Transformer Architecture

Summer

Introduction

The Transformer architecture is a breakthrough in natural language processing (NLP), introduced by Vaswani et al. in 2017 in the paper *Attention Is All You Need*. This document summarizes key concepts and advantages of the Transformer, including its architecture and applications.

Historical Context

- **Perceptron**: Introduced in 1957, capable of simple linear classification.
- **Feed-Forward Networks (FFNs)**: Limited by fixed input sizes and lack of temporal context.
- **Recurrent Neural Networks (RNNs)**: Introduced to handle sequences with arbitrary lengths but faced vanishing and exploding gradient issues.
- **Long Short-Term Memory (LSTM)**: Designed to mitigate RNN issues, though computationally intensive.

Key Innovations of Transformers

- **Self-Attention Mechanism**: Allows the model to relate different positions of a sequence to compute its representation.
- **Parallel Processing**: Unlike RNNs, Transformers process tokens simultaneously, enabling faster computation.
- **Positional Encodings**: Adds positional information to token embeddings since Transformers process tokens without intrinsic order.
- **Residual Connections and Layer Normalization**: Enhance gradient flow and stabilize training.

Self-Attention Mechanism

- Queries (Q), Keys (K), and Values (V) are derived from input embeddings.
- **Attention Scores**: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$.

- Multi-head attention allows the model to focus on different parts of the input simultaneously.

Transformer Architecture

- **Encoder**: Processes input sequence and outputs contextual representations.
- **Decoder**: Generates output sequence based on encoder representations and previous tokens.
- **Variants**:
  - Encoder-only models (e.g., BERT) for tasks like classification.
  - Decoder-only models (e.g., GPT) for text generation.
  - Encoder-Decoder models (e.g., T5) for sequence-to-sequence tasks.

Advantages of Transformers

- **Scalability**: Suitable for large-scale data and transfer learning.
- **State-of-the-Art Performance**: Excels across NLP, computer vision, and multimodal tasks.
- **Efficient Handling of Long-Range Dependencies**: Self-attention effectively captures context over long sequences.

Applications

- Machine Translation (e.g., Google Translate).
- Text Summarization and Question Answering.
- Pre-trained Language Models (e.g., BERT, GPT).

References

- Vaswani, A. et al. (2017). *Attention Is All You Need.* `https://arxiv.org/pdf/1706.03762`.
- Alammar, J. *The Illustrated Transformer.* `https://jalammar.github.io/illustrated-transf`
- He, K. et al. (2016). *Deep Residual Learning for Image Recognition.*