Attention Is All You Need

CS-E4070 – Advanced Topics in Deep Learning

Overview

Background	Neural networks
	Attention
Transformer	Architecture
	Results
	Implications

Neural networks

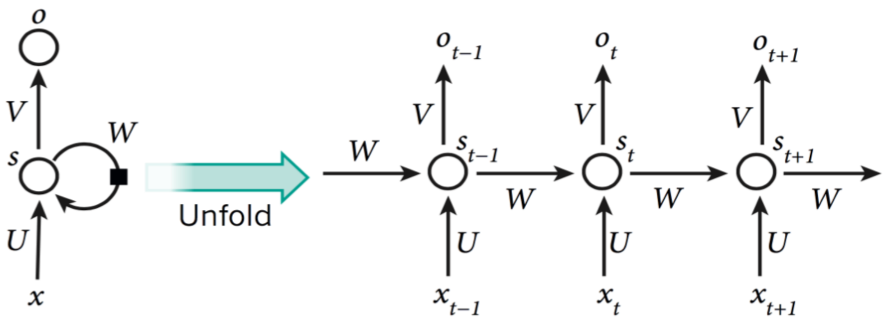
Computer vision, Natural language processing…

Machine translation, Image captioning, Speech recognition …

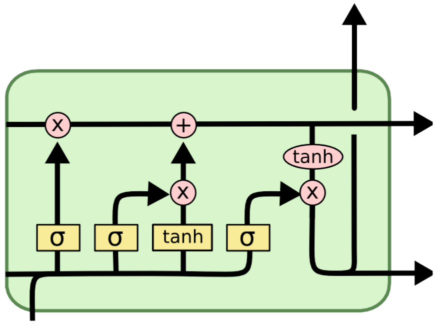
Feedforward networks
Autoencoder networks
Convolutional networks
Recurrent networks
Generative adversarial networks

Neural networks – recurrent

RNN

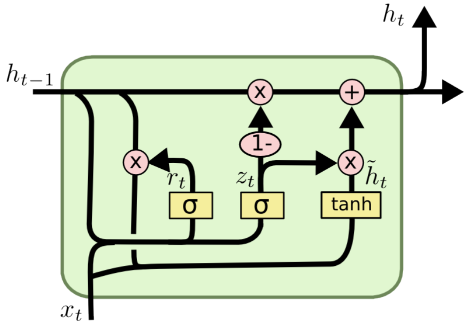


LSTM



$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{x}_t\mathbf{U}^i + \mathbf{h}_{t-1}\mathbf{W}^i + \mathbf{b}_i)\\
\mathbf{f}_t &= \sigma(\mathbf{x}_t\mathbf{U}^f + \mathbf{h}_{t-1}\mathbf{W}^f + \mathbf{b}_f)\\
\mathbf{o}_t &= \sigma(\mathbf{x}_t\mathbf{U}^o + \mathbf{h}_{t-1}\mathbf{W}^o + \mathbf{b}_o)\\
\mathbf{q}_t &= \tanh(\mathbf{x}_t\mathbf{U}^q + \mathbf{h}_{t-1}\mathbf{W}^q + \mathbf{b}_q)\\
\mathbf{p}_t &= \mathbf{f}_t * \mathbf{p}_{t-1} + \mathbf{i}_t * \mathbf{q}_t\\
\mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{p}_t)
\end{aligned}
$$

GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

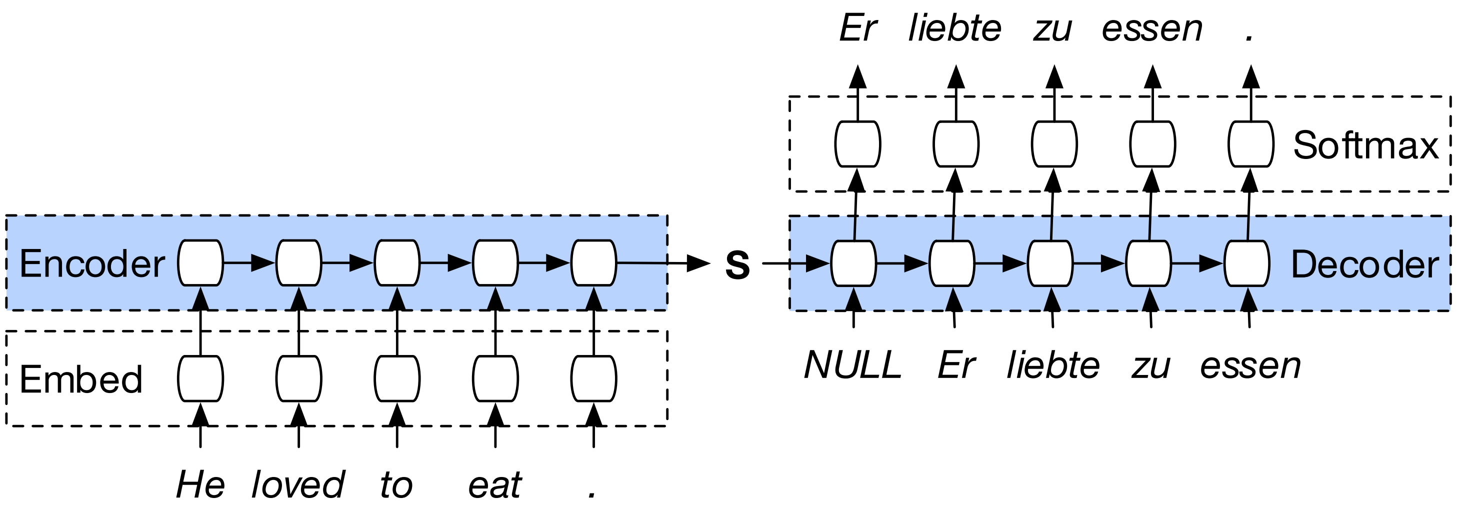
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

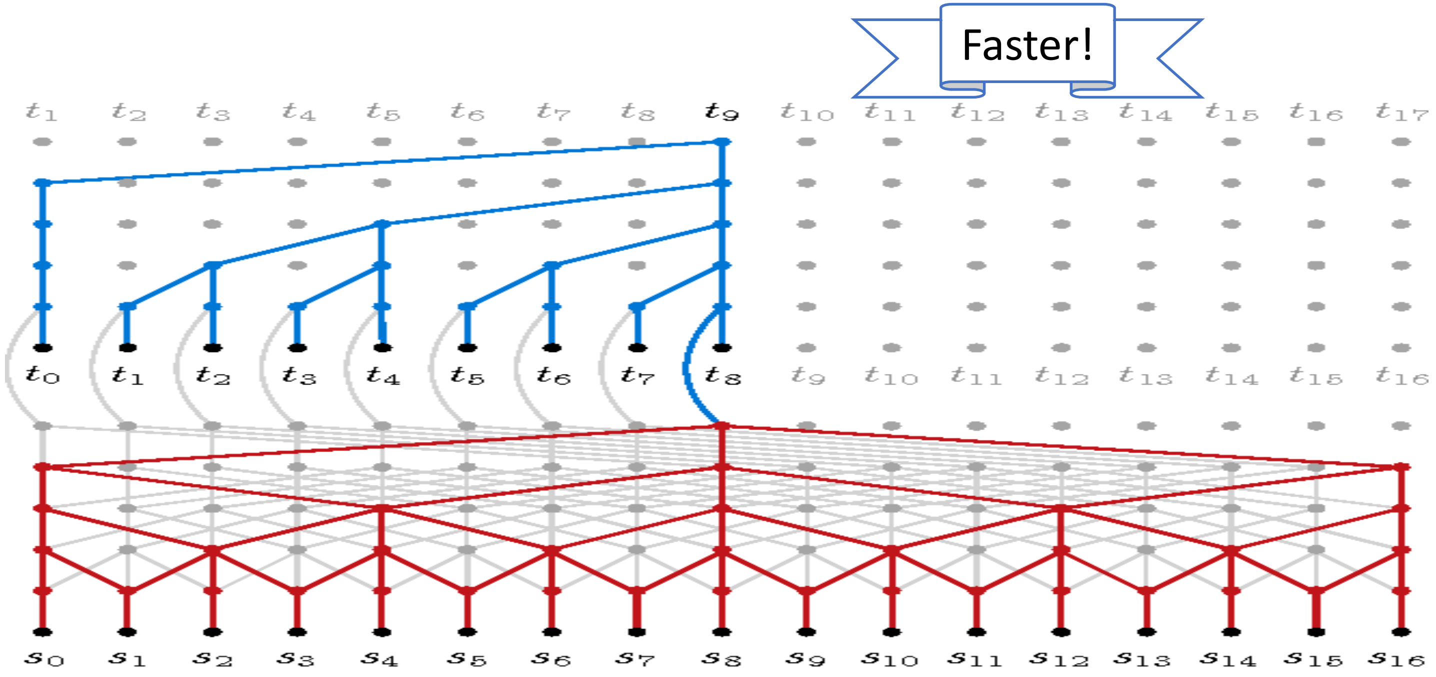
Neural networks – seq2seq

- Word embeddings
- Temporal information
- Context (subjective)
- Logits
- Teacher forcing



