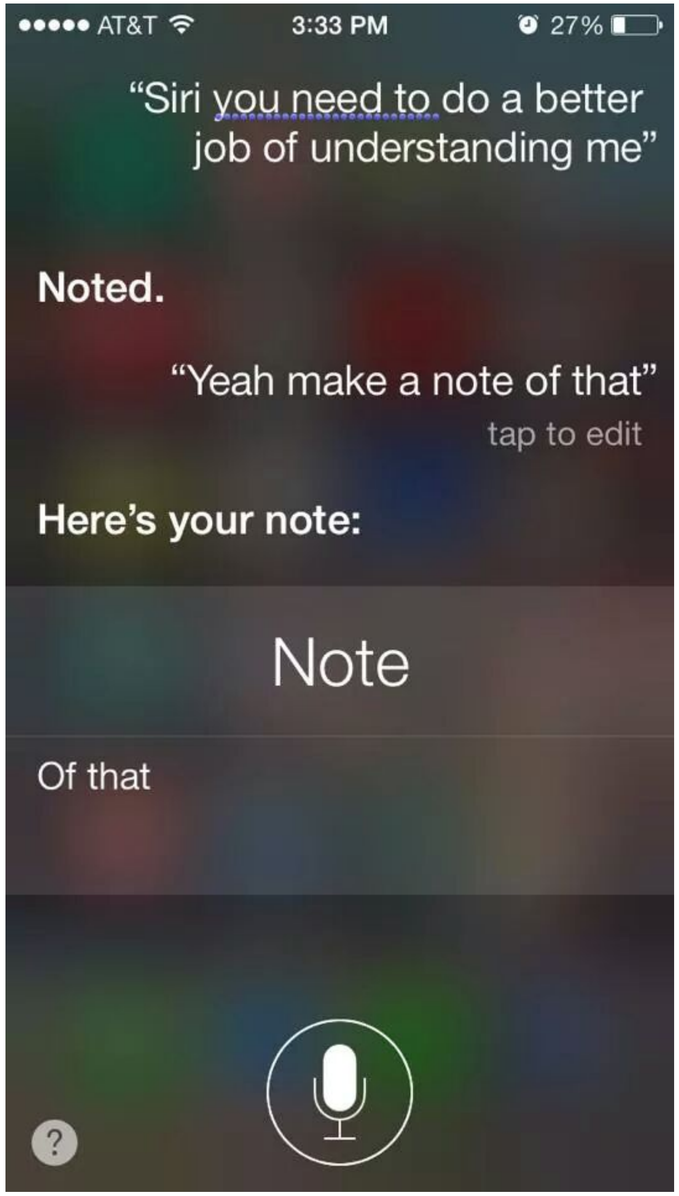
NLP: Attention & Transformer

Seminar in Deep Neural Networks

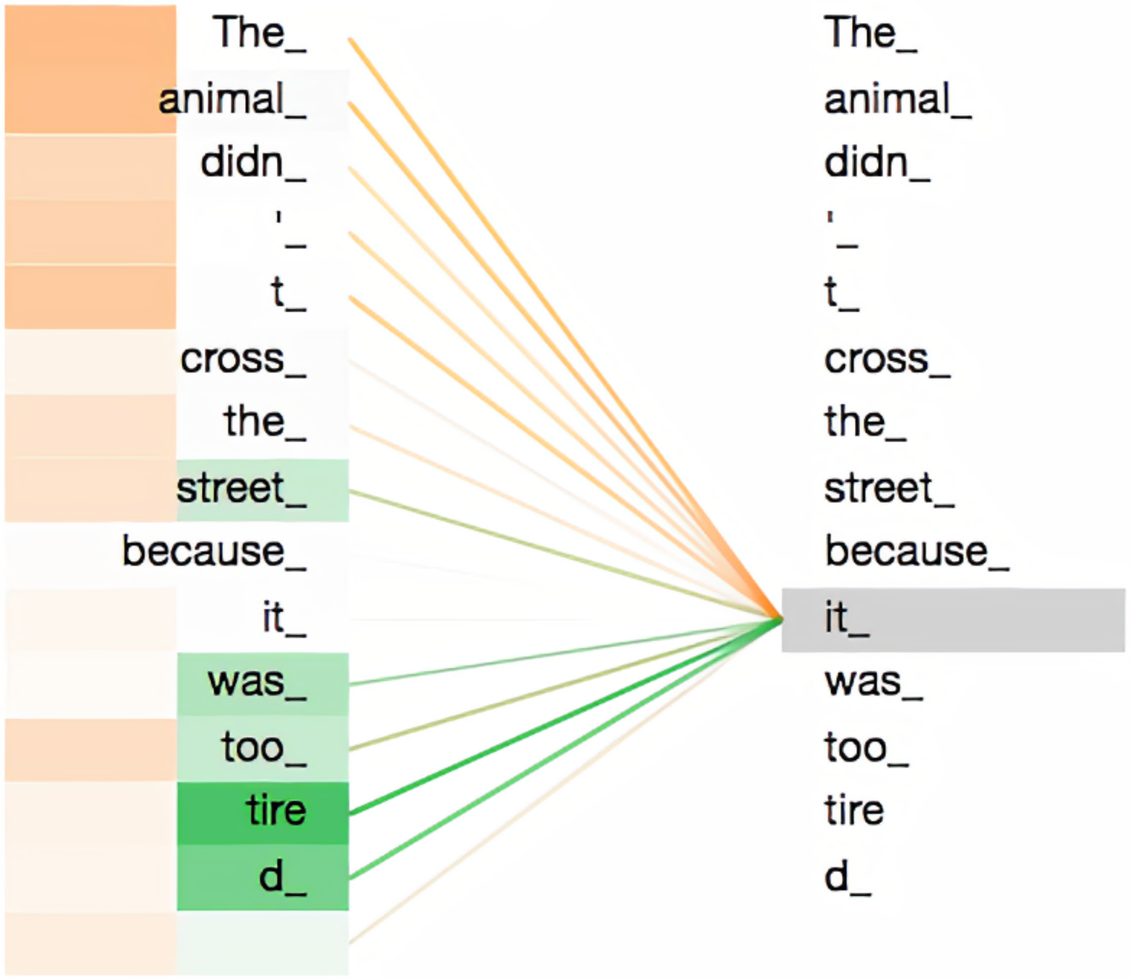
Mihai Zorca, 16.03.2021

Source:
https://data.embeddedcomputing.com/uploads/articles/wp/4992/files-aHViPTYzODY3JmNtZD1pdGVtZWRpdG9yaW1hZ2UmZmlsZW5hbWU9aXRlbWVkaXRvcmltYWdlXzVhZjlmNWM3MWQ2MGMuanBnJnZlcnNpb249MDAwMCZzaWc9NjlhZjdjMTYxZTNmZWViMDJlMDJiYjQzM2VmZGU0ZTc253D



