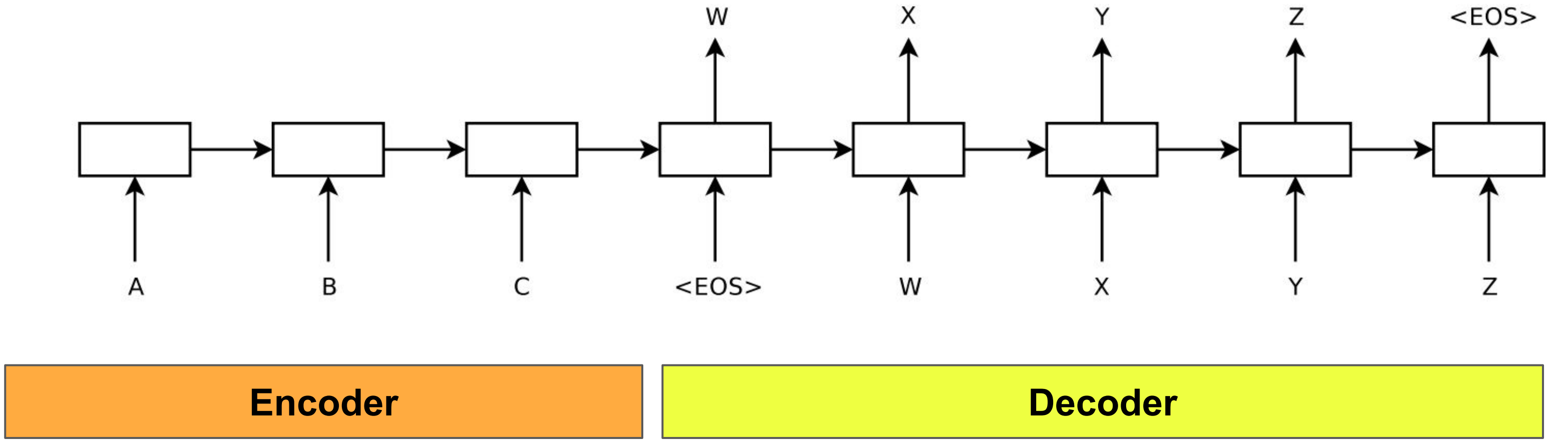
Attention is all you need

Vaswani et al, NeurIPS 2017

The need for attention in NMT



Sutskever et al, NeurIPS 2014

Contributions and claims

One contribution:
Model purely based on attention

Claims:
- Better performance on NMT tasks
- Faster to train (easy to parallelize)
- Long-range dependencies

Attention mechanism -- Overview

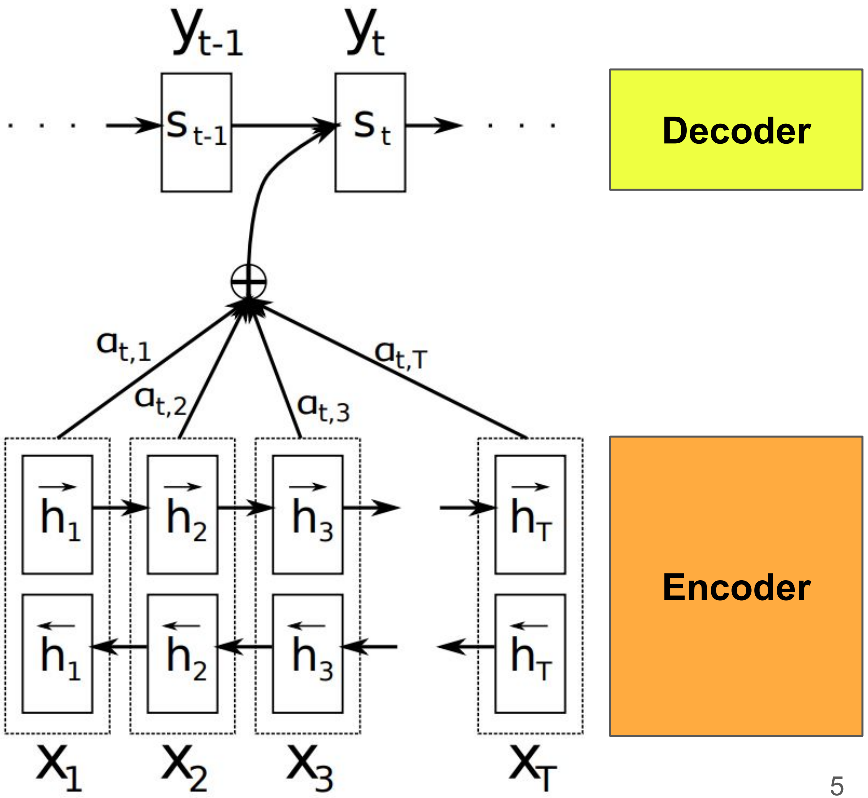
Great blog post:

http://ruder.io/deep-learning-nlp-best-practices/index.html#attention

Attention mechanism -- Overview

$$\mathbf{c}_i = \sum_j a_{ij} \mathbf{h}_j$$

$$\mathbf{a}_i = \mathrm{softmax}(f_{att}(\mathbf{s}_i, \mathbf{h}_j))$$



Bahdanau et al, ICLR 2015

Attention mechanism -- Additive

$$f_{att}(\mathbf{s}_{i-1}, \mathbf{h}_j) = \mathbf{v}_a^\top \tanh(\mathbf{W}_1 \mathbf{s}_{i-1} + \mathbf{W}_2 \mathbf{h}_j)$$

Bahdanau et al, ICLR 2015

Attention mechanism -- Multiplicative

$$f_{att}(s_i, h_j) = s_i^\top \mathbf{W}_a h_j$$

Luong et al, EMNLP 2015

Attention mechanism -- Self-attention

$$f_{att}(\mathbf{h}_j) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a \mathbf{h}_j)$$

$$\mathbf{a} = \mathrm{softmax}(\mathbf{v}_a \tanh(\mathbf{W}_a \mathbf{H}^\top))$$

$$\mathbf{c} = \mathbf{H}\mathbf{a}^\top$$

Lin et al, ICLR 2017

Attention mechanism -- Self-attention

$$\mathbf{A} = \mathrm{softmax}(\mathbf{V}_a \tanh(\mathbf{W}_a \mathbf{H}^\top))$$
$$\mathbf{C} = \mathbf{AH}$$

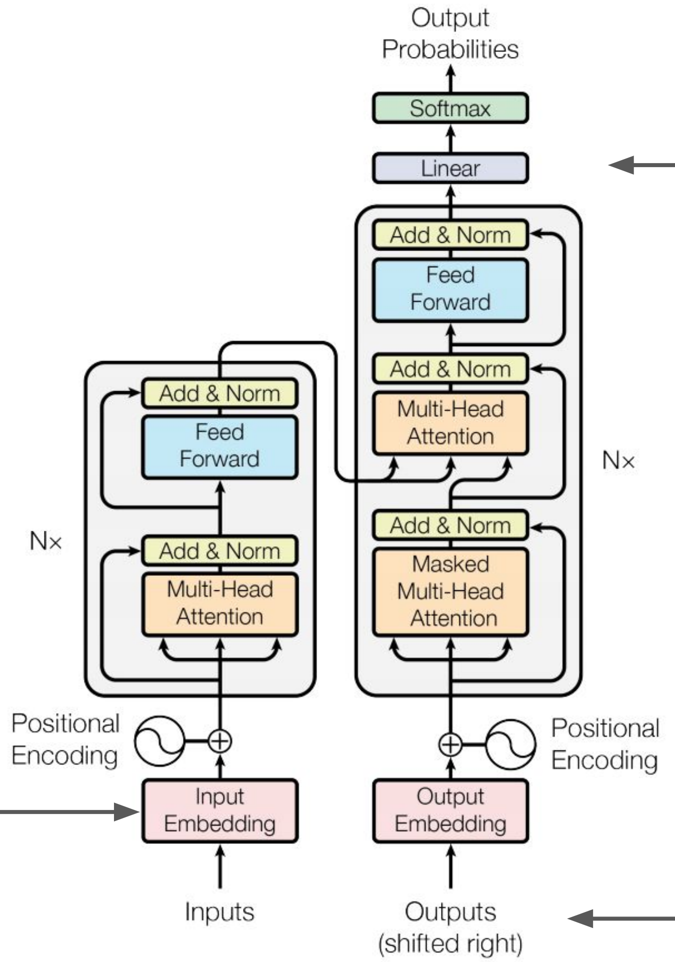
Lin et al, ICLR 2017

Attention mechanism -- Key-value

$$\mathbf{h}_i = [\mathbf{k}_i; \mathbf{v}_i]$$
$$\mathbf{a}_i = \mathrm{softmax}(\mathbf{v}_a^\top \tanh(\mathbf{W}_1[\mathbf{k}_{i-L}; \ldots; \mathbf{k}_{i-1}] + (\mathbf{W}_2\mathbf{k}_i)\mathbf{1}^\top))$$
$$\mathbf{c}_i = [\mathbf{v}_{i-L}; \ldots; \mathbf{v}_{i-1}]\mathbf{a}^\top$$