Neural networks – convolutional (for text)

- Embeddings matrix
- Filter widths equals embedding size
- Height states $\boldsymbol{h}$-gram
- Representations
- ByteNet, WaveNet



Neural networks – problems

- Hard to parallelize *#bottleneck*
- Limited by convolution filter sizes *#bottleneck*
- Source sequences compressed as fixed length vectors *#bottleneck*

- Increase in length of sequences, decreases performance *#case*
- Alignment problems, local-global information *#case*

Attention

- everything be pre cook and dry its crazy most Filipino people be used to very cheap ingredient and they do know quality the food be disgusting I have eat at least 20 different Filipino family home this not even medioc

- seriously f *** this place disgust food and shitty service ambience be great if you like dine in a hot cellar eng stagnate air truly it be over rate over price and they just under deliver forget try order a drink here it will take forever get and when it finally do arrive you will be ready pass out from heat exhaustion and lack of oxygen be that a head change you do not even have pay for it I will not disgust you with the detailed review of ever have try here but make it simple it all suck and after you get the bill you will be walk out with a sore ass sav money and spare your self the disappointment

- i be so angry about my horrible experience at Medusa today my previous visit be amaze 5/5 however my g of town and I land an appointment with Stephanie I go in with a picture of roughly what I want and come out absolutely nothing like it my hair be a horrible ashy blonde not anywhere close to the platinum blonde I requ she will not do any of the pop of colour I want and even after specifically tell her I do not like blunt cut my ha have lot of straight edge she do not listen to a single thing I want and when I tell her I be unhappy with the c she basically tell me I be wrong and I have do it this way no no I do not if I can go from Little Mermaid red to golden blonde in 1 sitting that leave my hair fine I shall be able go from golden blonde to a shade of platinu blonde in 1 sitting thanks for ruin my New Year's with 1 the bad hair job I have ever have

1 star reviews

- i really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do go to MI because of the quality of the highlight and the price the price be very affordable the highlight fanta thank Ashley i highly recommend you and ill be back

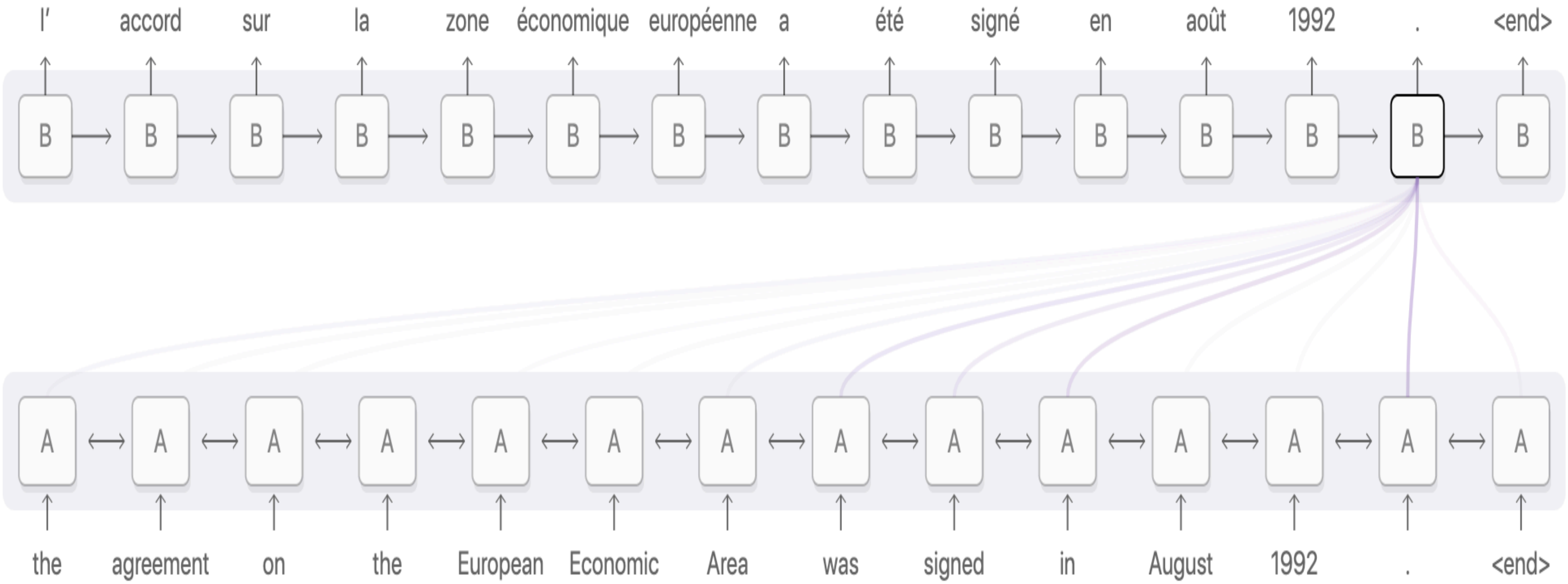
- love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can tas steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I ha had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Co

- great food and good service .... what else can you ask for everything that I have ever try here have be grea

- first off I hardly remember waiter name because its rare you have an unforgettable experience the day I go celebrate my birthday and let me say I leave feel extra special our waiter be the best ever Carlos and the st well I be with a party of 4 and we order the potato salad shrimp cocktail lobster amongst other thing and bo the food great the lobster be the good lobster I have ever eat if you eat a dessert I will recommend the chee cake that be also the good I have ever have it be expensive but so worth every penny I will definitely be bac there go again for the second time in a week and it be even good ...... this place be amazing

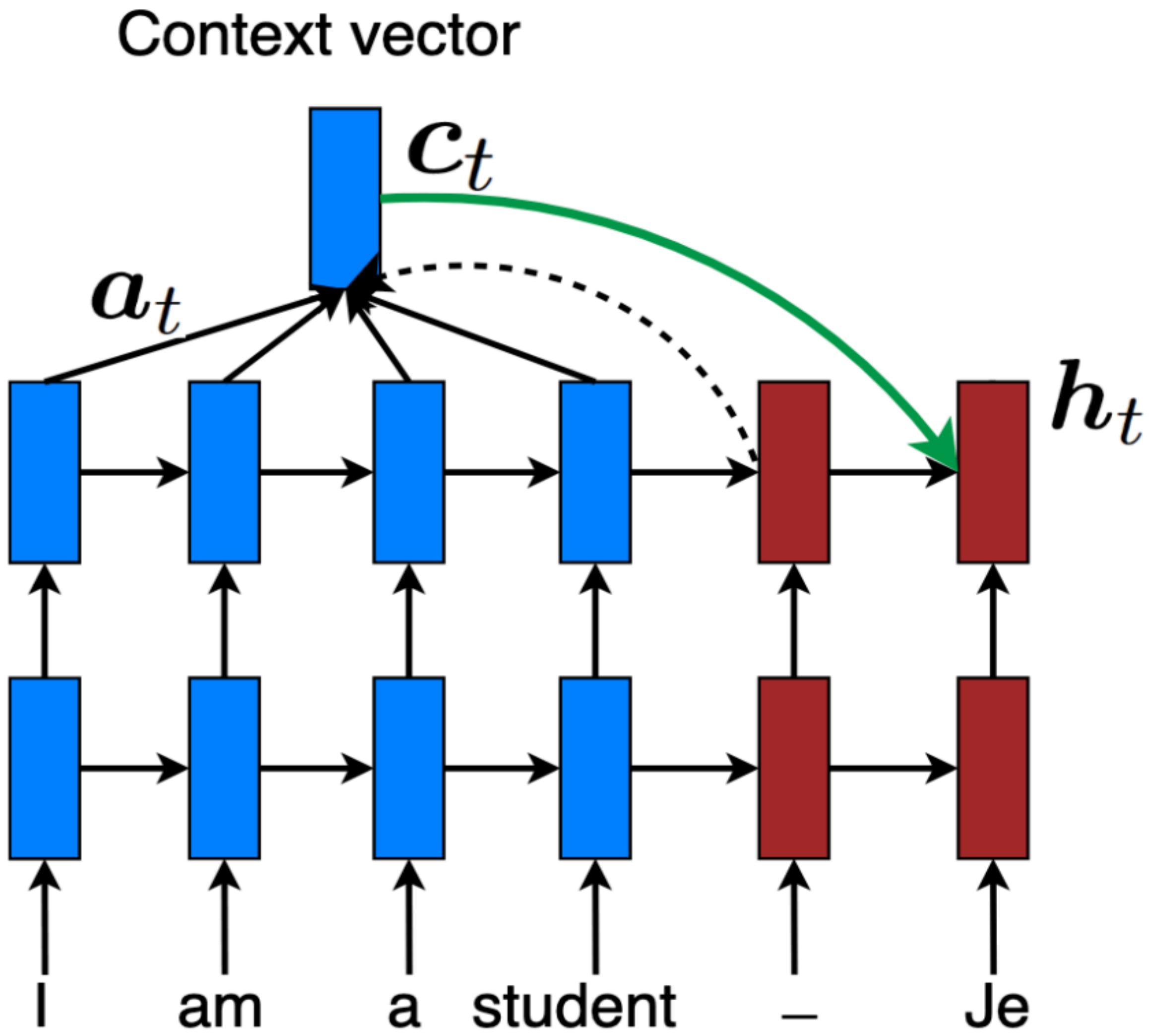
5 star reviews

Attention – seq2seq



Attention – seq2seq

- ✓Scoring
- ✓Source and target hidden states
- ✓Happens with every source state
- ✓SoftMax
- ✓Context vector
- ✓Next hidden state



