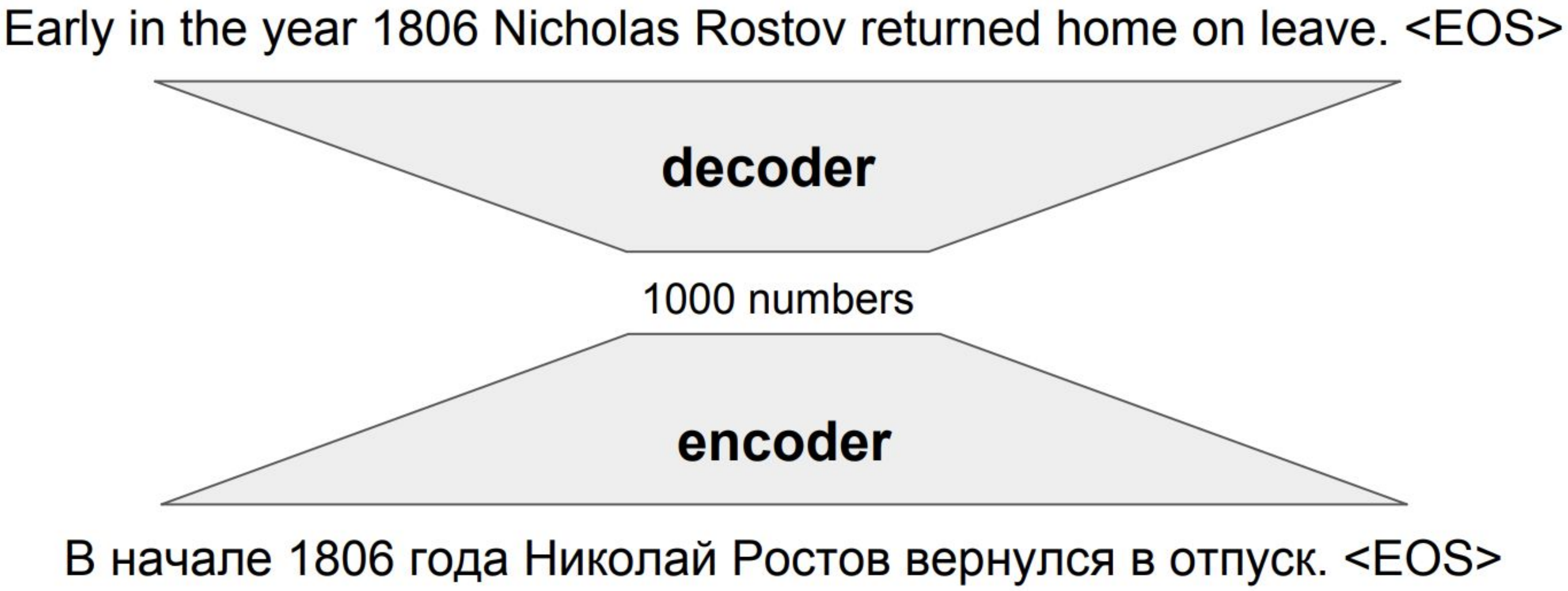
Attention

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



Attention for Machine Translation

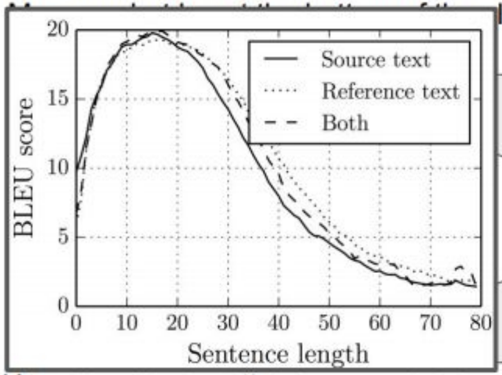
Translation

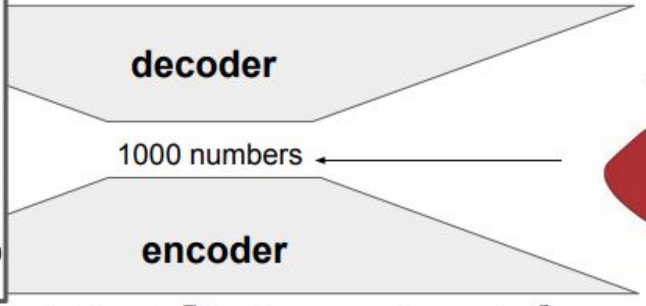


Neural machine translation and long sentences

Meeting a comrade at the last post station but one before Moscow, Denisov had drunk three bottles of wine with him and, despite the jolting ruts across the snow-covered road, did not once wake up on the way to

