Attention Is All You Need

Rem Lohinov

Before the age of ML

- **Rule-based approaches**
- **Statistical approaches**

The Problems with Traditional Architectures

Recurrent Neural Networks (RNNs):

- **The vanishing and exploding gradient problems – It is hard to control gradients during backpropagation**

Convolutional Neural Networks (CNNs)

- **Struggle with capturing long-range dependencies**

What is the attention?

Global attention – takes into account all elements in the input data when calculating the attention weights

Local attention – uses smaller window of input elements

Self-attention – attending to elements within the same sequence (either the input or the output)

Attention weights – how important the element in the input sequence relative to the *current context*

Differences between attentions:

English: The cat sat on the mat.

French: Le chat était assis sur le tapis.

Attention Mechanisms: Query, Key, and Value

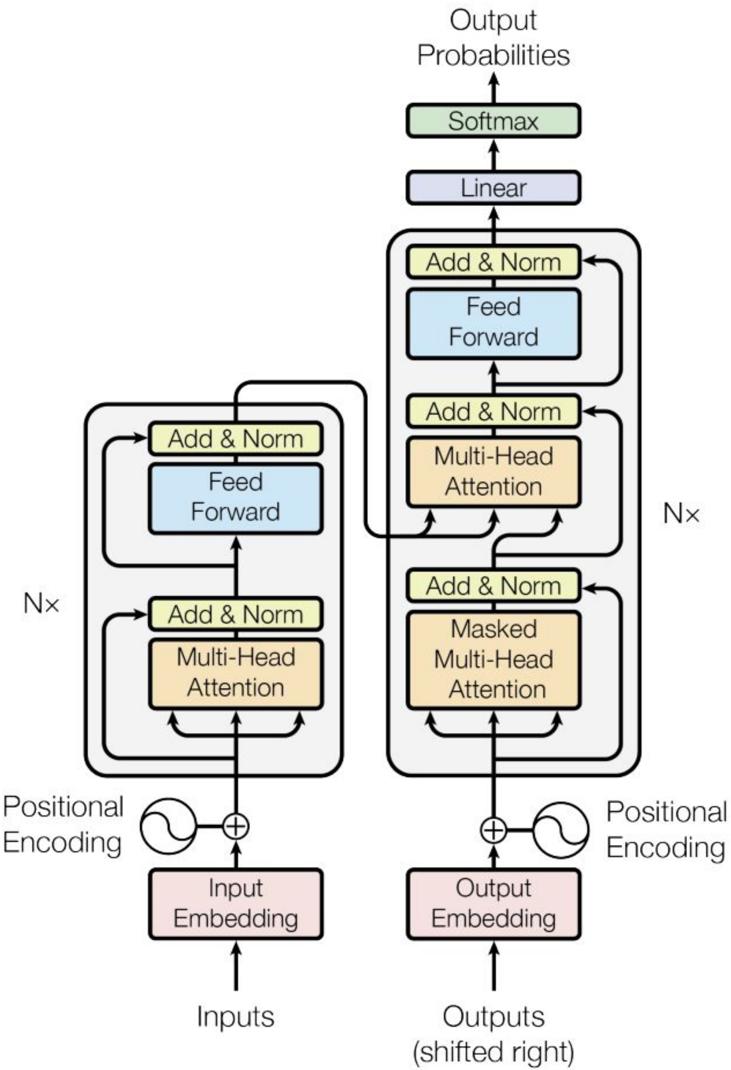
Query — the element we are currently focusing on

Key — other elements in the input data

Value – is associated with each Key, representing the information to be aggregated

The Transformer Architecture

1. Encoder-decoder structure
2. Self-attention layers for capturing relationships
3. Feed-forward networks
4. Layer normalization and residual connections
5. Multi-head attention