Attention – types of scorers

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

Attention – scoring

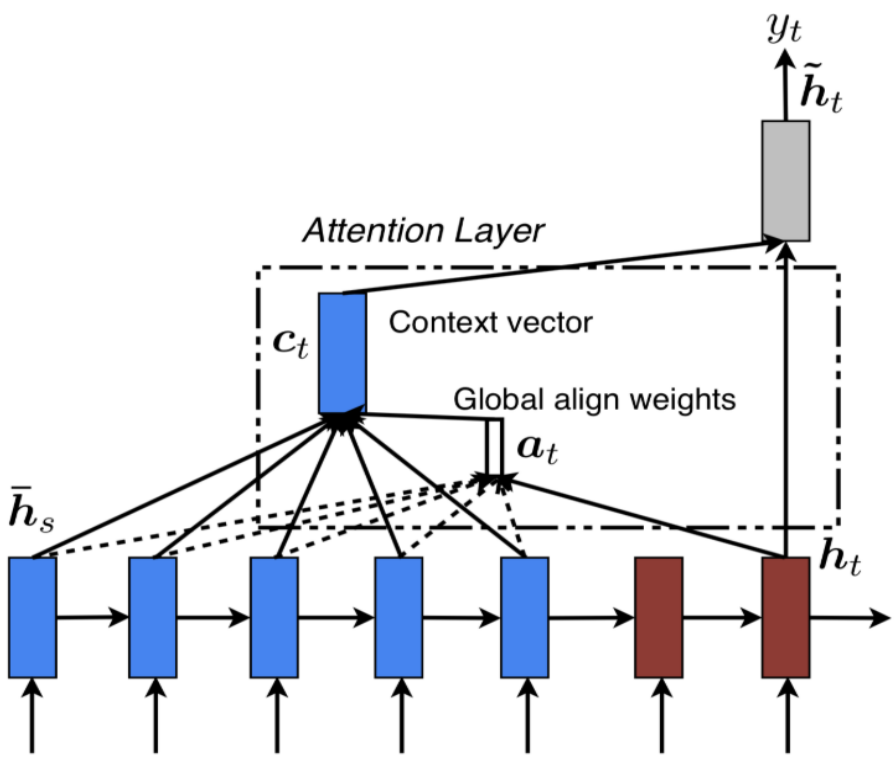
Scoring:

- $score\left(h_{target}^{t-1}, h_{source}^{\star}\right) = (h_{target}^{t-1}).\, h_{source}^{\star}$
- Another neural network that learns the function

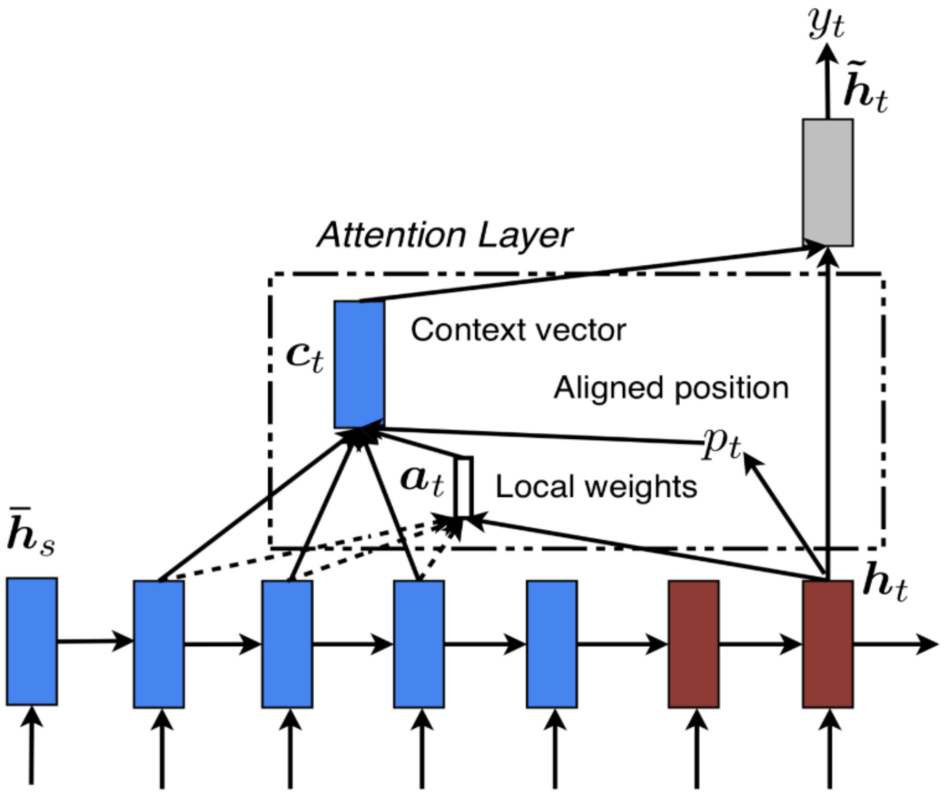
Attention vectors → Scoring network/layer → Encoder states

$\left.\begin{array}{l}\text{Global attention (soft)}\\ \text{Local attention (hard)}\end{array}\right\}$ basic variants

Attention – global vs local (1)