leigh beside Rostov, who grew more and more impatient the nearer

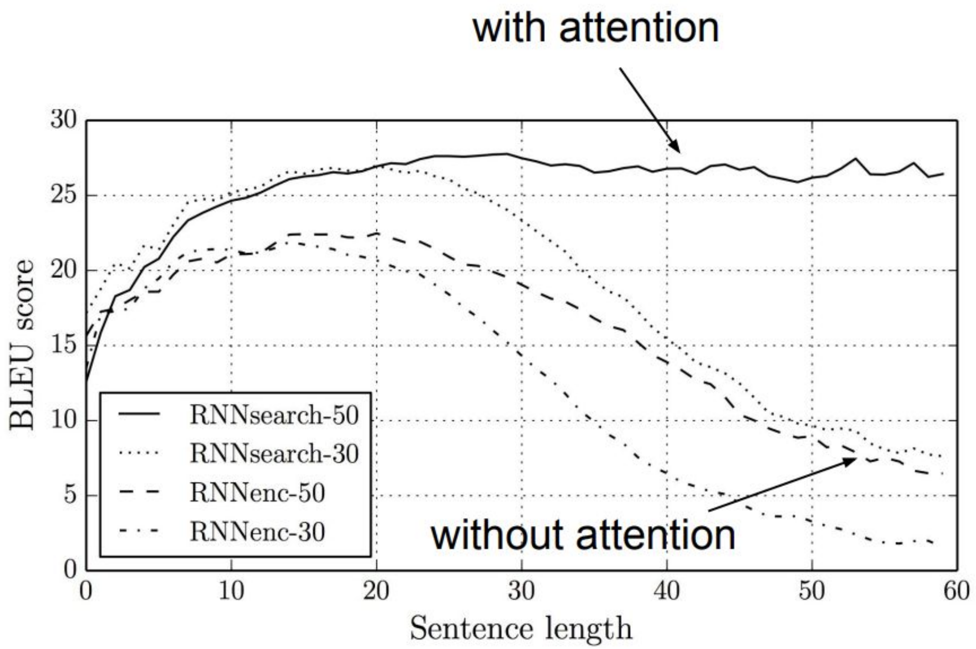




На предпоследней станции, встретив товарища, Денисов выпил с ним три бутылки вина и подъезжая к Москве, несмотря на ухабы дороги, не просыпался, лежа на дне перекладных саней, подле Ростова, который, по мере приближения к Москве, приходил все более и более в нетерпение. <EOS>

Leo Tolstoy "War and Peace", 1869
Cho et al. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", 2014

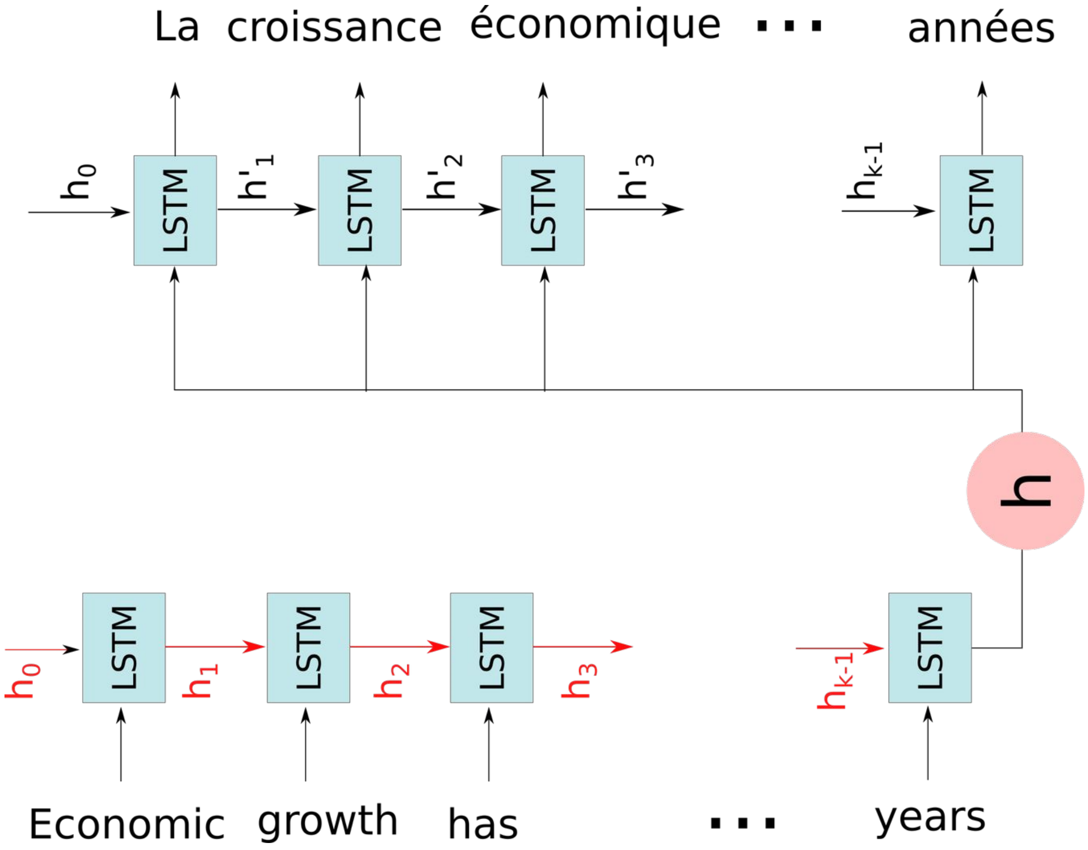
Translation with attention



Bahdanau et al. "Neural Machine Translation by Jointly Learning to Align and Translate", 2014