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Positional Encoding

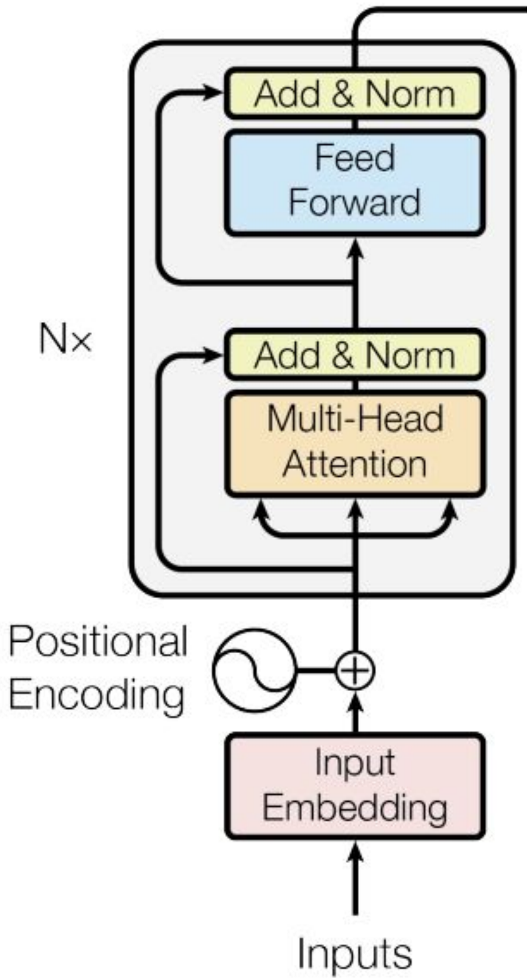
- Self-attention mechanisms does not consider the order of elements in a sequence.

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

- Different formulas for even and odd elements, $2i+1<d_{model}$
- Pos is the position in sequence

Encoder

- Embedding
- Positional Encoding
- Multi-head self-attention
- Feed-forward network
- Residual connections + layer normalization



Multi-Head Attention

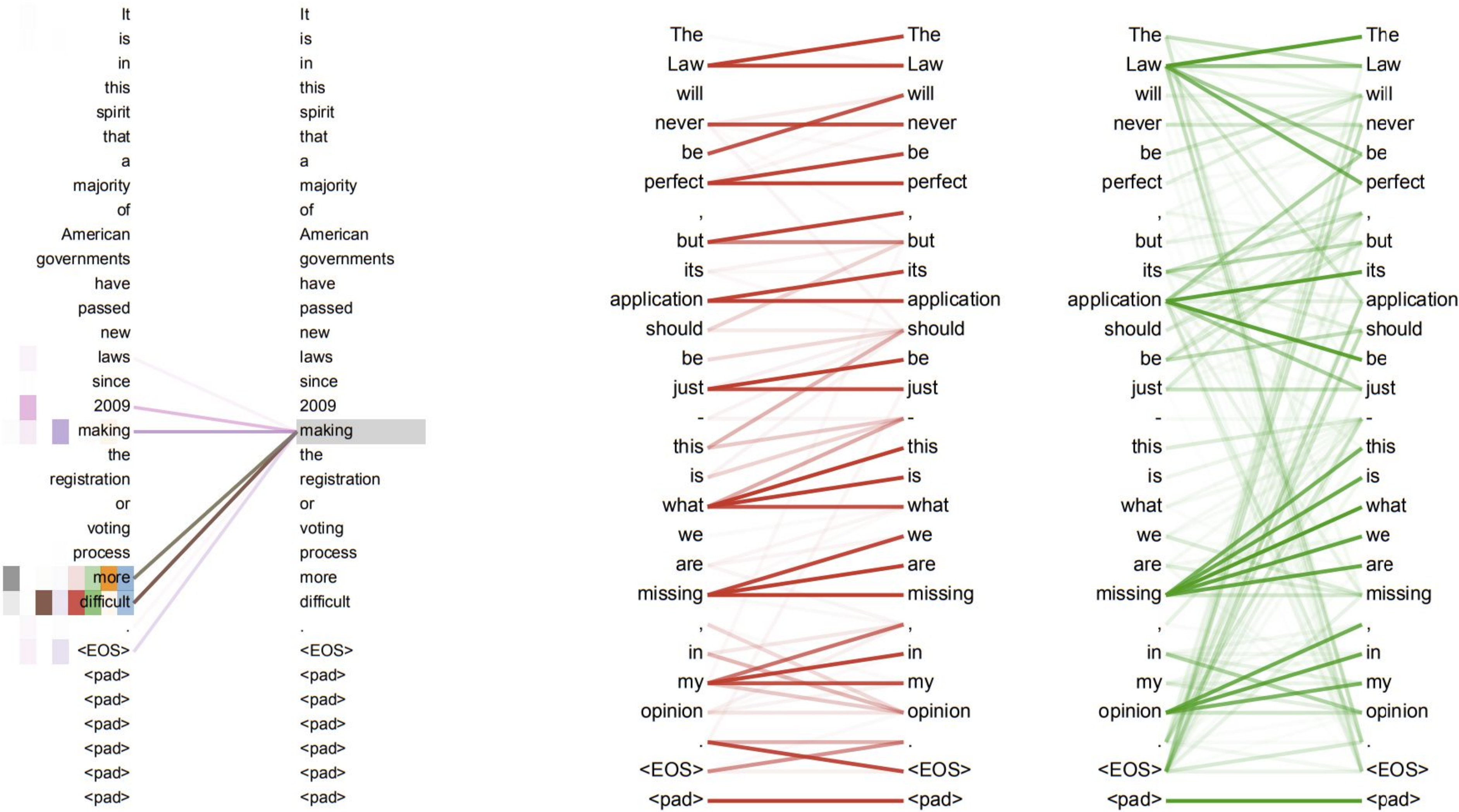
Each head generates the **attention weights** that determine the relevance of each element in the input sequence for the current context.