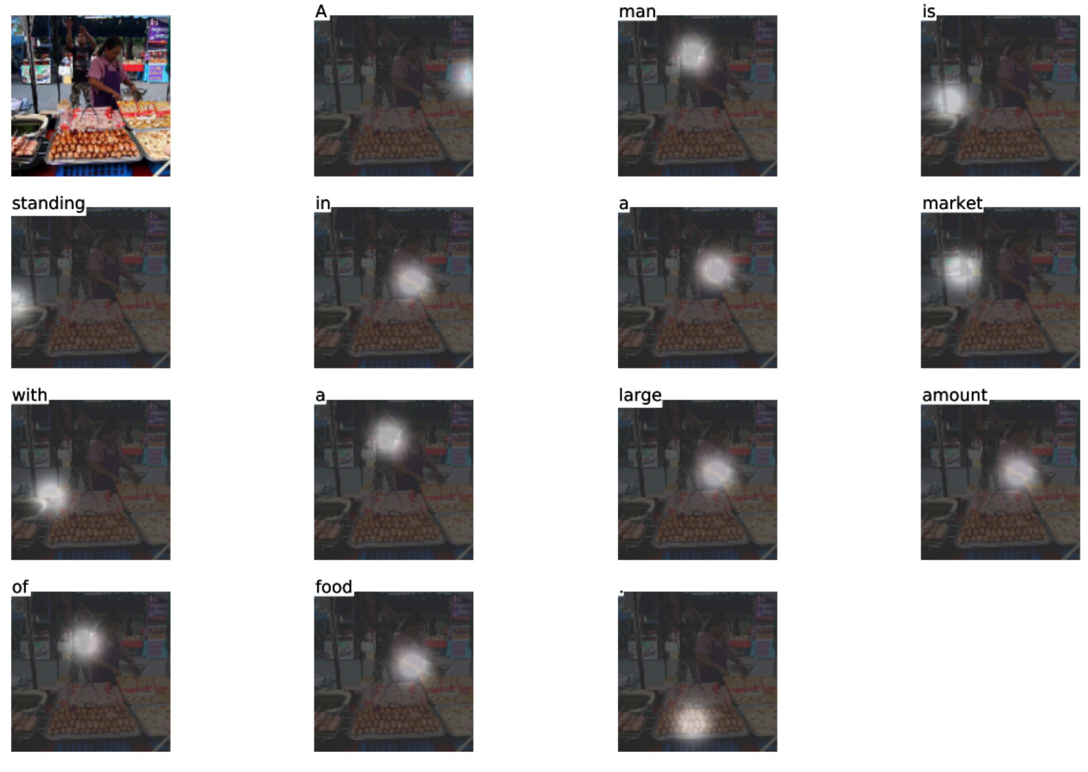


Global Attention Model

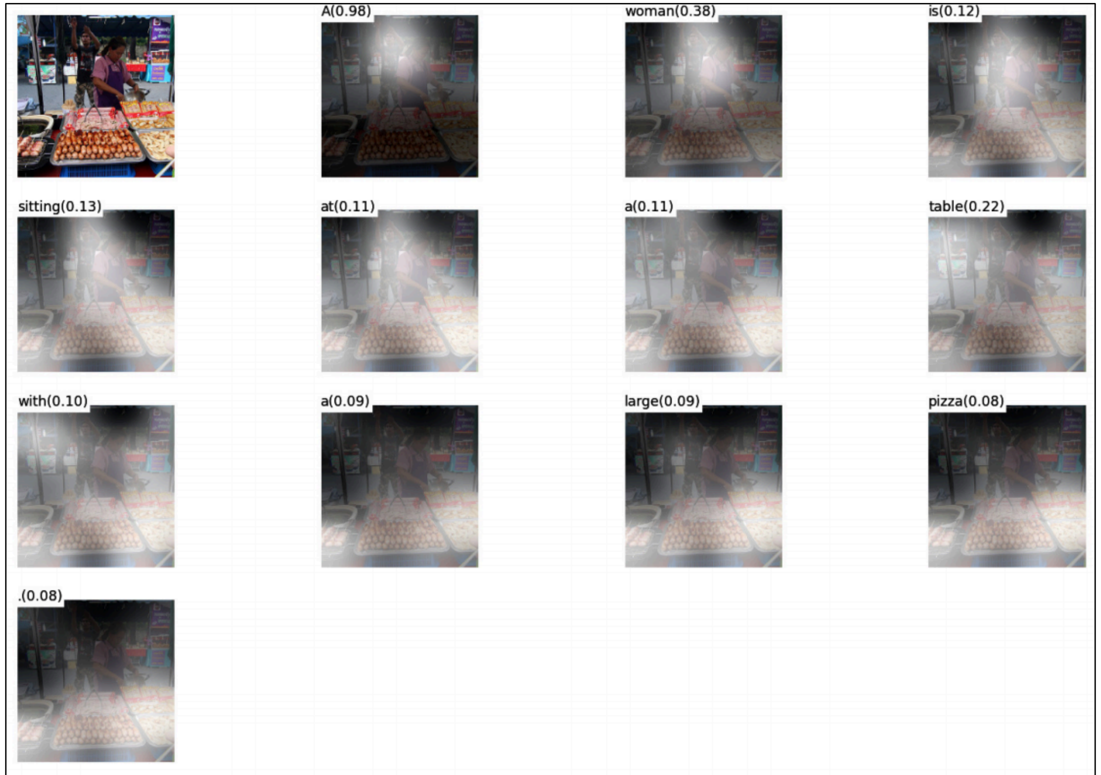


Local Attention Model

Attention – global vs local (2)



(a) A man is standing in a market with a large amount of food.



(b) A woman is sitting at a table with a large pizza.

Attention – notes

- Hard to parallelize *#bottleneck*
- ~~Limited by convolution filter sizes *#bottleneck*~~
- ~~Source sequences compressed as fixed length vectors *#bottleneck*~~

- ~~Increase in length of sequences, decreases performance *#case*~~
- Alignment problems, local-global information *#case*
- Improves performance
- Better interpretability

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

NIPS
2017

Architecture – at a glance

- No recurrence
- Relies completely on attention
- Retains known structure
- Outperforms RNNs & CNNs
- Novelties:
    1. Scaled dot product attention
    2. Multi-head attention

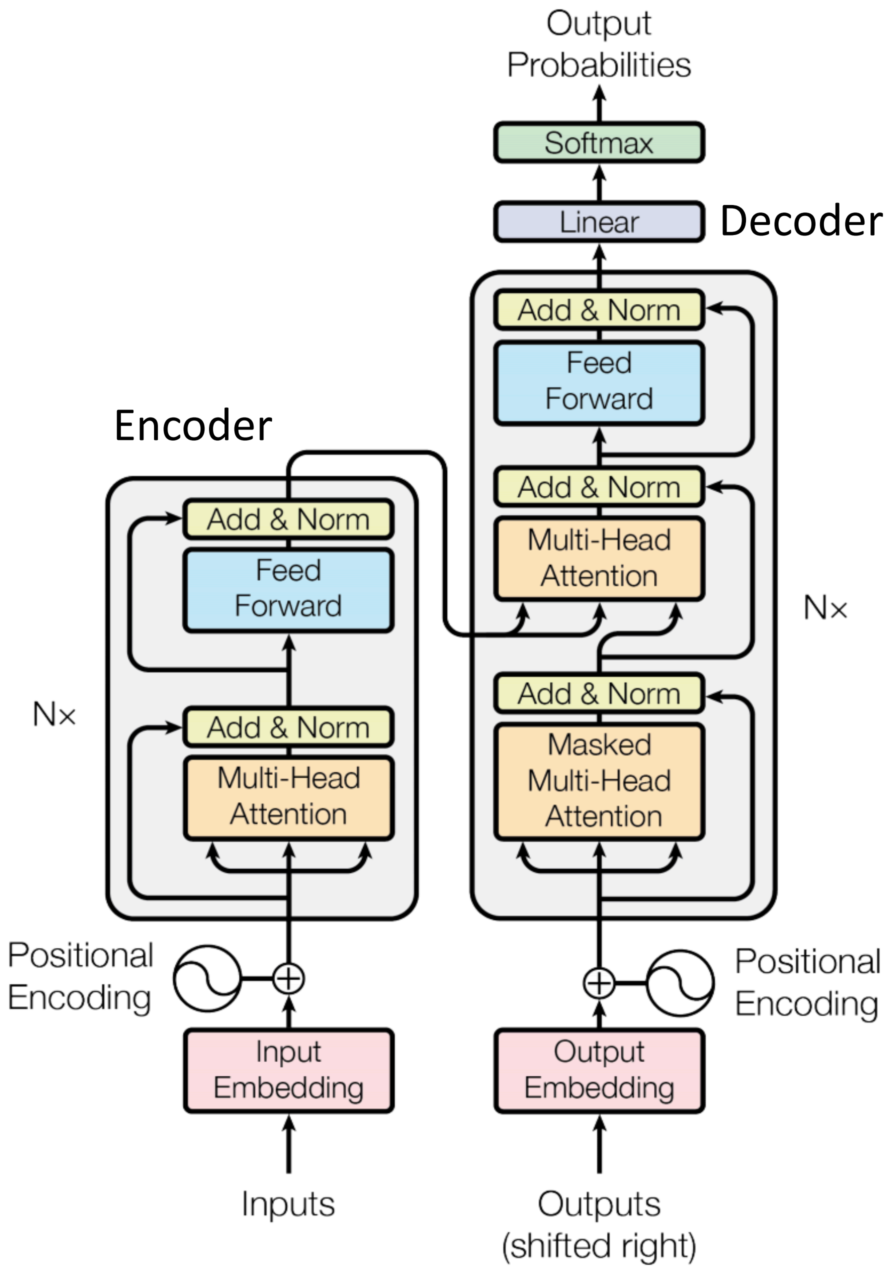


Figure 1: The Transformer - model architecture.

Architecture – scope

- Machine translation
  je suis étudiant $\longrightarrow$ *I am a student*
- Sentences length ~ 100
- Resources for parallelization
- Stacked

