

Attention Is All You Need

Author: Vaswani et al.

created by paper2slides

2017-06-12

<https://arxiv.org/abs/1706.03762>

Executive Summary

- Introduces the Transformer architecture, based solely on attention mechanisms.
- Removes the need for recurrence and convolutions in sequence transduction tasks.
- Demonstrates superior results on tasks like machine translation, with more parallelizability and shorter training times.
- Establishes new state-of-the-art BLEU scores on WMT 2014 English-to-German and English-to-French tasks.

Background and Motivation

- RNNs and CNNs are commonly used in sequence transduction models, containing encoder-decoder architectures.
- Limitations:
 - Sequential computation in RNNs precludes parallelization.
 - CNNs struggle to capture long-range dependencies.
- Goal: Develop a simpler architecture that leverages only attention mechanisms for superior performance.

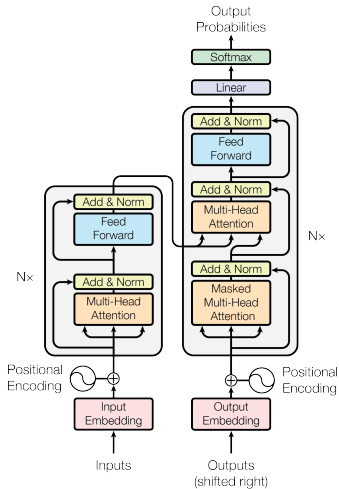


Figure: The Transformer model architecture uses attention mechanisms exclusively.

Transformer Architecture

- The Transformer has an encoder-decoder structure.
- Both encoder and decoder are composed of stacked layers (6 layers each for base model).
- Each layer has two sub-layers:
 - Multi-head self-attention mechanism.
 - Position-wise fully connected feed-forward network.
- Residual connections and layer normalization are applied to each sub-layer.

Self-Attention Mechanism

- Maps a query and key-value pairs to an output.
- **Scaled Dot-Product Attention:**

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

- Each output is a weighted sum of values, scaled by the relationship between queries and keys.
- This allows the model to focus on different parts of the input sequence.

Multi-Head Attention

- Uses multiple heads to attend to information from different representation subspaces.
- For each head:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

- Outputs from heads are concatenated and linearly transformed.
- Provides diversified attention and better learning from multiple representation spaces.

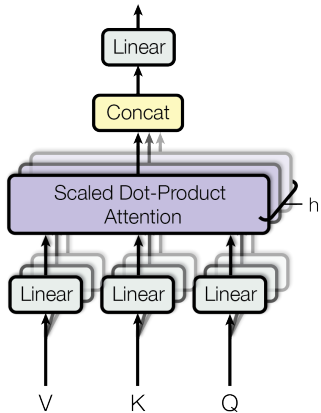


Figure: Visualization of the multi-head attention mechanism.

Positional Encoding

- Transformer adds position information since it lacks recurrence.
- Positional encodings use sinusoidal functions:

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

- Helps the model use sequence order.
- Sinusoidal encodings generalize better for longer sequences.

Training and Regularization

- Datasets: WMT 2014 English-to-German and English-to-French.
- Optimizer: Adam with a custom learning rate schedule.
- Regularization techniques:
 - Residual Dropout $P_{\text{drop}} = 0.1$
 - Label Smoothing $\epsilon_{\text{ls}} = 0.1$
- Training Hardware: 8 NVIDIA P100 GPUs.

Results: Machine Translation

- **WMT 2014 English-to-German Translation Task:**
 - Base model achieves 27.3 BLEU.
 - Big model achieves 28.4 BLEU, setting state-of-the-art.
- **WMT 2014 English-to-French Translation Task:**
 - Base model achieves 38.1 BLEU.
 - Big model achieves 41.8 BLEU, setting state-of-the-art.
- Outperforms previous best models and ensembles at lower training costs.

Performance Comparison

- **Base model:** 100,000 steps, 12 hours on 8 P100 GPUs.
- **Big model:** 300,000 steps, 3.5 days on 8 P100 GPUs.
- Lower computational costs and higher efficiency compared to competitive models.

Input-Input Layer5

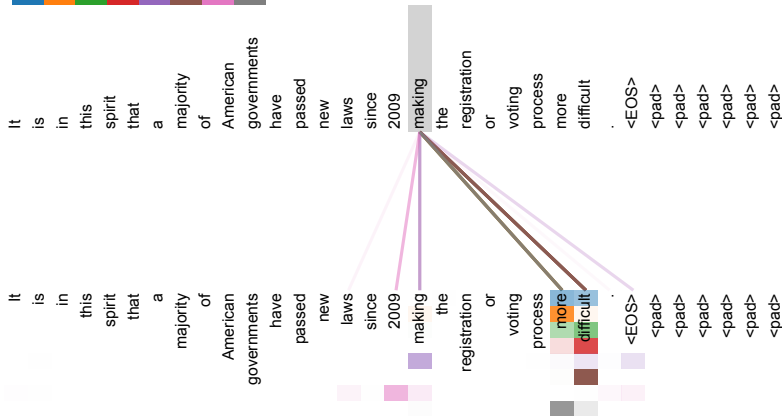


Figure: Attention mechanism following long-distance dependencies in the encoder self-attention.

Conclusion

- Transformer eliminates recurrence and convolutions, relying only on attention.
- Achieves state-of-the-art results in machine translation tasks with superior parallelizability and efficiency.
- Future work:
 - Extend Transformer to other modalities: images, audio, video.
 - Explore local restricted attention for large inputs/outputs.
 - Make generation less sequential.