Attention Is All You Need

Authors

? Ashish Vaswani ? (Google Brain)

? Noam Shazeer ? (Google Brain)

? Niki Parmar ? (Google Research)

? Jakob Uszkoreit (Google Research)

? Llion Jones ? (Google Research)

? Aidan N. Gomez ? ? (University of Toronto)

? ?ukasz Kaiser ? (Google Brain)

? Illia Polosukhin ? ?

Abstract

The paper introduces the Transformer, a novel network architecture for sequence transduction tasks that relies entirely on attention mechanisms, eliminating the need for recurrence and convolutions. This model demonstrates superior performance in machine translation tasks, achieving state-of-the-art BLEU scores on the WMT 2014 English-to-German and English-to-French translation tasks. The Transformer is more parallelizable and requires significantly less training time compared to existing models.

Introduction

Traditional sequence modeling and transduction models utilize recurrent neural networks (RNNs) with attention mechanisms. However, these models face challenges in parallelization due to their sequential nature. The Transformer model addresses this by using self-attention mechanisms to compute dependencies, allowing for greater parallelization.

Model Architecture

The Transformer consists of an encoder-decoder structure:

- Encoder: Composed of 6 identical layers, each with a multi-head self-attention mechanism and a p

- Decoder: Similar to the encoder but includes an additional sub-layer for multi-head attention over

Attention Mechanisms

- Scaled Dot-Product Attention: Computes attention as a scaled dot product of queries, keys, and v

- Multi-Head Attention: Allows the model to attend to information from different subspaces using m

Position-wise Feed-Forward Networks

Each layer in the encoder and decoder includes a feed-forward network applied identically to each position.

Embeddings and Positional Encoding

- Embeddings: Convert input and output tokens into vectors.

- Positional Encoding: Injects information about the sequence order using sine and cosine functior

Training

The Transformer was trained on the WMT 2014 English-German and English-French datasets using the Adam optimizer with specific regularization techniques like dropout and label smoothing.

Results

The Transformer outperforms previous models in translation tasks, achieving higher BLEU scores with reduced training costs. The model's architecture allows for efficient learning of long-range dependencies due to its self-attention mechanism.

Conclusion

The Transformer model provides a robust alternative to traditional sequence transduction models by leveraging attention mechanisms exclusively. The authors plan to extend the model to other tasks and input/output modalities beyond text.

References

? The paper includes references to foundational works on neural networks, attention mechanisms, and machine translation, highlighting contributions from various researchers in the field.

The code for the Transformer model is available at [GitHub](https://github.com/tensorflow/tensor2tensor).