Daniluk et al, ICLR 2017

Transformer networks



Normalized linear projection

Layer normalization

2-layer NLP

Layer normalization

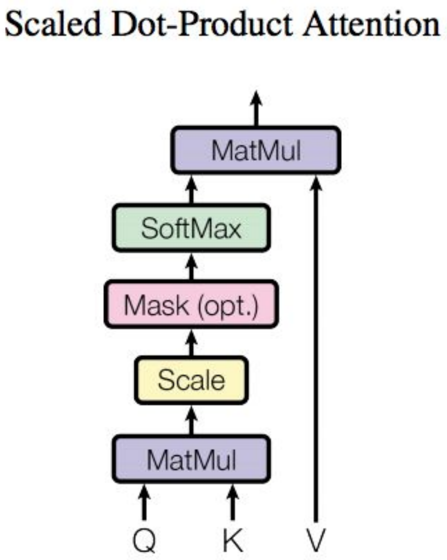
Encode the absolute position

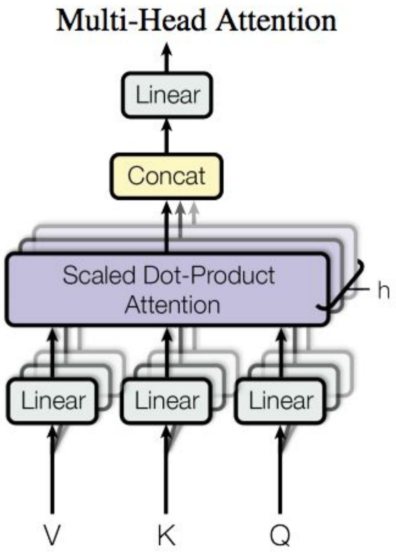
Scaled linear projection

Outputs of the previous timesteps

Transformer networks

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





Subspaces

Previous states of the decoder

Hidden states of the encoder

Ablation study on NMT

	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

Other NLP tasks

Subject-verb agreement (long-range dependency)

English:	[...] plan will be approved
German:	[...] **Plan** verabschiedet **wird**
Contrast:	[...] **Plan** verabschiedet **werden**

RNNs are better than CNNs and Transformers

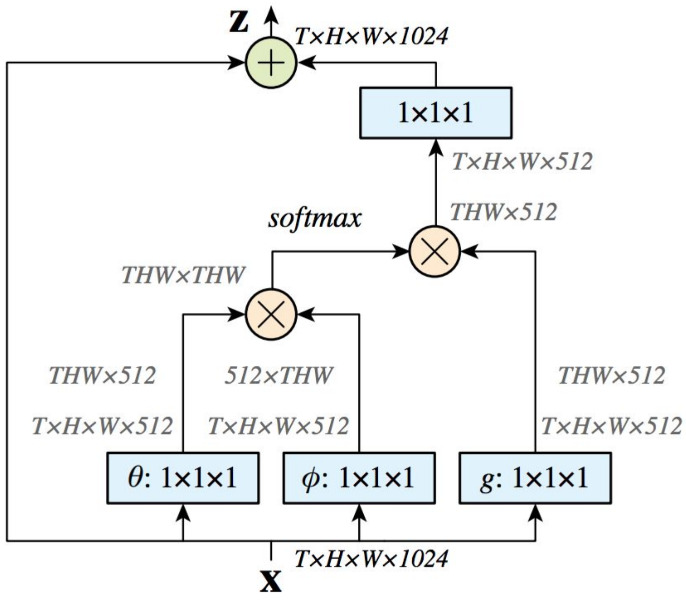
Word sense disambiguation (semantic features)

source:	*Er hat zwar schnell den Finger am* ***Abzug****, aber er ist eben neu.*
reference:	*Il a la* ***gâchette*** *facile mais c'est parce qu'il débute.*
contrastive:	*Il a la* ***soustraction*** *facile mais c'est parce qu'il débute.*
contrastive:	*Il a la* ***déduction*** *facile mais c'est parce qu'il débute.*
contrastive:	*Il a la* ***sortie*** *facile mais c'est parce qu'il débute.*
contrastive:	*Il a la* ***rétraction*** *facile mais c'est parce qu'il débute.*

Transformers are better than CNNs and RNNs

Tang et al, EMNLP 2018

Non-local networks, Wang et al CVPR 2018



Key **Query** **Value**

Applied to action recognition in videos
(Kinetics and Charades datasets)

Applied to object segmentation/detection and keypoint detection in images
(MS-COCO dataset)

Wang et al, CVPR 2018