Source:
https://i.imgur.com/t7zJVq9.jpg

Attention



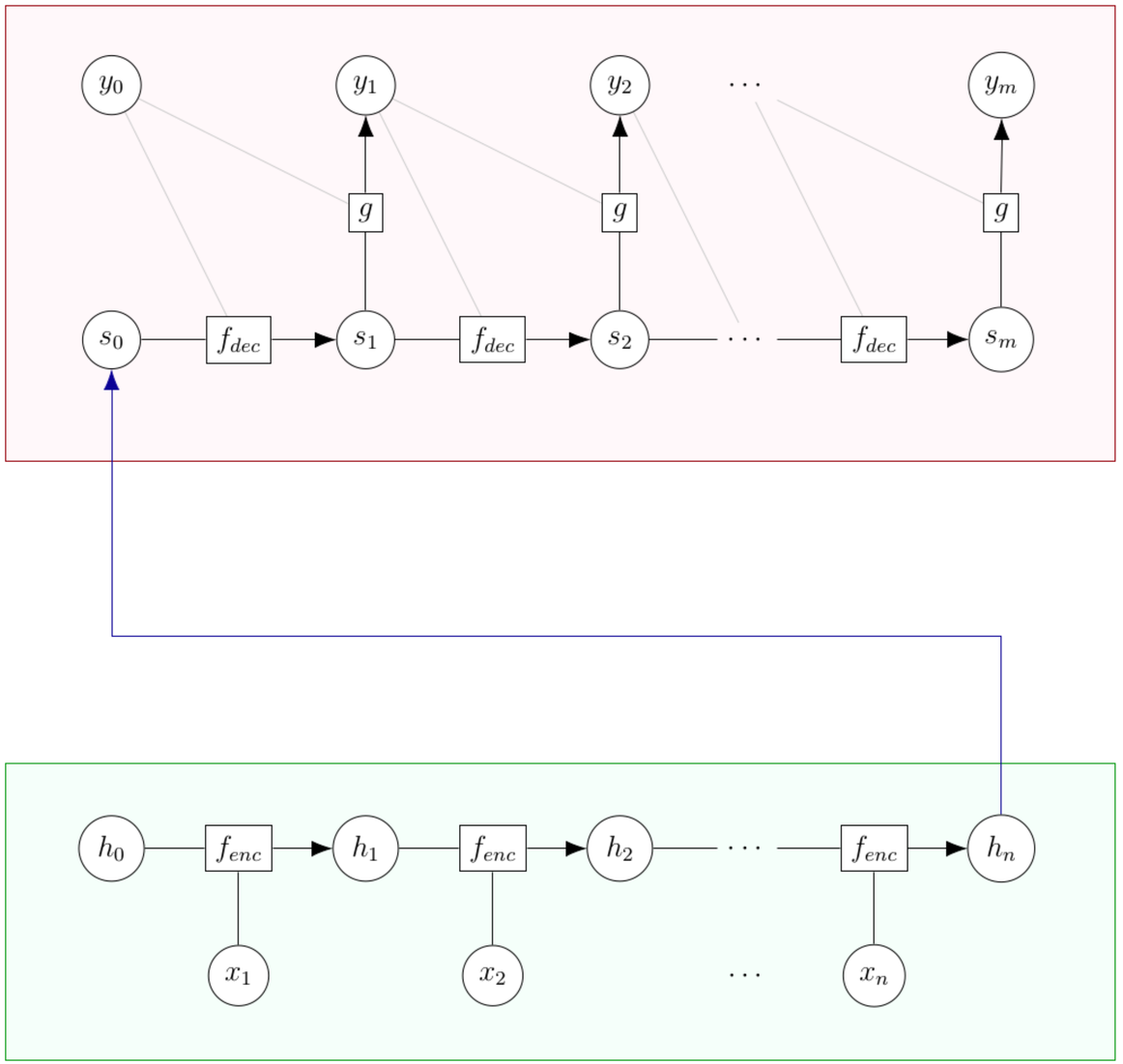
Source: http://jalammar.github.io/illustrated-transformer/