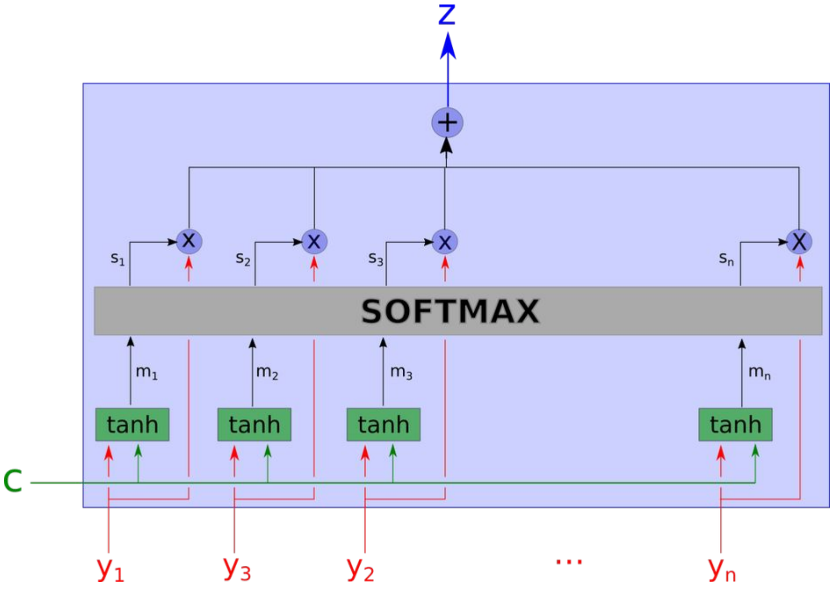
Machine translation

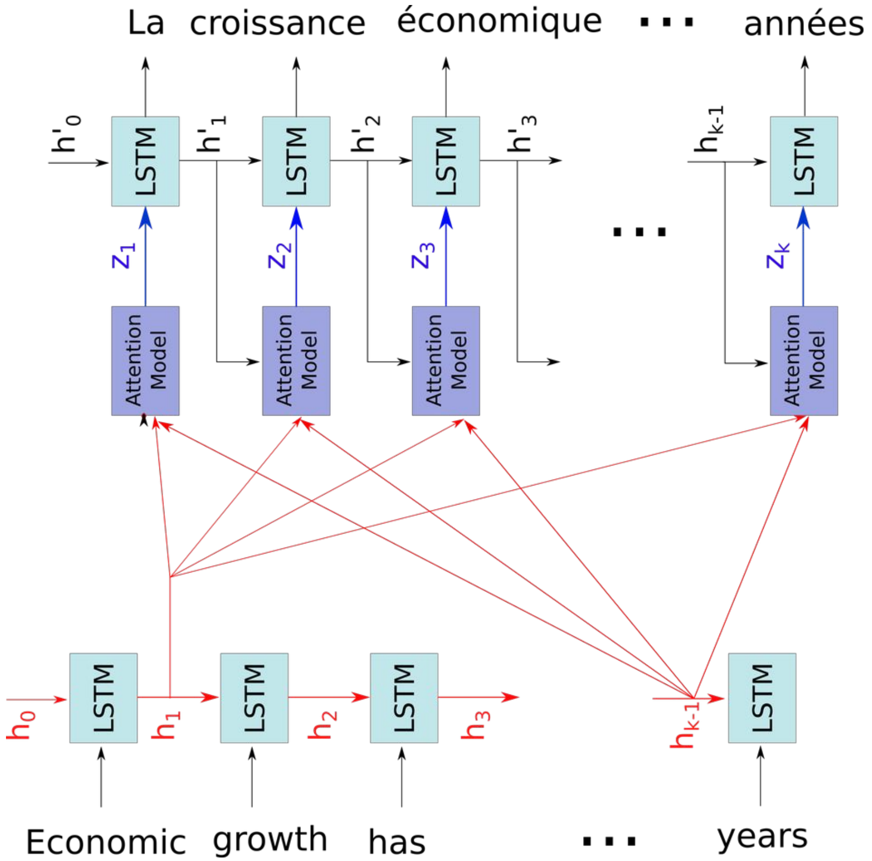
- 2 LSTMs
- Encoder-decoder structure
- Generation per token (word)

