Summary of "Attention is All You Need"

In this work, the authors propose the Transformer, a model architecture that completely eliminates recurrence and relies entirely on attention mechanisms to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and achieves state-of-the-art performance in machine translation after being trained for as little as 12 hours on eight P100 GPUs.

Performance Highlights

- The model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving upon the existing best results, including ensembles, by over 2 BLEU.

- On the WMT 2014 English-to-French task, the Transformer establishes a new single-model state-of-the-art BLEU score of 41.8, training for only 3.5 days on eight GPUs, which is a small fraction of the cost compared to previous best models.

Motivation

The inherently sequential nature of recurrent models precludes parallelization within training examples. This limitation becomes critical at longer sequence lengths, where memory constraints restrict batching across examples.

Model Architecture

The Transformer follows the standard encoder-decoder architecture commonly used in neural sequence transduction models. The encoder maps an input sequence of symbol representations to a continuous sequence, while the decoder generates an output sequence auto-regressively, consuming previously generated symbols as input.

Encoder

- Composed of a stack of **N = 6 identical layers**.

- Each layer includes:

    - A **multi-head self-attention mechanism**.
    - A **position-wise fully connected feed-forward network**.
- Each sub-layer is followed by a residual connection and layer normalization.
- All sub-layers and embeddings produce outputs of dimension **512**.

Decoder

- Also composed of a stack of **N = 6 identical layers**.
- In addition to the two sub-layers in each encoder layer, the decoder includes a third sub-layer that performs multi-head attention over the encoder's output.
- Residual connections and layer normalization are applied similarly.
- A masking mechanism in the self-attention sub-layer prevents positions from attending to subsequent positions, ensuring that predictions at position *i* only depend on outputs at positions less than *i*.

Attention Mechanisms

Scaled Dot-Product Attention

- Maps a query and a set of key-value pairs to an output, computed as a weighted sum of the values.
- Weights are derived using a compatibility function between the query and each key.

Multi-Head Attention

- Instead of a single attention function, multiple attention heads are used.
- The queries, keys, and values are linearly projected *h* times into different subspaces.
- Attention is computed in parallel across these heads, then concatenated and projected again.

- This enables the model to jointly attend to information from different representation subspaces.

Other Components

Position-wise Feed-Forward Networks

- Each layer contains a fully connected feed-forward network.
- While transformations are the same across positions, the parameters differ across layers.

Embeddings and Softmax

- Learned embeddings convert input and output tokens into **512-dimensional vectors**.
- A learned linear transformation and softmax function produce predicted next-token probabilities.

Positional Encoding

- Since the model uses no recurrence or convolution, **positional encodings** are added to the input embeddings to convey sequence order.
- These encodings have the same dimensionality as the embeddings.

Why Self-Attention?

- **Computational complexity per layer** is reduced.
- **Parallelization** is significantly improved by reducing the number of sequential operations.
- **Path lengths** between long-range dependencies are shorter, making it easier to learn such relationships.

Training Details

- **English-German**: WMT 2014 dataset with ~4.5 million sentence pairs.
- **English-French**: WMT 2014 dataset with ~36 million sentence pairs.
- **Vocabulary**: Tokenized into 32,000 word pieces.
- **Optimizer**: Adam.
- **Regularization**: Includes residual dropout, label smoothing, and others.

Generalization to Other Tasks

To test the generality of the Transformer, the authors also applied it to **English constituency parsing**, which poses additional challenges due to strict structural constraints and longer outputs than inputs.

Conclusion

The Transformer achieves new state-of-the-art results on major translation tasks and does so with significantly **faster training times** than recurrent or convolutional architectures. Its parallelizable design and effectiveness in capturing long-range dependencies make it a foundational architecture in modern deep learning.