Before Attention

Input sequence encoded into a single vector

Encoding has fixed length

RNNs have trouble remembering long-range dependencies



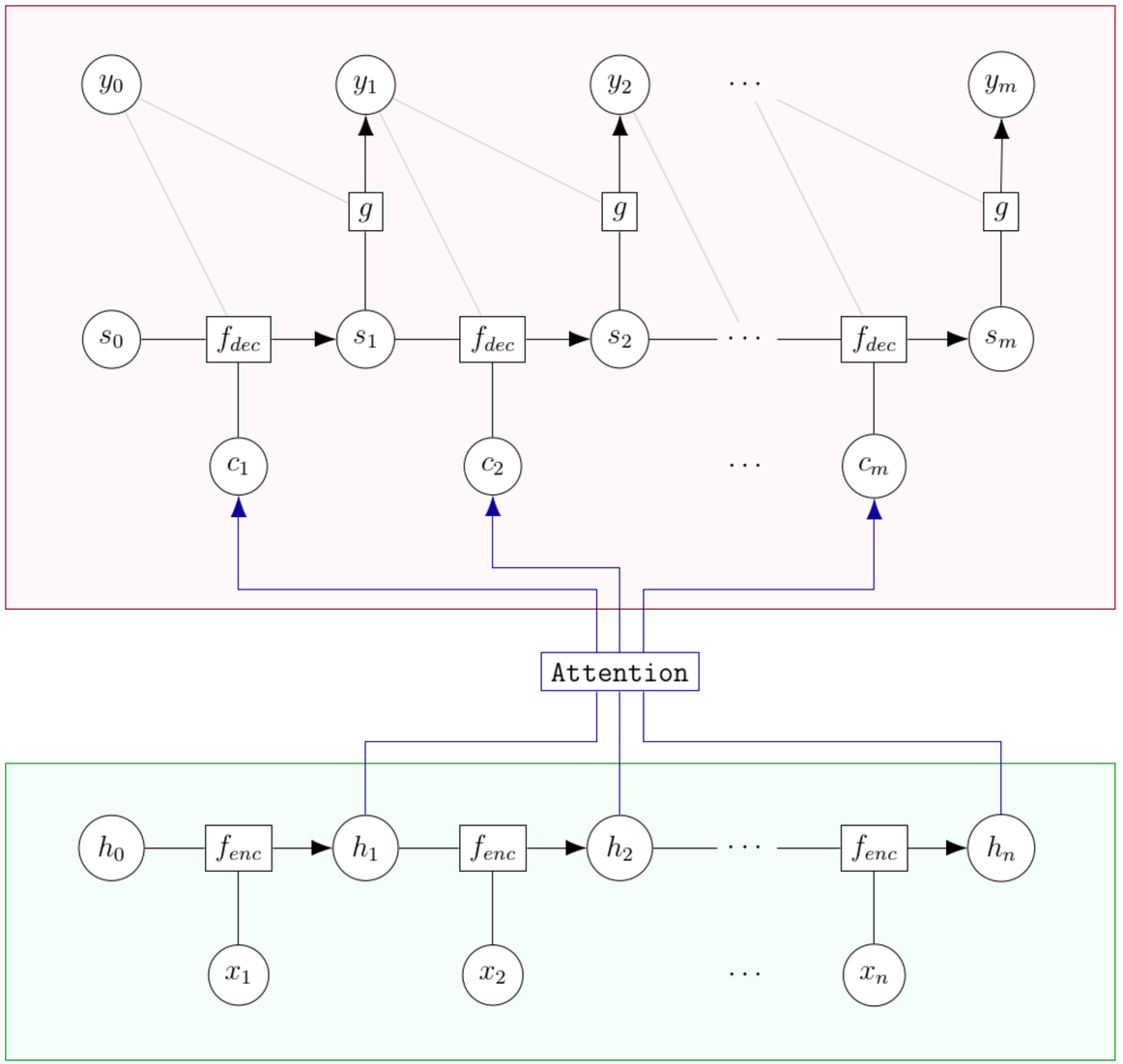
With Attention

Input sequence encoded into a sequence of vectors

Attention is a weighted sum

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

$$\alpha_{ij} = \text{softmax}(e_{ij})$$

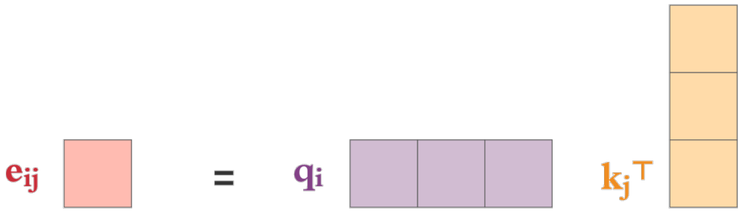


Dot-Product Attention

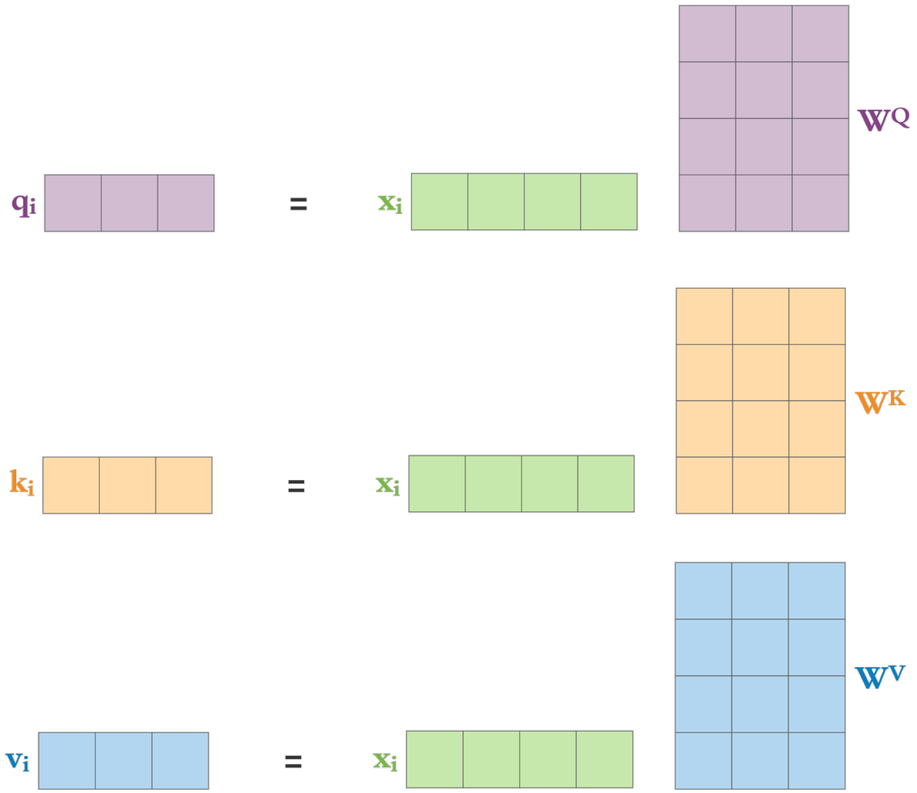
Element-wise

Turn each input x_i into query, key and value vectors

Then $e_{ij} = q_i k_j^\top$



$$c_i = \sum_{j=1}^{n} \text{softmax}(e_{ij}) v_j$$

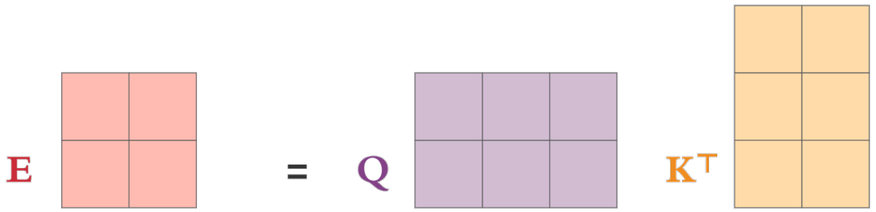


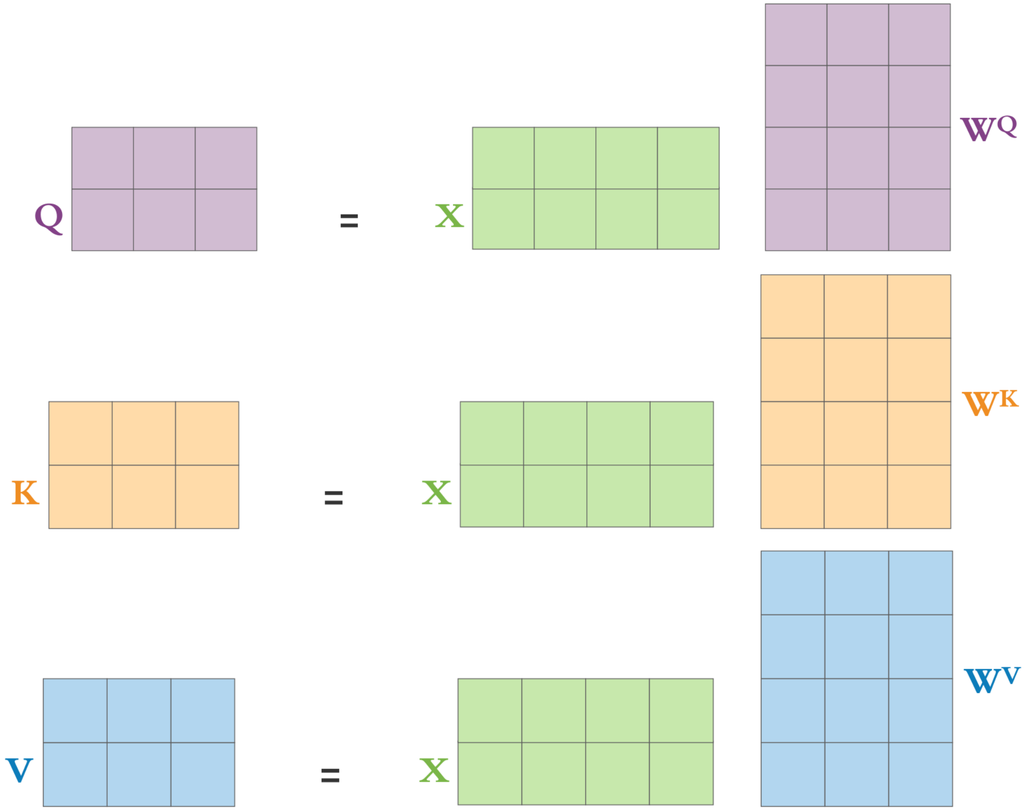
Dot-Product Attention

Matrix notation

Stack all inputs x_i into query, key and value matrices.

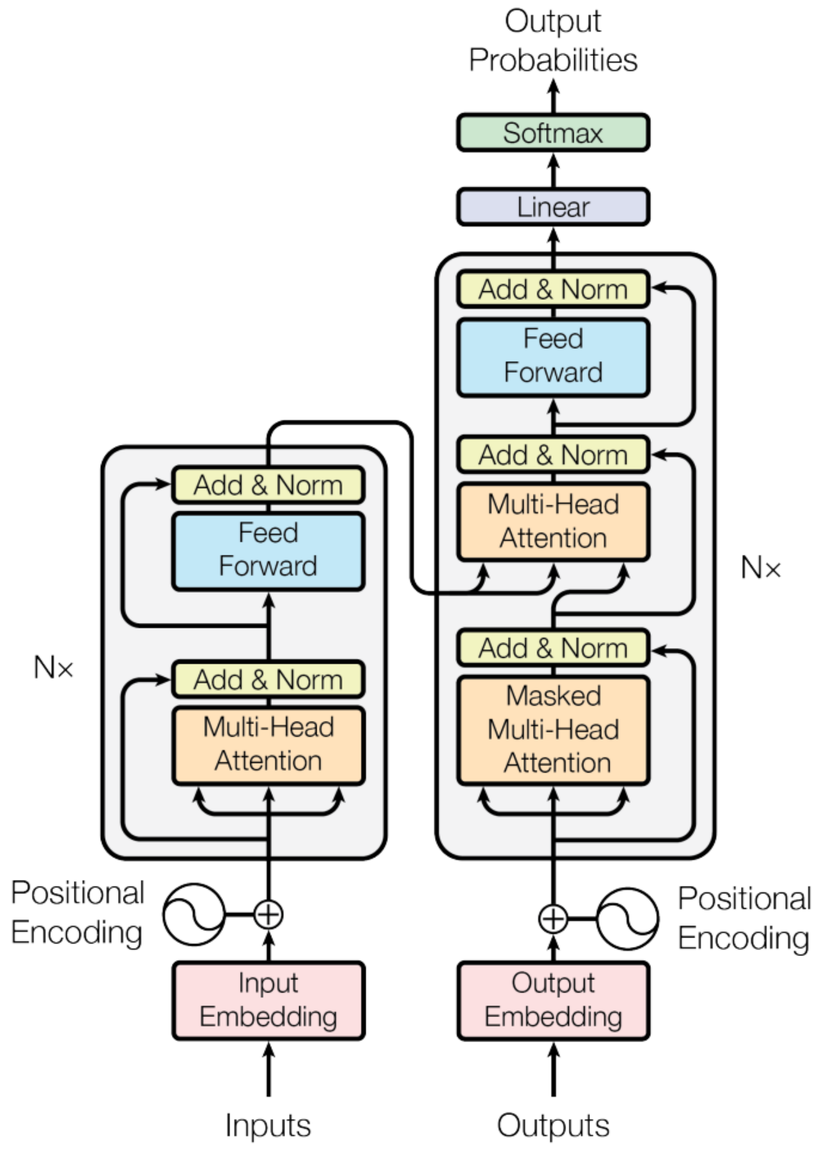
Then $E = QK^\top$





Transformer

"I heard you like Attention…"



Source:

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Transformer

Some Definitions

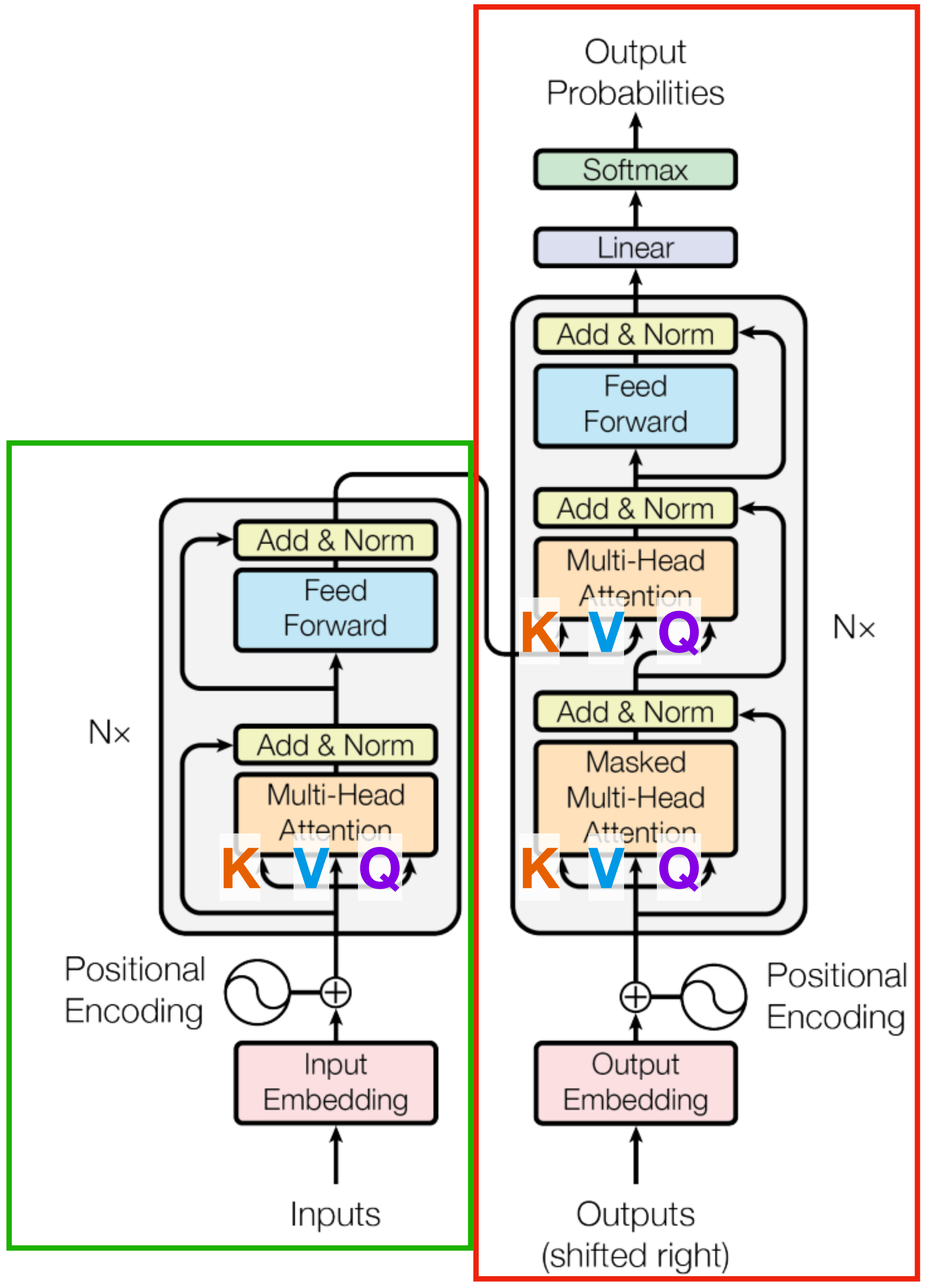
Self-Attention:

Q, **K**, **V** generated from the **same** input **X**

Multi-Head Attention:

Attention multiple times in parallel

Scaled Dot-Product Attention:



Source: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Transformer

More Model Details

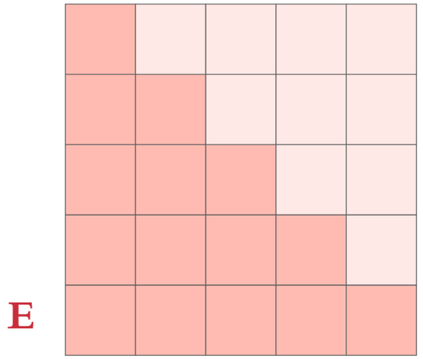
Have N = 6 layers each for encoder and decoder

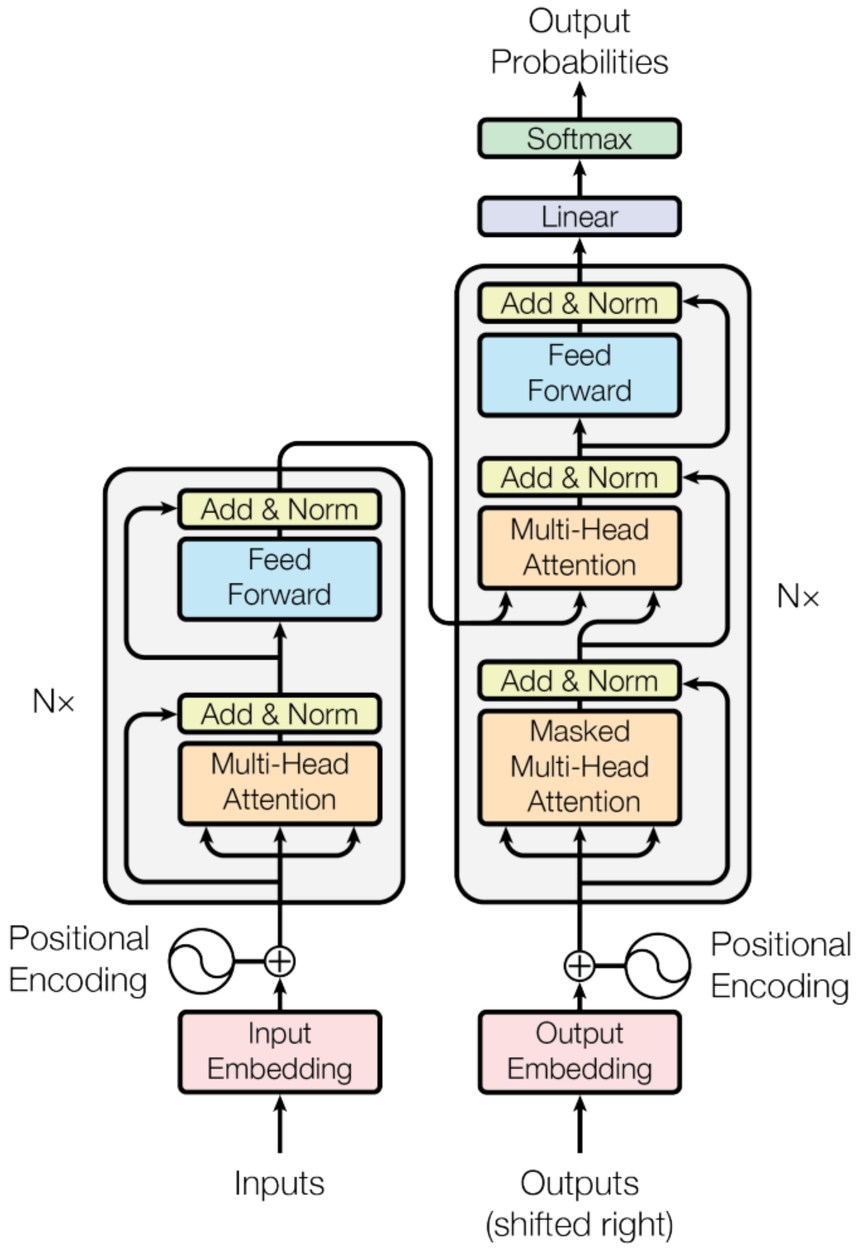
Residual connections parallel to all attention/ff layers

Positional Encoding

Add position-dependent value to the embedding

Masked Attention:





Source: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Multi-Head Attention

Run attention heads in parallel

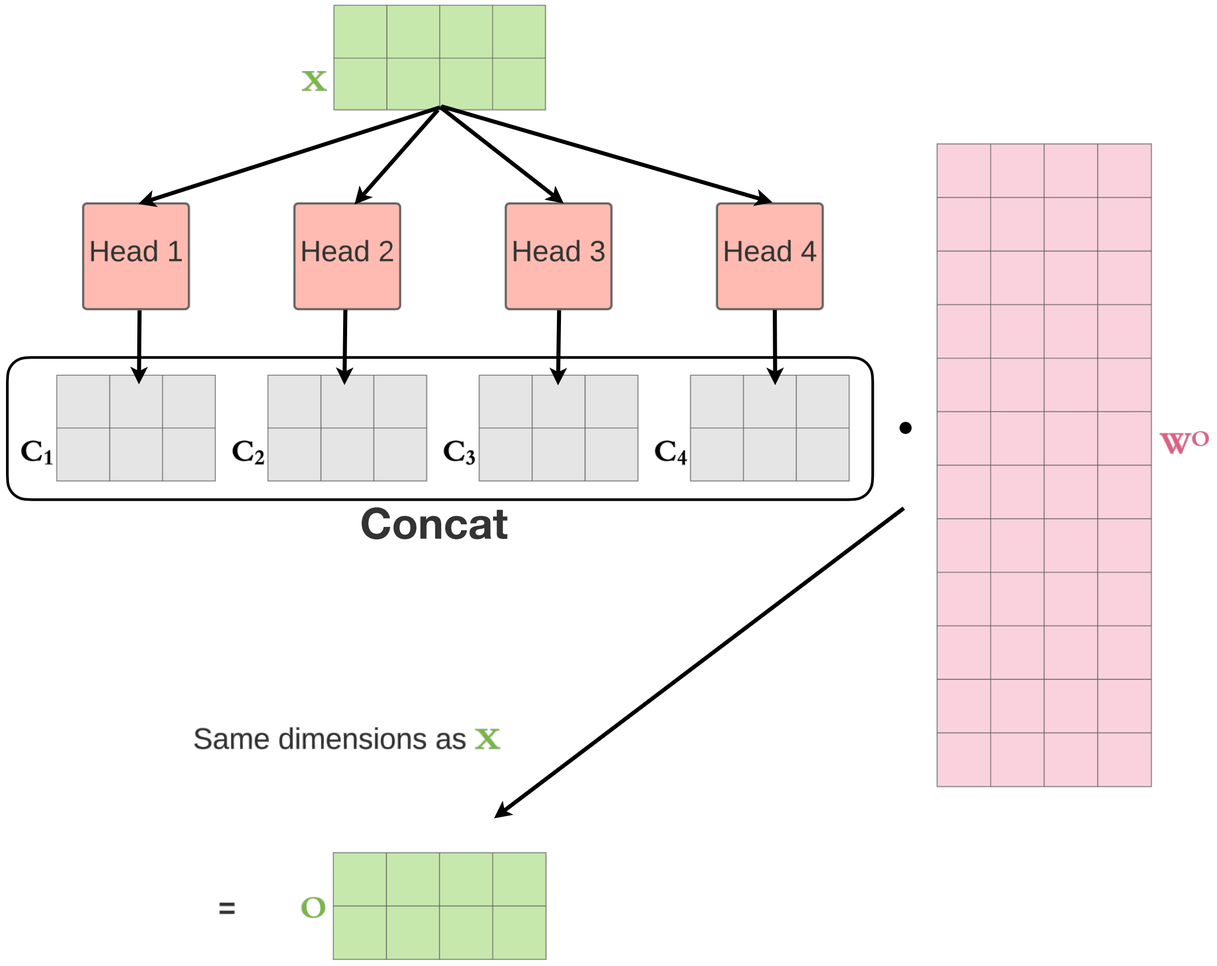
Concatenate their output

Obtain the output **O** with the same dimensions as **X**

Concrete values:

dim(**X**) = 512

dim(**Head i**) = 64



Why Transformer?

Efficient algorithmically (for small n)

$O\left(n^2 d\right)$ vs. $O\left(nd^2\right)$ for RNNs

Removes sequential bottlenecks

Attention is just Mat-Mat-mul :-)

Captures long-range dependencies

Source:
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Transformer Results

Trained on EN-DE (4.5mio sentences) and EN-FR (36mio sentences)

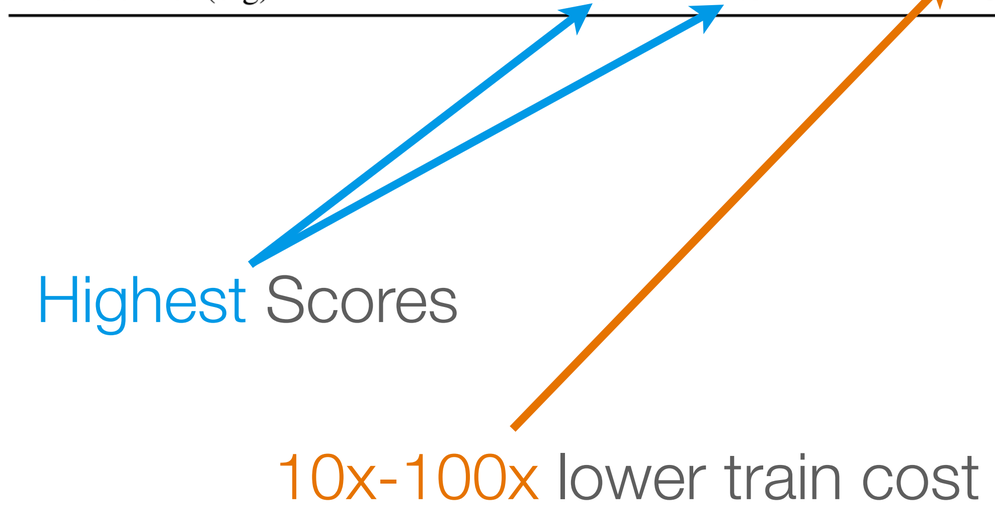
No pre-training

New state-of-the-art, with less training:

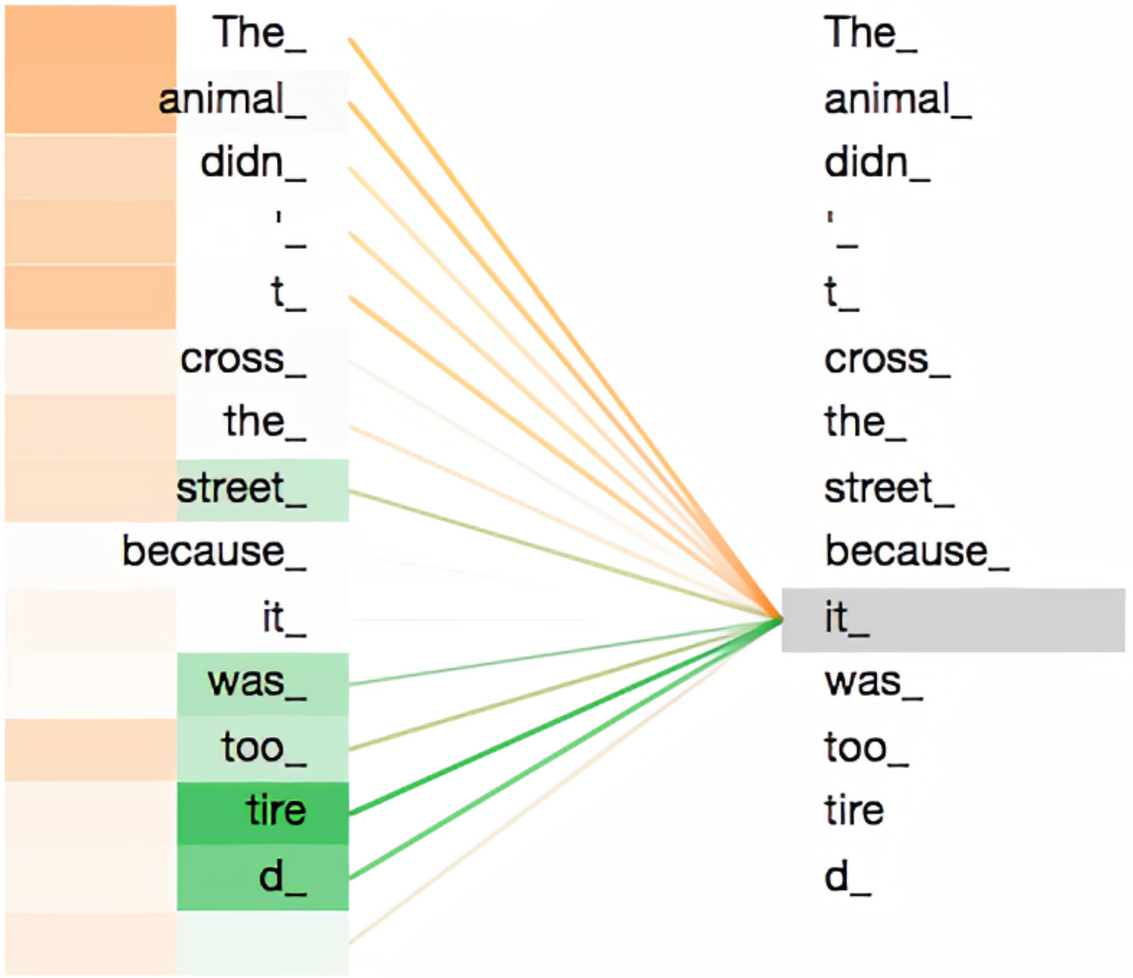
EN-DE: 28.4 BLEU
EN-FR: 41.8 BLEU

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	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Source: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