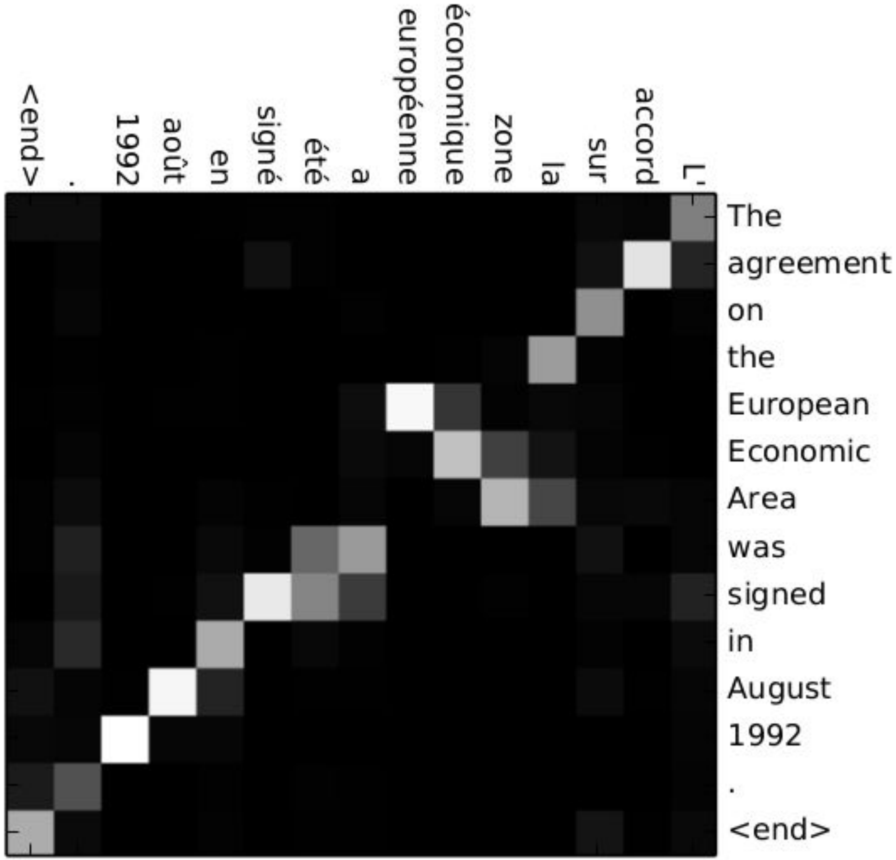


Translation with attention

- Add attention block
- Each h — attention input
- Each h’ — attention context






<end>
.
1992
août
en
signé
été
a
européenne
économique
zone
la
sur
accord
L'
The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>

Attention is all you need

Attention Is All You Need

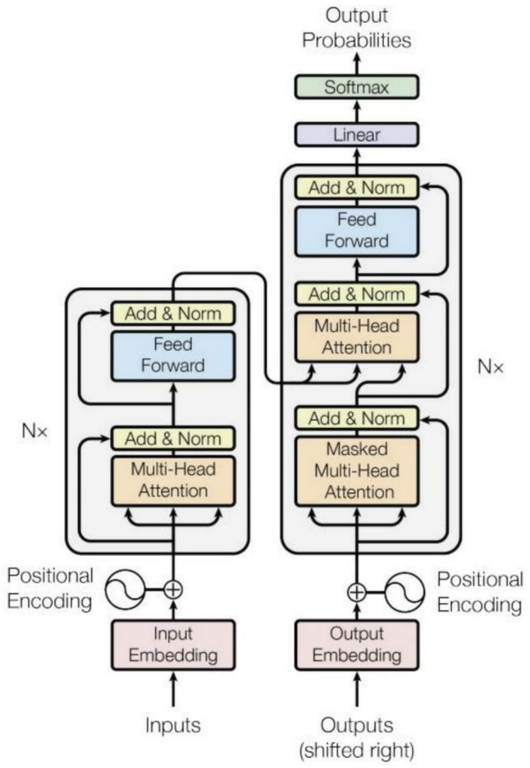
Replace LSTMs with a lot of attention! Apply to neural machine translation

- State-of-the art results
- Much less computation for training

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [17]	23.75			
Deep-Att + PosUnk [37]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [36]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [31]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [37]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [36]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Transformer architecture

- Encoder: 6 layers of self-attention + feed-forward network
- Decoder: 6 layers of masked self-attention and output of encoder + feed-forward.



