Attention Is All You Need

Author: Vaswani et al.

created by paper2slides

2017-06-12

https://arxiv.org/abs/1706.03762

Executive Summary

- This paper introduces the Transformer model, a novel architecture that relies entirely on self-attention mechanisms.
- The Transformer outperforms previous state-of-the-art sequence transduction models in machine translation tasks.
- Benefits include superior quality, higher parallelization, and reduced training time.
- Achieved a new state-of-the-art BLEU score of 28.4 on the WMT 2014 English-to-German and 41.8 on the English-to-French translation tasks.

Motivation and Background

- Traditional sequence models use recurrent neural networks (RNNs) or convolutional neural networks (CNNs).
- RNNs have limited parallelization due to their sequential nature, hindering efficient training on long sequences.
- CNNs improve parallelization but still require multiple layers to capture long-range dependencies.
- Attention mechanisms have emerged as efficient for modeling dependencies regardless of their distance in the sequences.

Model Architecture Overview

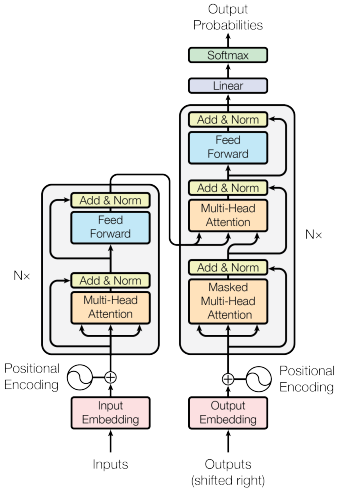

Figure: The Transformer - model architecture.

Model Architecture: Encoder and Decoder

- **Encoder:**
  - Comprises 6 identical layers.
  - Each layer has two sub-layers: multi-head self-attention and position-wise fully connected feed-forward network.
  - Residual connections and layer normalization are applied after each sub-layer.
- **Decoder:**
  - Comprises 6 identical layers, similar to the encoder.
  - Includes a third sub-layer performing multi-head attention over the encoder's output.
  - Incorporates masking to prevent attending to future positions, ensuring autoregressive property.

Attention Mechanism: Scaled Dot-Product Attention

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

- Queries (Q), keys (K), and values (V) are matrices.
- Compute dot products of the query with all keys, scale by $\sqrt{d_k}$, and apply a softmax function to get the weights.
- Compute the weighted sum of the values.

Attention Mechanism: Multi-Head Attention

- Instead of a single attention function, apply multiple attention mechanisms, called heads.
- Each head operates on linearly projected versions of the queries, keys, and values.
- Outputs are concatenated and once more projected to yield the final values.

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O \\ \text{head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

Positional Encoding

- Self-attention is agnostic to the positional order of tokens.
- Positional encodings are added to the input embeddings to incorporate sequence order.
- Use sine and cosine functions of different frequencies:

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

- Allows the model to learn and extrapolate from positional information.

Training and Regularization

Training Data and Hardware:

- Data: WMT 2014 English-German (4.5M sentence pairs), English-French (36M sentence pairs).
- Trained on 8 NVIDIA P100 GPUs.

Optimizers and Regularization:

- Optimizer: Adam with specific learning rate schedule.
- Regularization: Residual dropout, label smoothing ($\epsilon_{ls} = 0.1$).

Performance on Machine Translation

Table: Translation Results on WMT 2014

Model	EN-DE BLEU	EN-FR BLEU
ByteNet	23.75	-
GNMT + RL	24.6	39.92
ConvS2S	25.16	40.46
MoE	26.03	40.56
Transformer (Base)	27.3	38.1
Transformer (Big)	**28.4**	**41.8**

Ablation Studies

- **Number of heads:** 8 heads provide optimal performance.
- **Model dimensions:** Increasing dimensions d_k and d_v improves performance.
- **Dropout rates:** Adjusting dropout rates from 0.0 to 0.2 impacts BLEU scores.
- **Positional encoding:** Sinusoidal vs. learned encodings perform similarly.

English Constituency Parsing

- Evaluated on WSJ section of Penn Treebank.
- Baseline model: F1 score of 91.3 on WSJ.
- Semi-supervised model: F1 score of 92.7.
- Comparable to state-of-the-art results in constituency parsing.

Conclusion and Future Directions

- The Transformer establishes new state-of-the-art in machine translation.
- Advantages include better parallelization and reduced training time.
- Outperforms traditional sequence models like RNNs and CNNs.
- Future work: extending the Transformer to other modalities (e.g., images, audio), exploring local restricted attention.