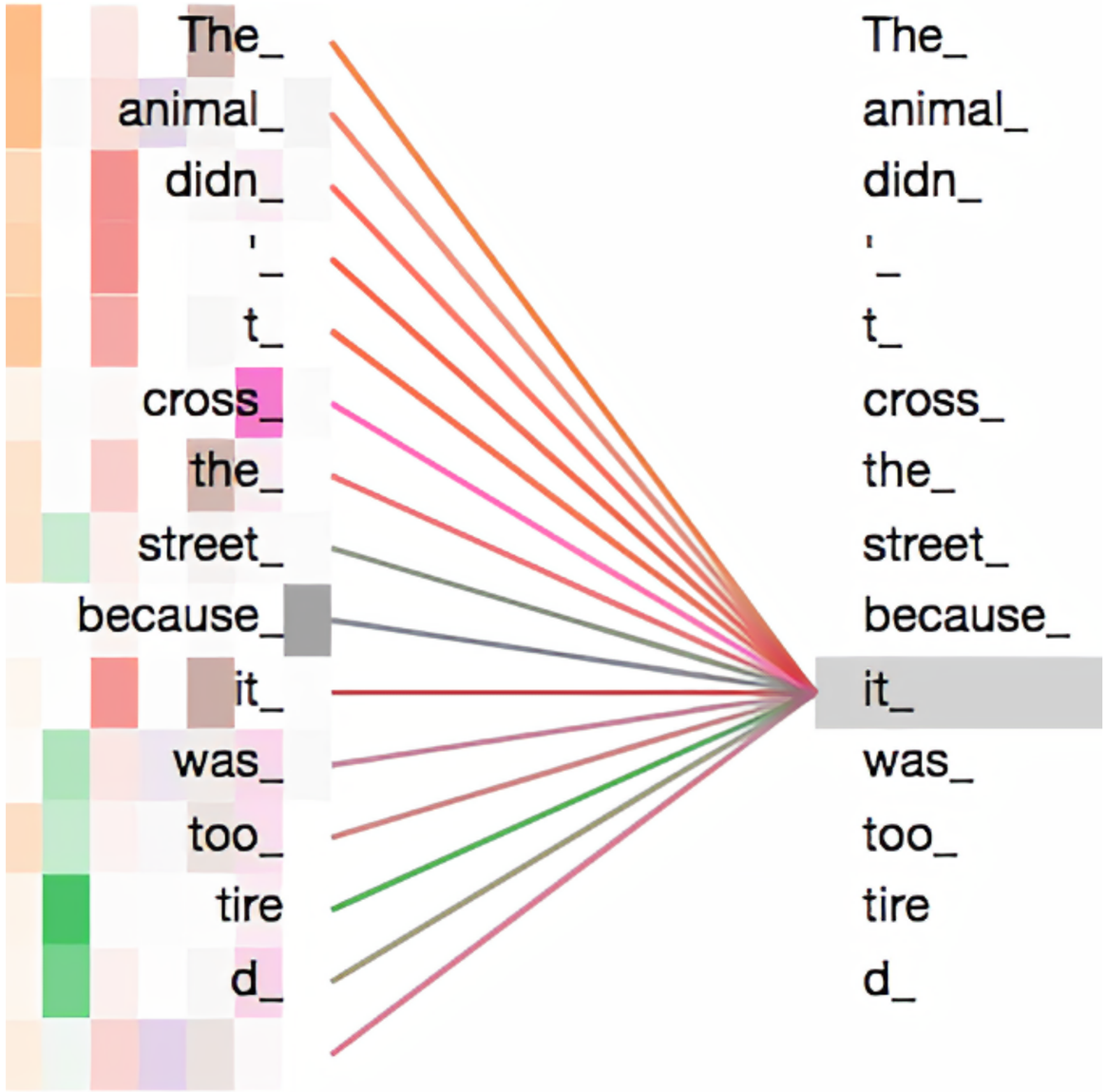


Interpretable



Source: http://jalammar.github.io/illustrated-transformer/

Interpretable?



Source: http://jalammar.github.io/illustrated-transformer/

Attention Identifiability

A short proof

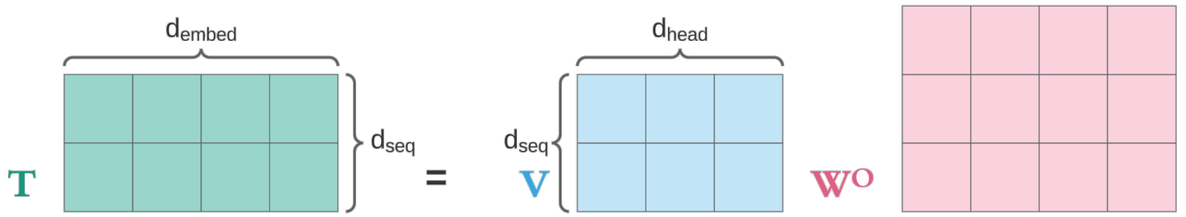
Does the model learn stable representations?

Can we **identify** the attention weights given input & output?

No! (if input seq. is long enough)

Let **A** = $\mathrm{softmax}($ E $)$

Then Attention(**Q**,**K**,**V**) = **A**·**T**



$$\mathrm{rank}(\mathbf{T}) \leq \mathrm{rank}(\mathbf{V}) \leq \min\left(d_{\mathrm{seq}}, d_{\mathrm{head}}\right)$$

=> nontrivial null space, if sequence long

For any **A** in the null space, we have:

(**A** + **A**)·**T** = **A**·**T**

=> **infinitely** many attention weights can give the same output!

Effective Attention

In the null space of **T**

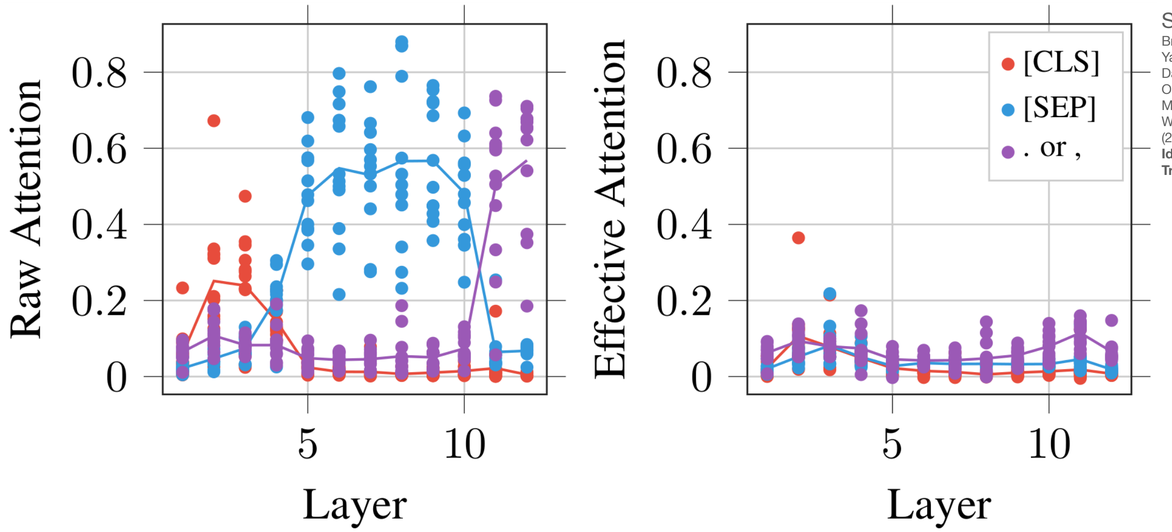
$$A = A^{\parallel} + A^{\perp}$$

Orthogonal to the null space of **T**

Decompose **A** in two components

Call $A^{\perp}$ the *effective* attention

Effective Attention correlates better with the output



Source:
Brunner, Gino & Liu, Yang & Pascual, Damian & Richter, Oliver & Ciaramita, Massimiliano & Wattenhofer, Roger. (2020). **On Identifiability in Transformers.**

Token Identifiability

Can we recover the input tokens from the hidden embeddings?