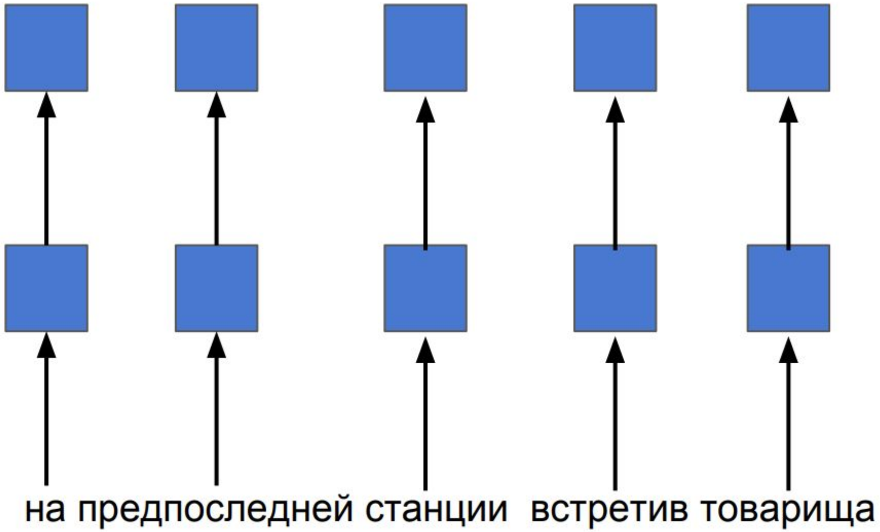
Layer types

- Input layer
- Per-word feedforward
- Self-attention
- Attention over encoder outputs

Input layer

add positional encoding
encode offsets between words
not used in LSTMs

embedding
pick a vector for every word



Positional encoding

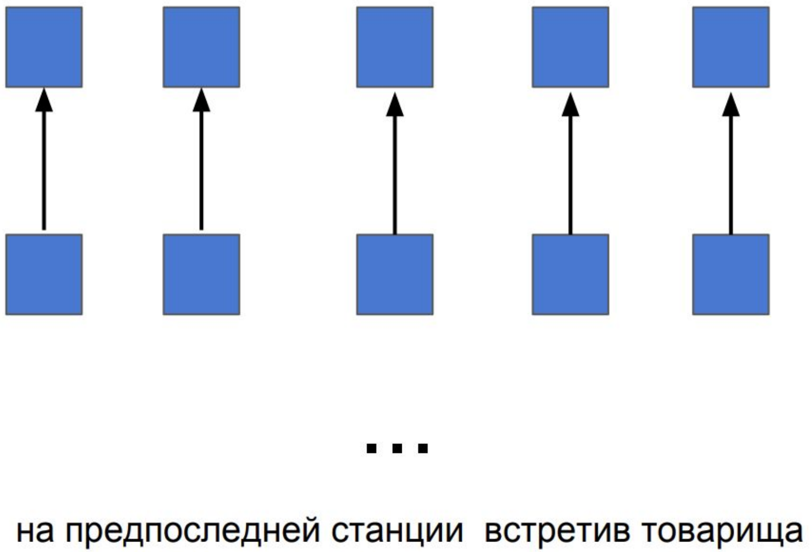
- Positional encoding provides relative or absolute position of given token
- Many options to select positional encoding, in this work:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

- Alternative, to learn positional embeddings

Per-word feedforward

fully-connected network
same for every word
G(x) = Dense(ReLU(Dense(x)))



Self-attention

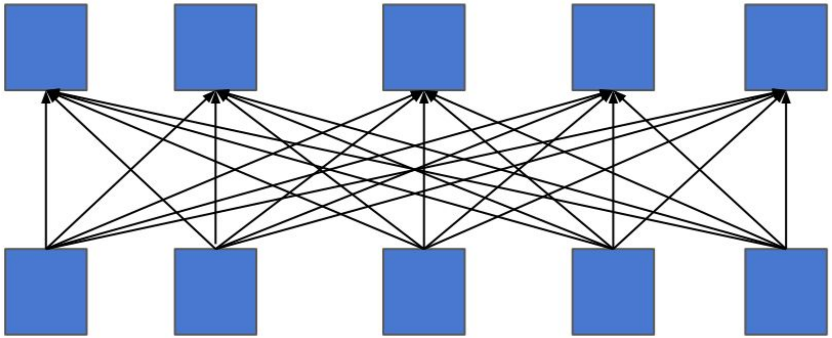
every word attends to
features of all words

replaces recurrence

$$\alpha_{ki} = \frac{\exp(\mathrm{score}(x_k, x_i))}{\sum_{j=1}^{L} \exp(\mathrm{score}(x_k, x_j))}$$