1. Encoder Self-attention
2. Decoder Self-attention
3. Encoder-Decoder attention (global attention)

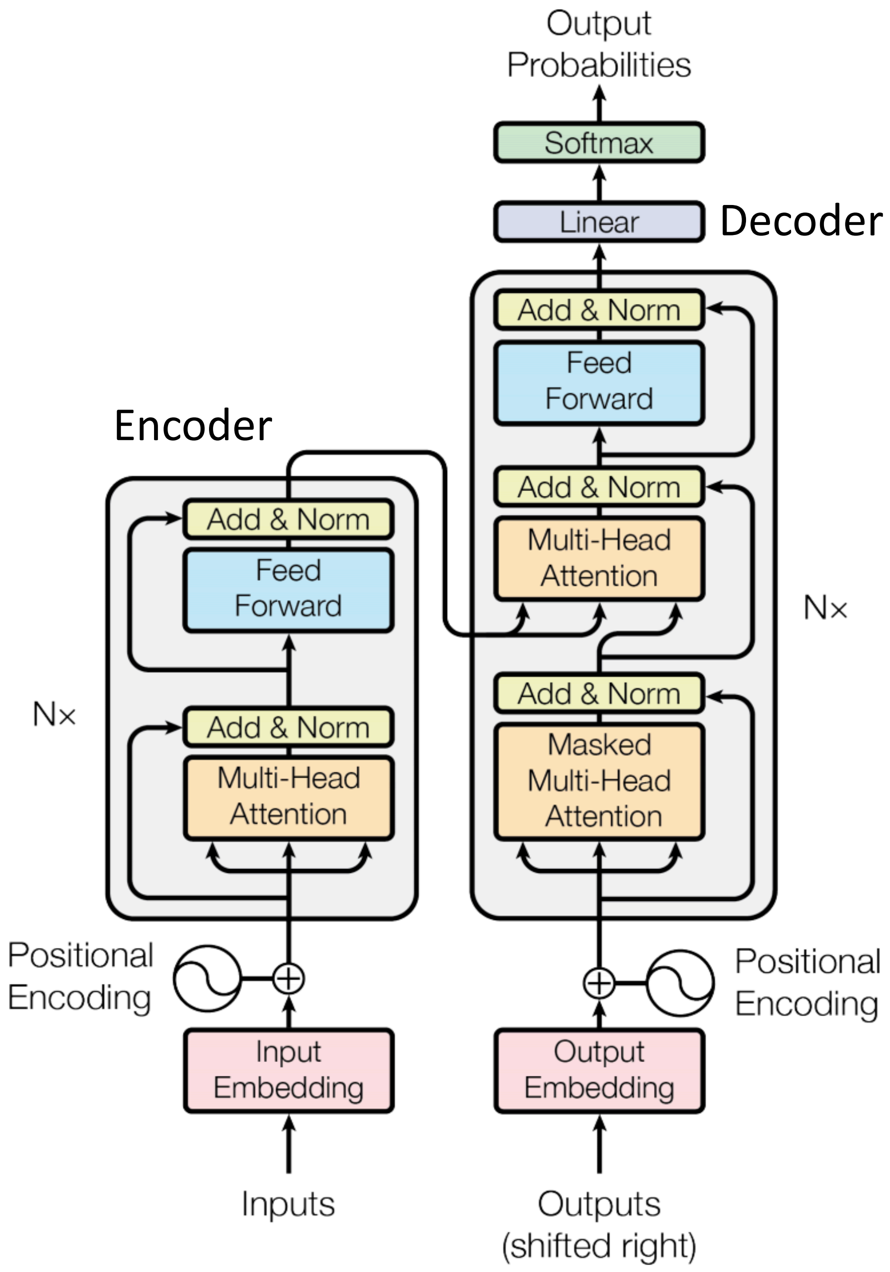
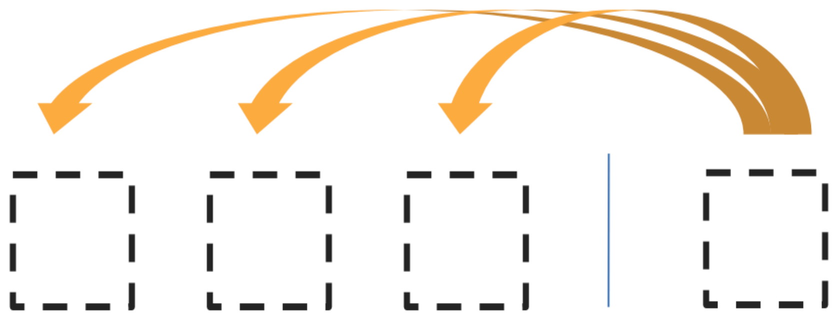
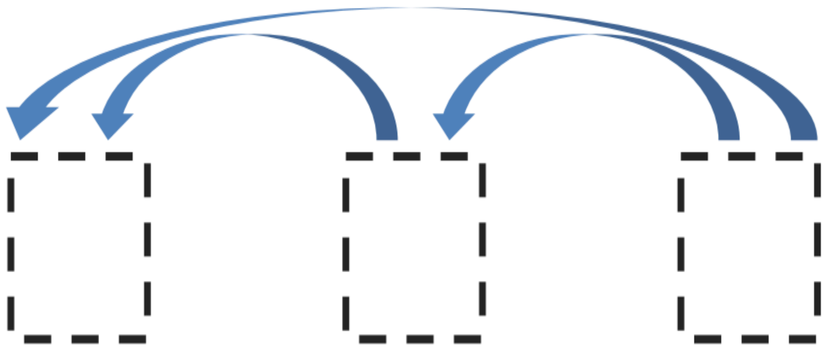


Figure 1: The Transformer - model architecture.