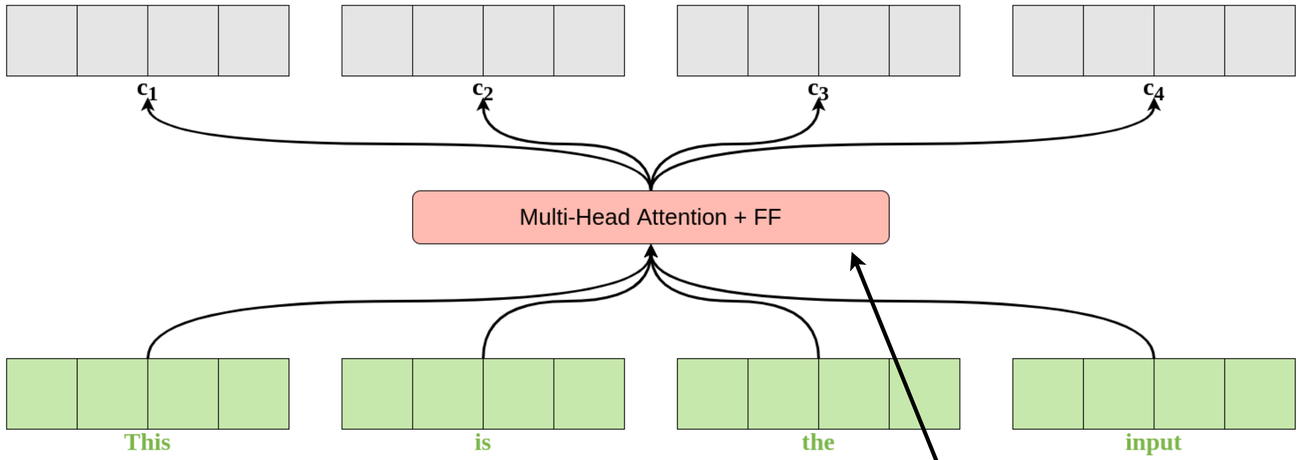
Yes if we find classifier $\hat{g}_l$ such that

$$x_i = NN\left(\hat{g}_l(c_i^l)\right)$$

The input token

Nearest Neighbor

Hidden embedding



Multiple Layers

Token Identifiability

Can we recover the input tokens from the hidden embeddings?

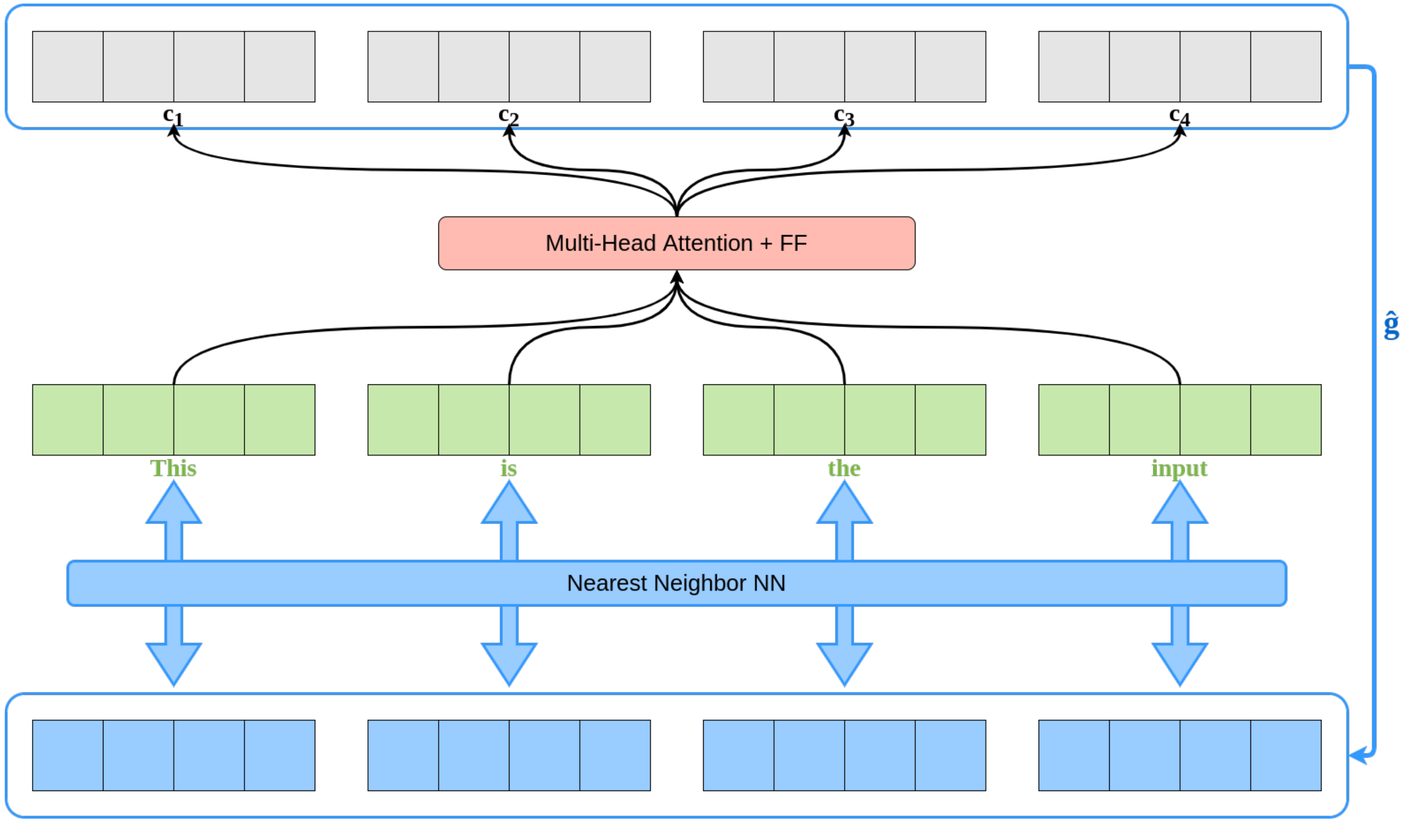
Yes if we find classifier $\hat{g}_l$ such that

$$x_i = NN\left(\hat{g}_l(c_i^l)\right)$$

The input token

Nearest Neighbor

Hidden embedding



Token Identifiability

Can we recover the input tokens from the hidden embeddings?

Yes if we find classifier $\hat{g}_l$ such that

$$x_i = NN\left(\hat{g}_l(c_i^l)\right)$$

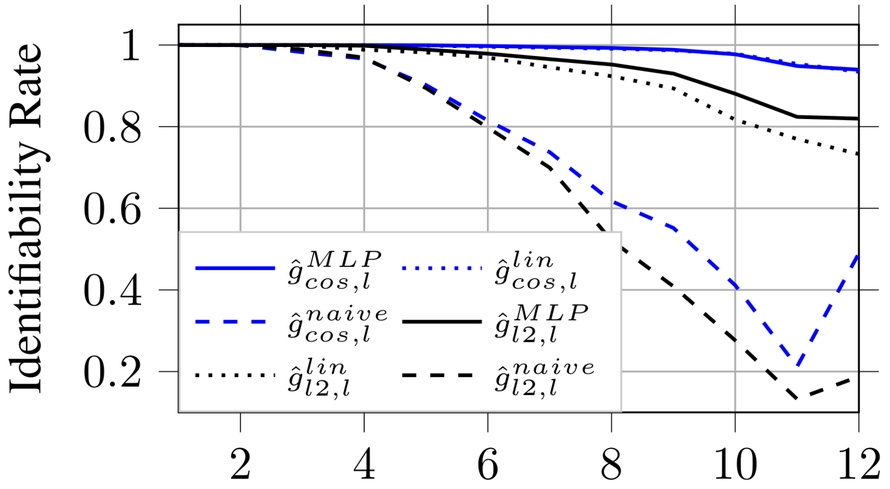
The input token

Nearest Neighbor

Hidden embedding

For NN, cosine distance works best

=> most information in the **angles**



Source:
Brunner, Gino & Liu, Yang & Pascual, Damian & Richter, Oliver & Ciaramita, Massimiliano & Wattenhofer, Roger. (2020). **On Identifiability in Transformers.**