$$\mathrm{score}(x_k, x_i) = x_k^T x_i / \sqrt{d},\ d = \dim(x_k)$$

$$y_k = \sum_{i=1}^{L} \alpha_{ki} x_i$$



x_i

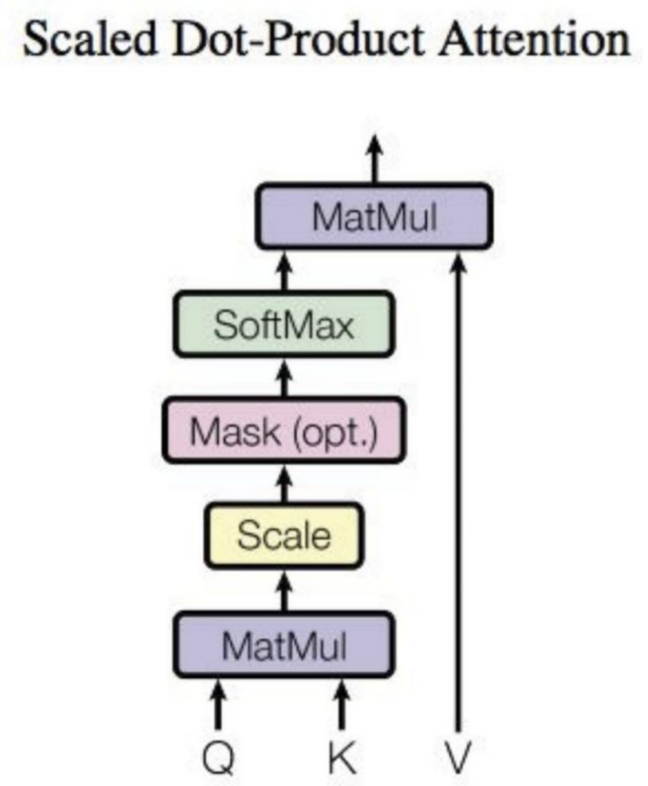
...

на предпоследней станции встретив товарища

Scaled dot-product attention

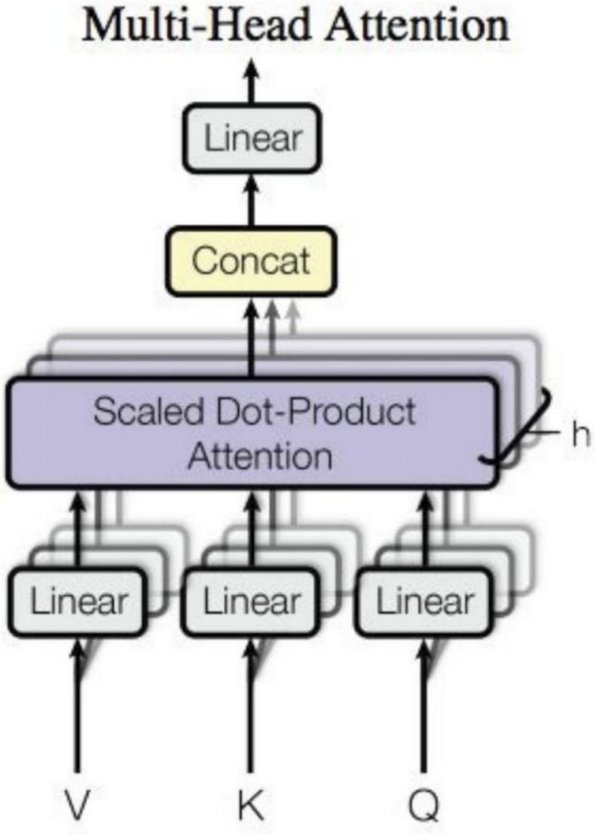
- How Values should pay attention on Query using Keys
- Self-attention: Query = Key = Value