->

The **attention weights** are then used to compute a **weighted sum of the Value matrices.**

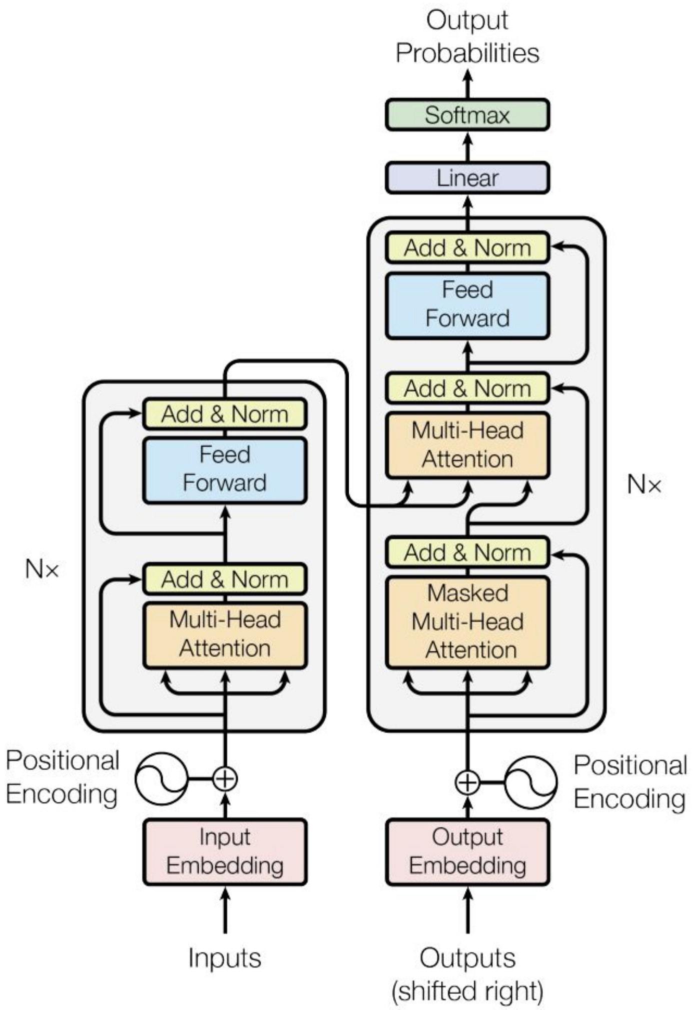
-> *Concatenation and Linear Projection* ->

SINGLE OUTPUT


It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>

Decoder

- Masked multi-head self-attention
- Multi-head attention over encoder output
- Skip/residual connections everywhere – against vanishing gradient problem



Training

- The dimensionality of the word embeddings and positional encodings (d_{model}).
    - Base: 512
    - Big: 1024

- The dimensionality of the feed-forward networks
    - Base: 2048
    - Big: 4096

- Number of attention heads:
    - Base: 8
    - Big: 16

- **Training time:**
    - **Base: 12 hours**
    - **Big: 3.5 days**

Results and Benchmarks

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**	$2.3 \cdot 10^{19}$	

Applications and Use Cases

- Machine Translation
- Text Summarization
- Sentiment Analysis: emotion expressed in a piece of text
- Question Answering :)
- Pretraining and Transfer Learning: BERT, RoBERTa
- Named Entity Recognition (NER): the objective is to identify and classify entities

Limitation, disadvantages

- Memory and Computational Requirements
- Lack of Interpretability: especially self-attention mechanisms
- Susceptibility to Adversarial Attacks: funny :), not funny outputs :(
- Ethical Considerations and Bias

Conclusion

Attention Is All You Need!