Encoder-Decoder Attention

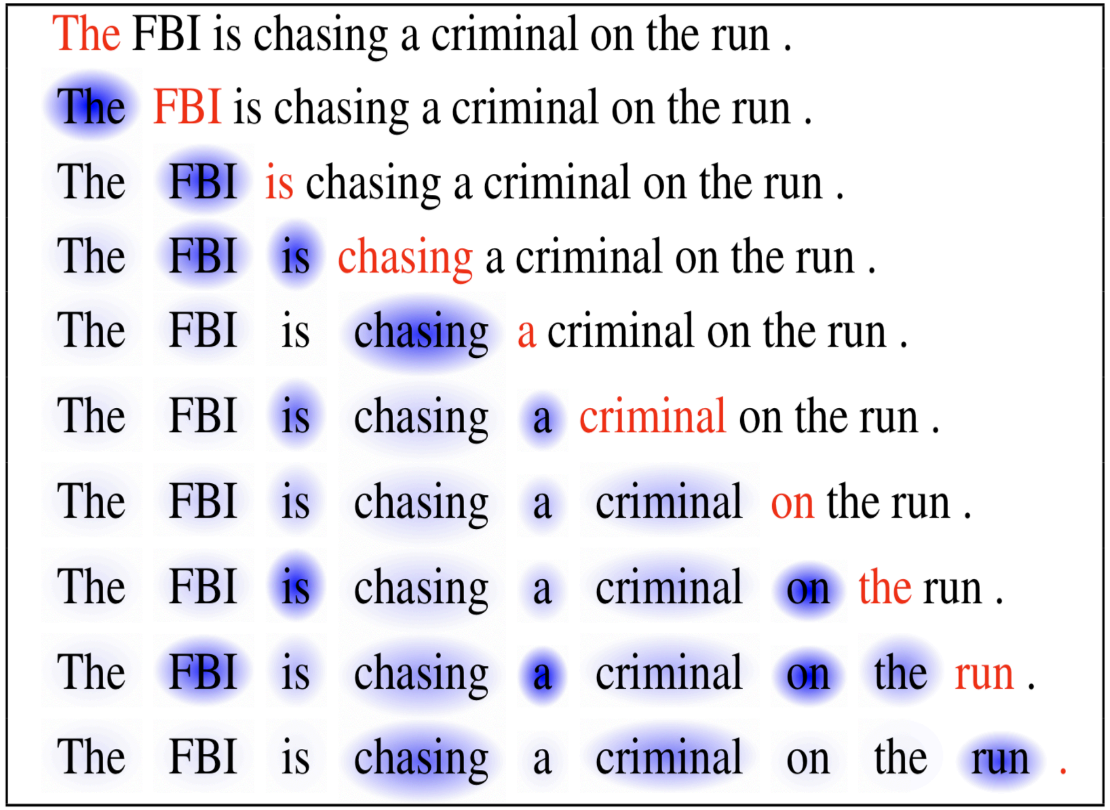
Encoder Self-Attention



MaskedDecoder Self-Attention

Architecture – self-attention

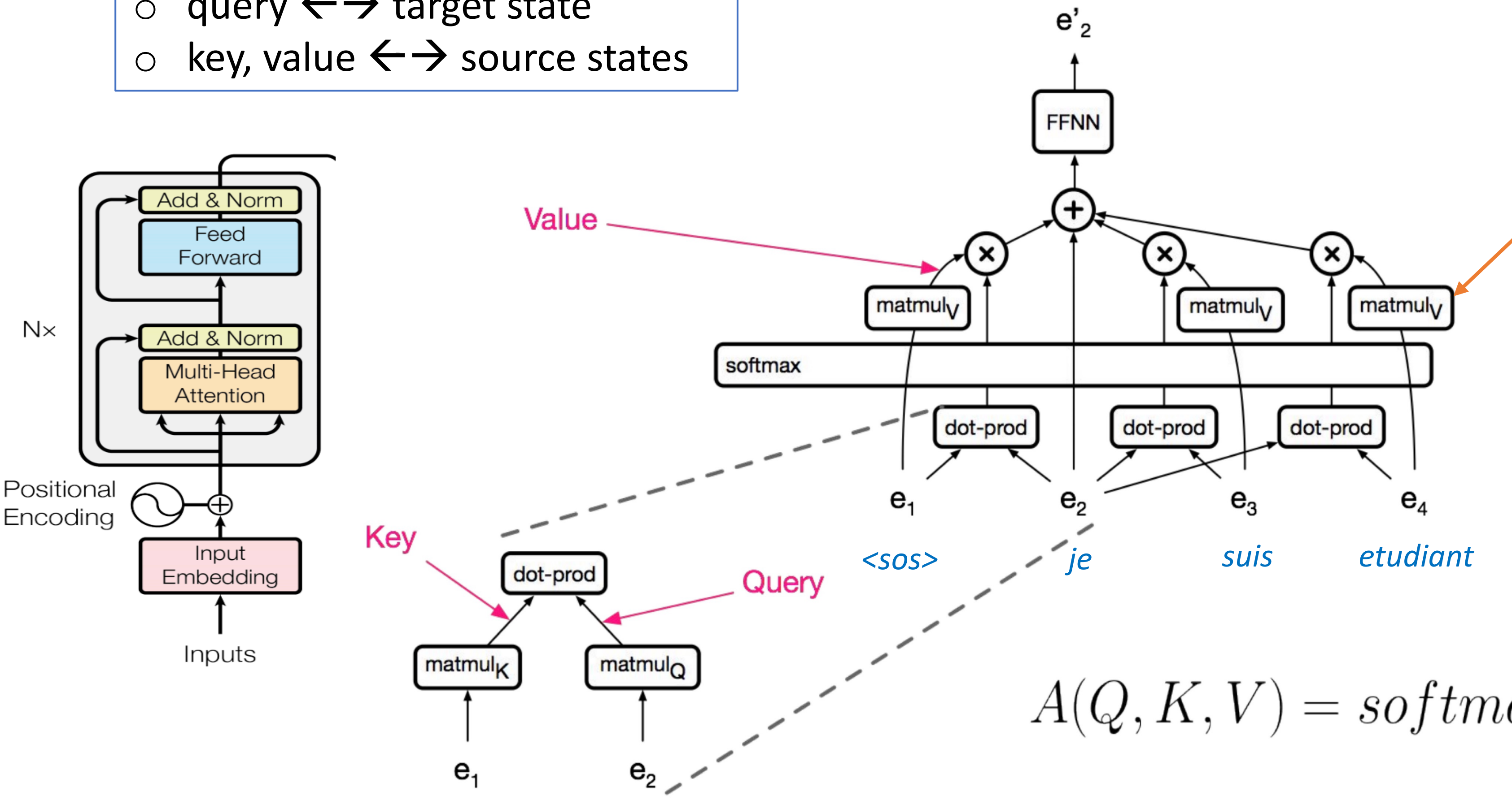
- Representations
- Intra-attention, RNNs
- Constant path length between any two positions *#intuition*
- Refer by content *#motivation*
- Multiplicative interactions *#motivation*



Removes recurrence completely!

Architecture – encoder self-attention

- query ←→ target state
- key, value ←→ source states

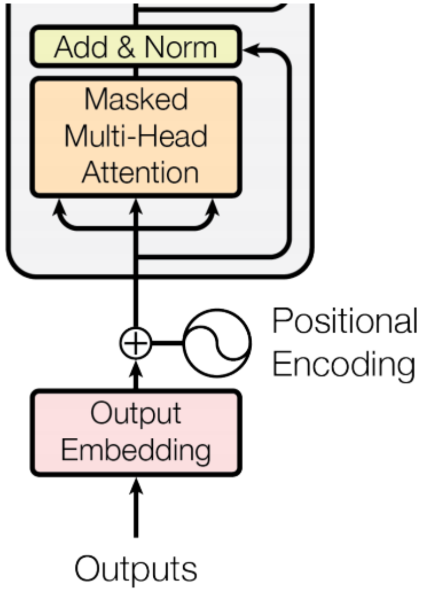


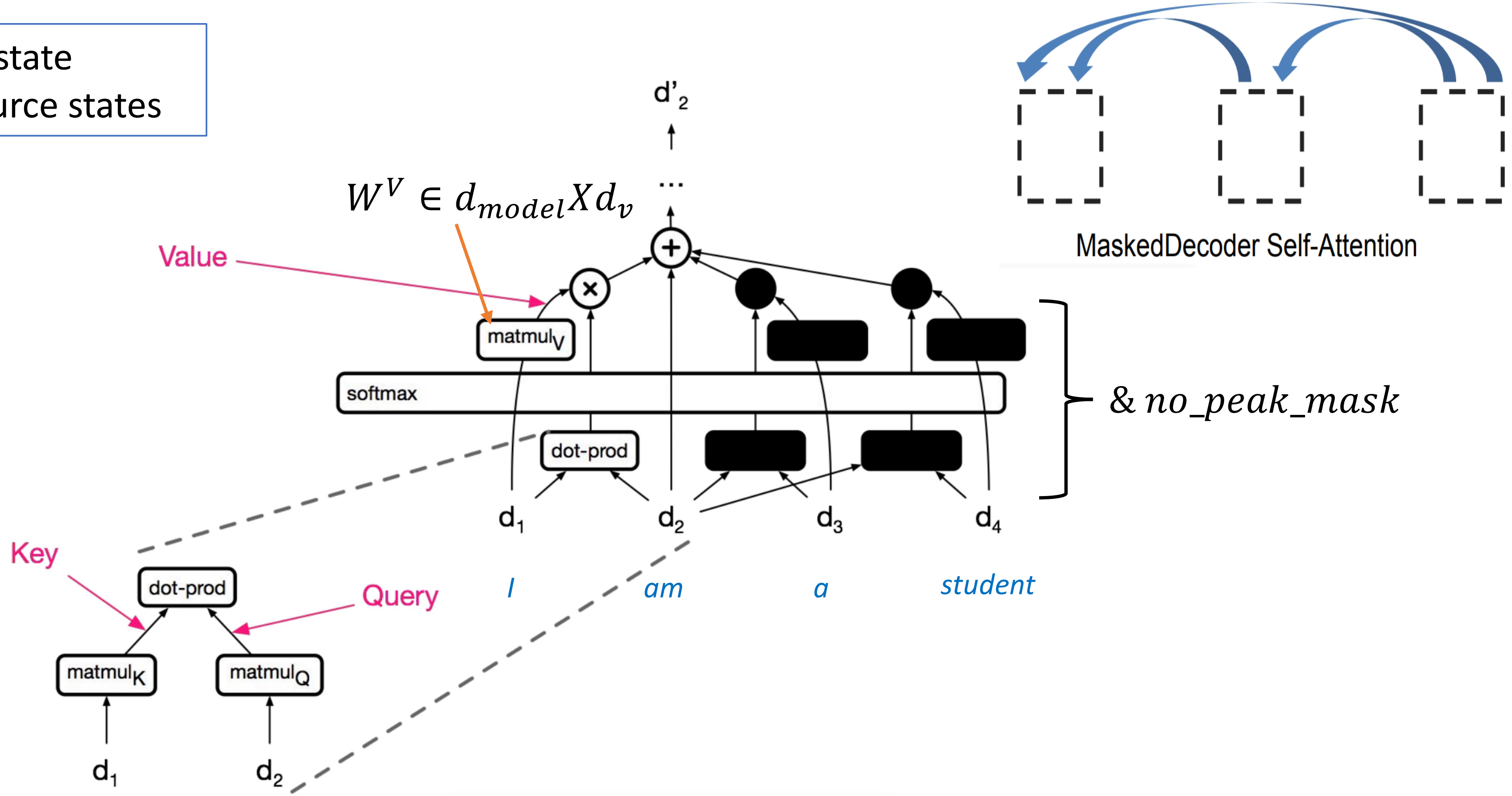
$W^V \in d_{model} X d_v$

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

Architecture – decoder self-attention

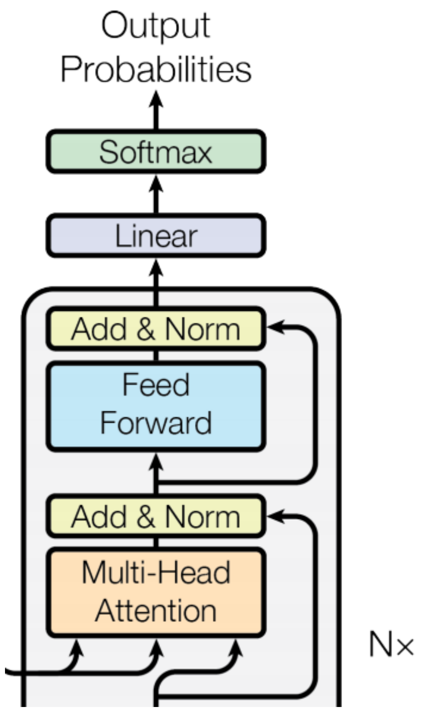
- query ←→ target state
- key, value ←→ source states