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

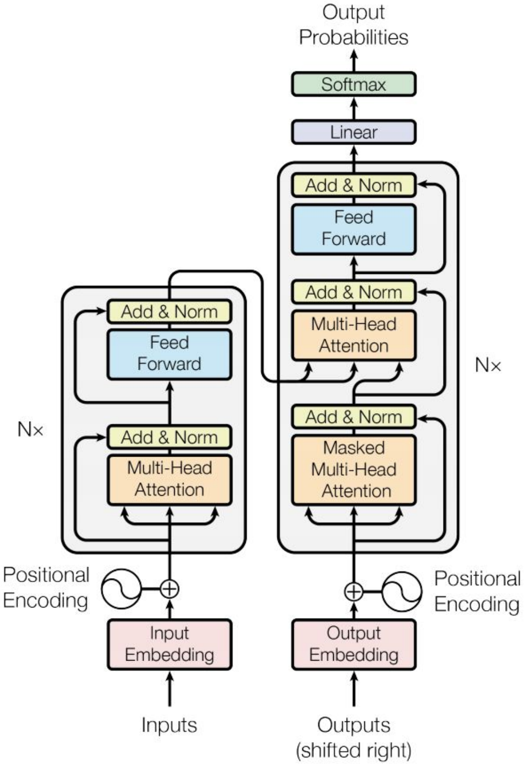


Multi-head attention

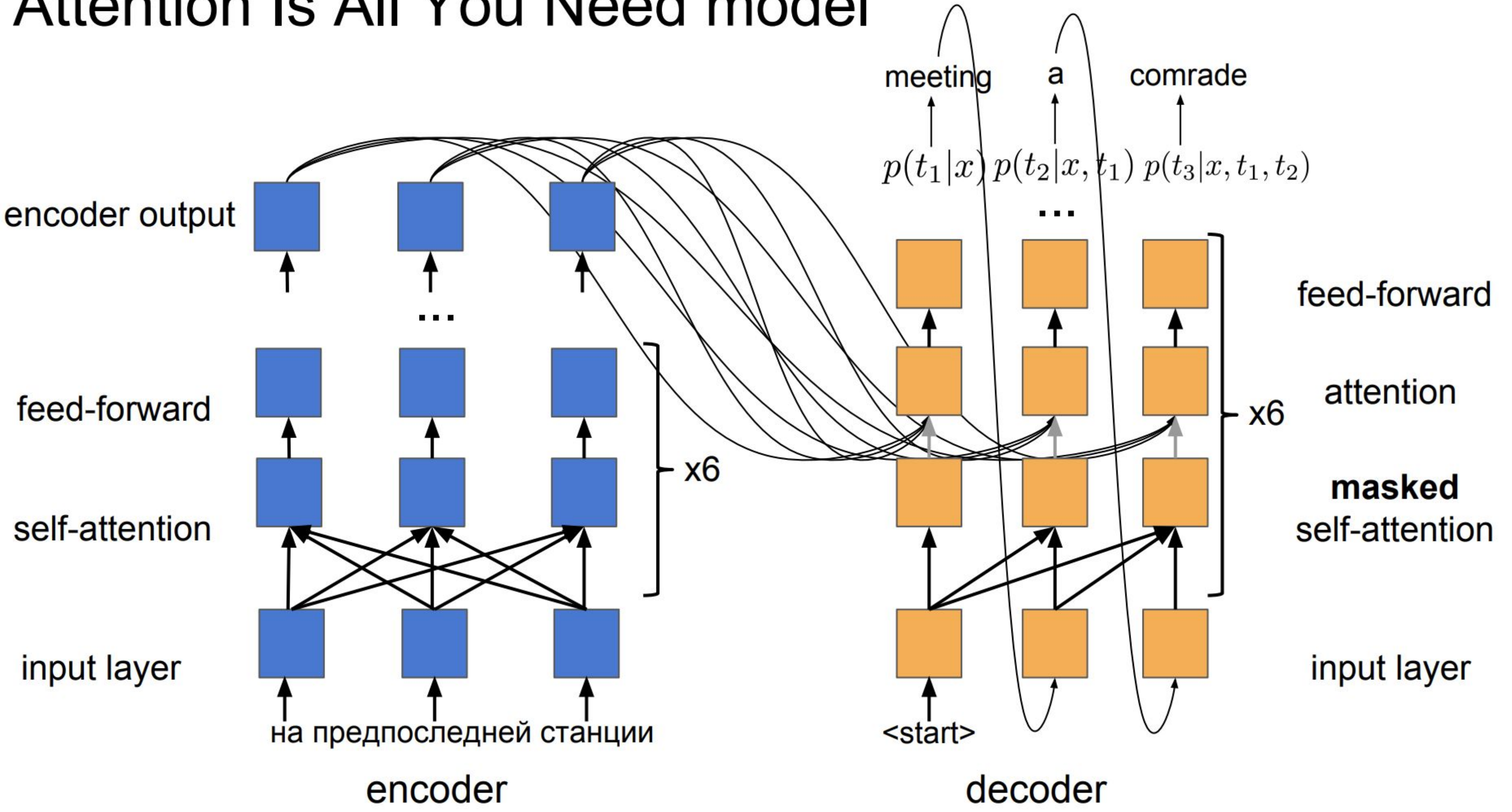
- Apply attention in K feature spaces;
- Concatenate results