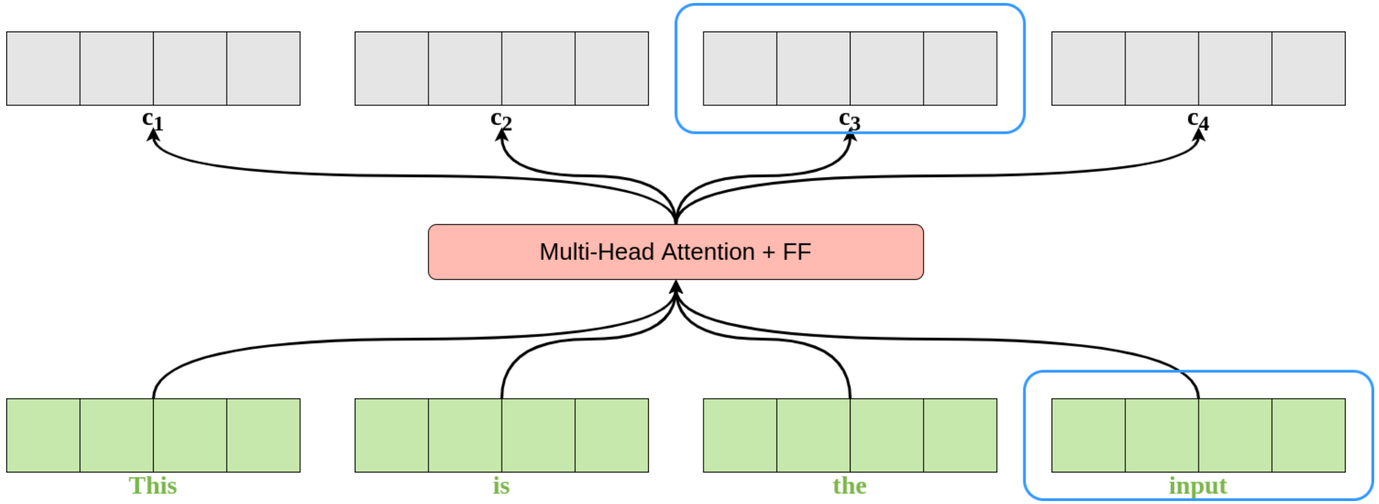
Hidden Token Attribution

How much of the input token is present in a hidden embedding?

=> use **Gradient Attribution**

$$\text{attr}(x_i, c_j^l) = \| \frac{\partial c_j^l}{\partial x_i} \|_2$$

$$\text{contribution}(x_i, c_j^l) = \frac{\text{attr}(x_i, c_j^l)}{\sum_{k=0}^{d_{seq}} \text{attr}(x_k, c_j^l)}$$

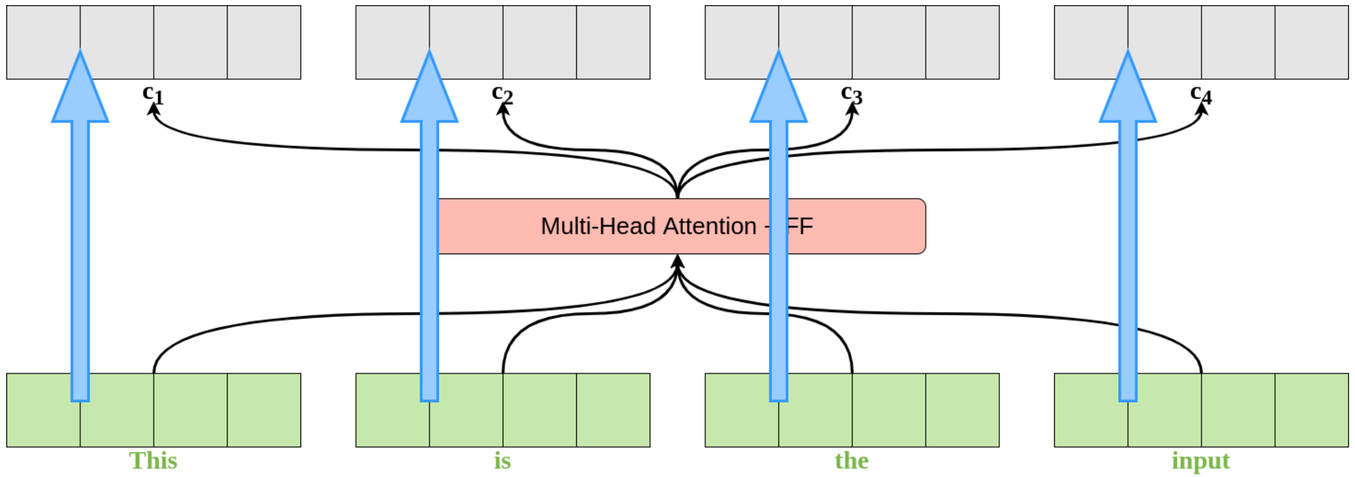


Hidden Token Attribution

How much of the input token is present in a hidden embedding?

Focus on corresponding hidden token:

$$\text{contribution}(x_i, c_i^l)$$

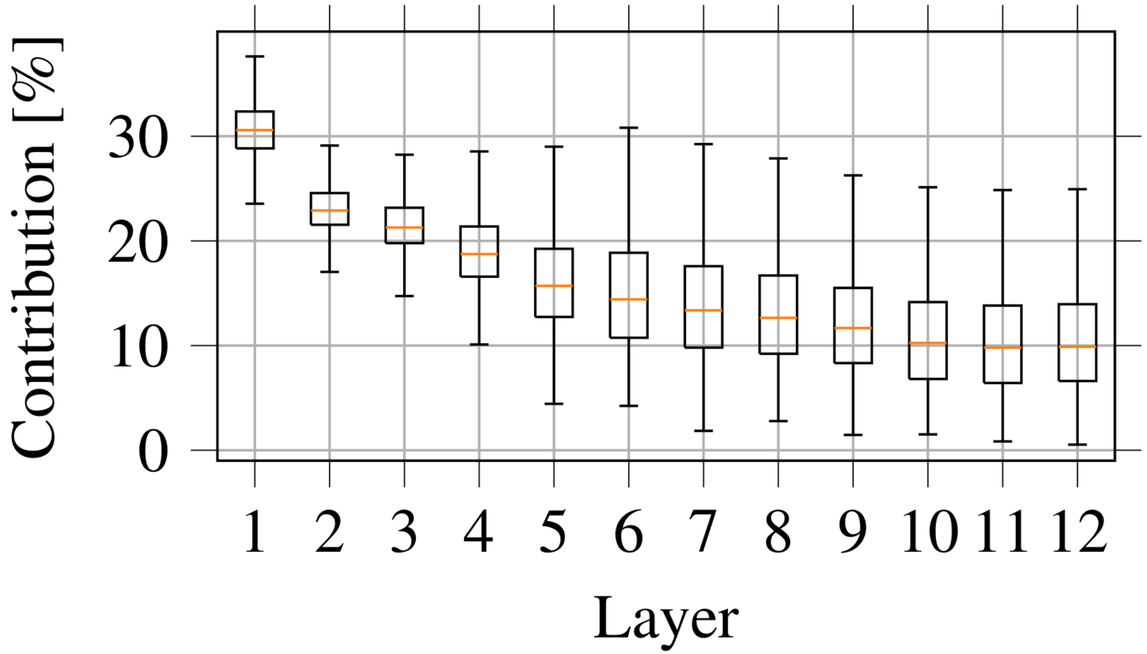


Hidden Token Attribution

Low contribution to corresponding hidden token

Strong mixing of token information

Contrast to high identifiability!



Source:
Brunner, Gino & Liu, Yang & Pascual, Damian & Richter, Oliver & Ciaramita, Massimiliano & Wattenhofer, Roger. (2020). **On Identifiability in Transformers.**

Takeaways

Attention is just a (clever) weighted sum.

Transformer models replace RNNs with layers of attention.

Hidden tokens remain (mostly) identifiable, Attention weights don't.

The papers:

[1]: **Neural Machine Translation by Jointly Learning to Align and Translate** by Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio
[2]: **Attention Is All You Need** by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
[3]: **On Identifiability in Transformers** by Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, Roger Wattenhofer

Thank you for your attention!