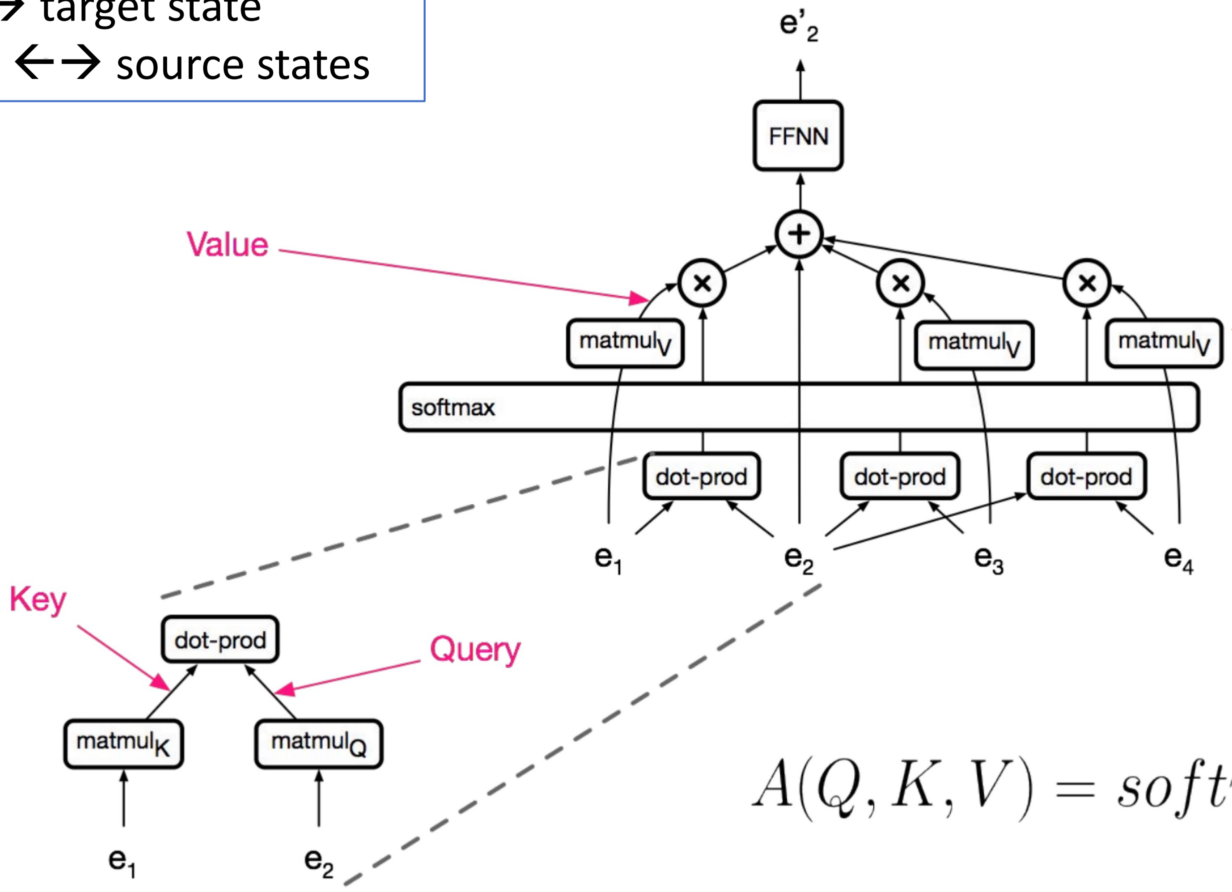




Architecture – encoder-decoder attention

- query ←→ target state
- key, value ←→ source states





$W^V \in d_{model} X d_v$

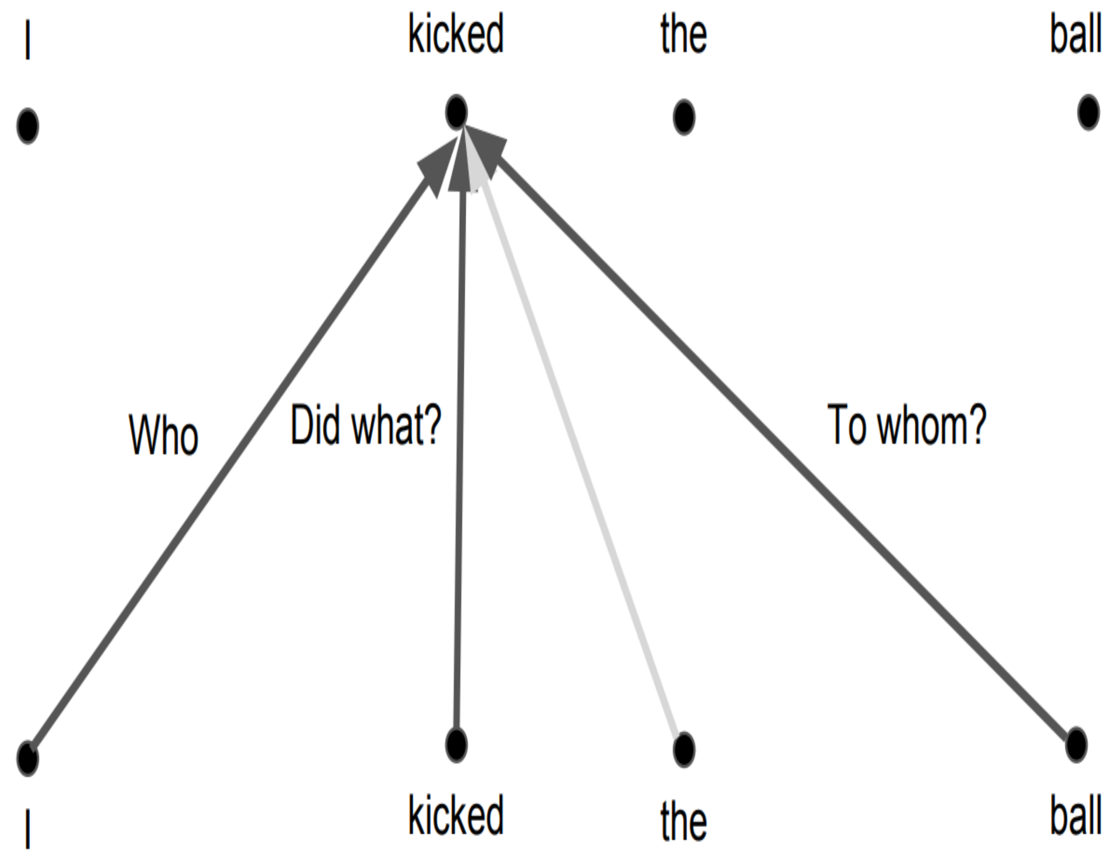
- global attention
- queries → masked output representations
- key, value ←→ encoder states

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

Architecture – scaled dot-product attention

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

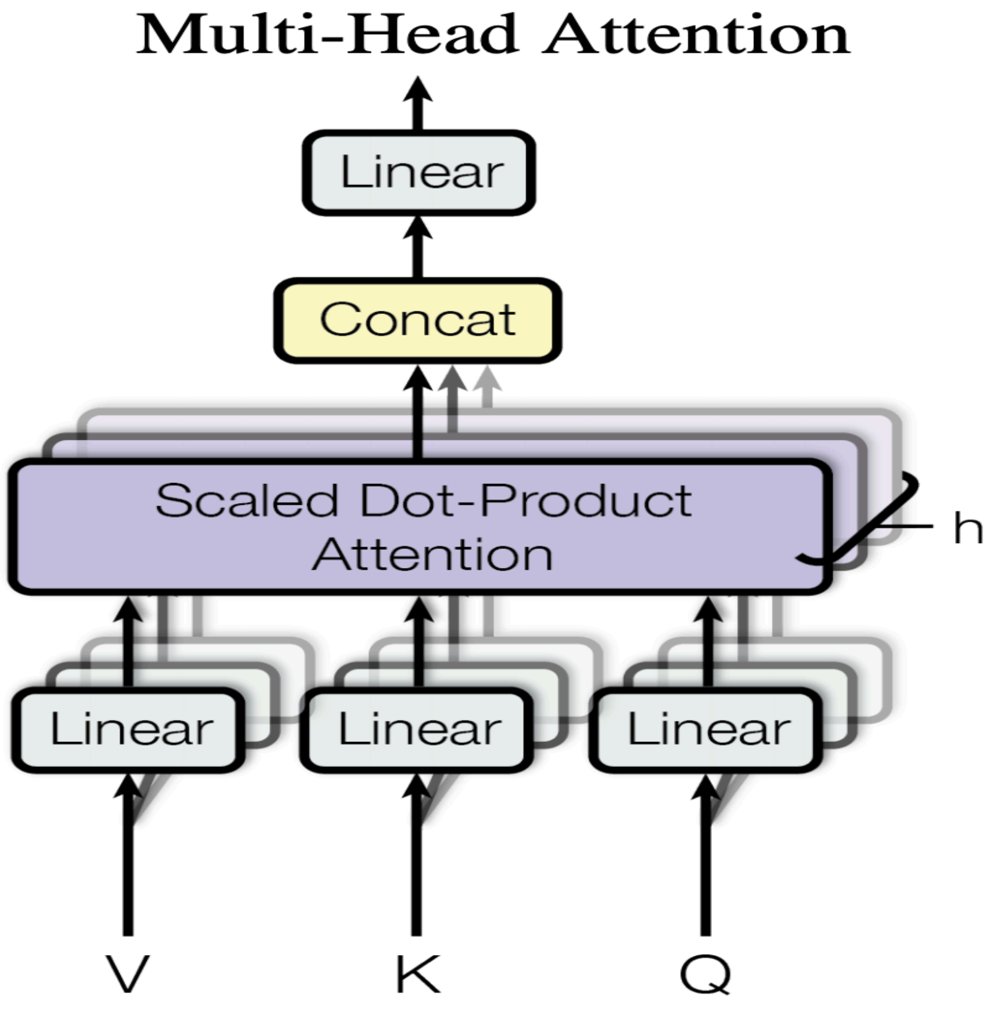
- $\frac{1}{\sqrt{\{d_k\}}}$ to account for large inputs
- ~~Hard to parallelize *#bottleneck*~~
- Alignment problems, local-global information *#case*
- Same linear projection → head