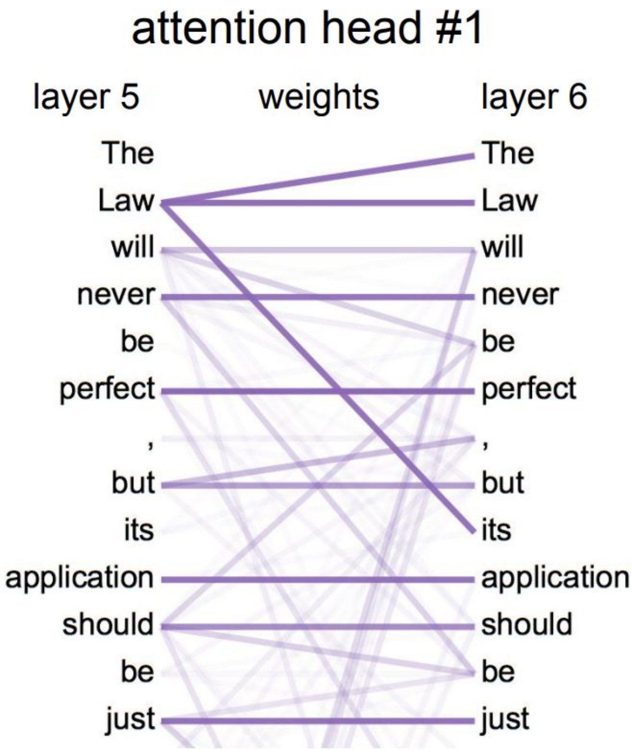
Transformer

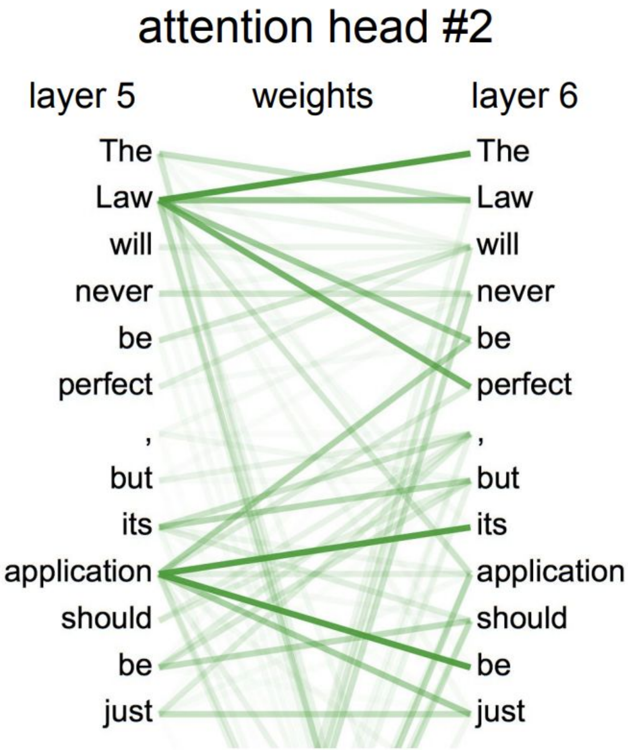


Attention Is All You Need model

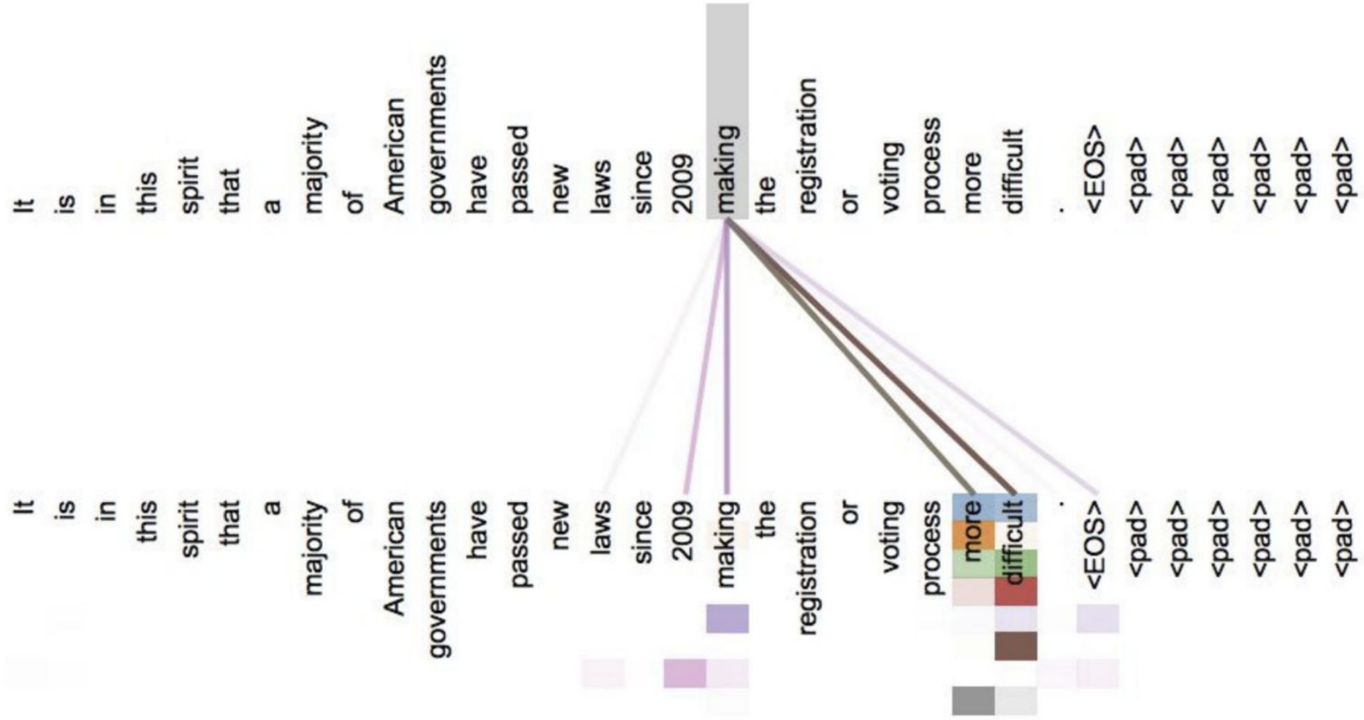


Self-attention





Multi-head attention



Attention for Point Clouds