Architecture – multi-head attention

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

- Parallel heads (distributions)
- W^o =accounts for capturing information from all attention heads
- Overhead = Linear projection + SoftMax



Architecture – other details

- Positional encoding – sinusoids
- Residual connections
- Layer normalization
- 8 heads, 6 layers
- Adam optimizer
- Label smoothing

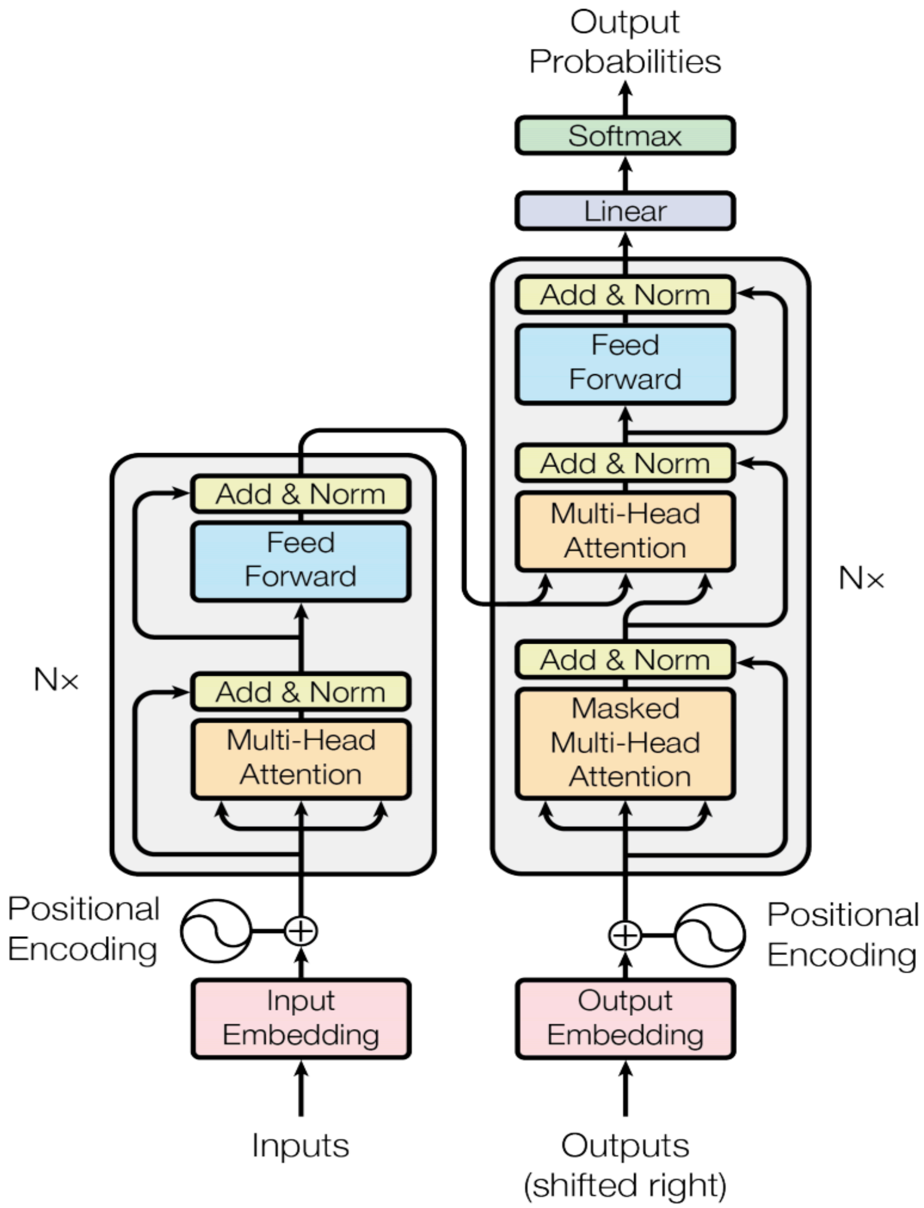


Figure 1: The Transformer - model architecture.

Results

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

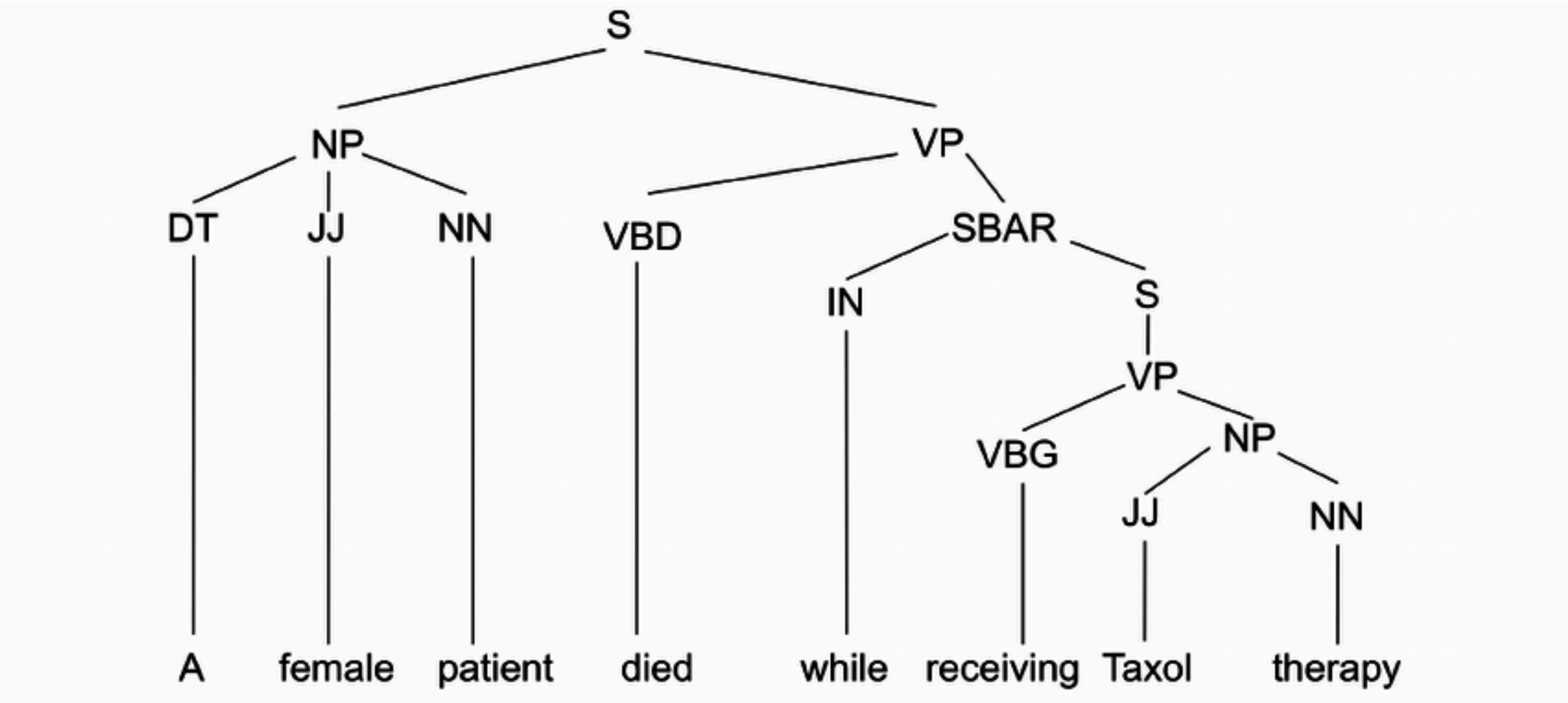
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	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**	**41.8**	$2.3 \cdot 10^{19}$	

Results

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Results


S
NP
VP
DT
JJ
NN
VBD
SBAR
IN
S
VP
VBG
NP
JJ
NN
A
female
patient
died
while
receiving
Taxol
therapy

Implications

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Universal Transformers
- A Study of Reinforcement Learning for Neural Machine Translation

References

https://arxiv.org/abs/1706.03762 – publication

http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture14-transformers.pdf – lecture by author

http://jalammar.github.io/illustrated-transformer/ – blog post

http://nlp.seas.harvard.edu/2018/04/03/attention.html – tutorial

Thank you for your “attention!”