AI VIET NAM – COURSE 2022

Attention Is All You Need

November 30, 2022

Date of publication:	12/06/2017
Authors:	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
Sources:	Attention Is All You Need
Data sources:	Transformer
Keywords:	Attention, Transformer, Seq2Seq
Summary by:	Quoc Viet

1. **Purpose/Output:**
   On academic English to German and English to French translation benchmarks, the Transformer [3] surpasses both recurrent and convolutional models, as shown in Figure 1. Aside from improved translation quality, the Transformer requires less computation to train and is a far better fit for modern machine learning hardware, allowing training to be completed in an order of magnitude less time.

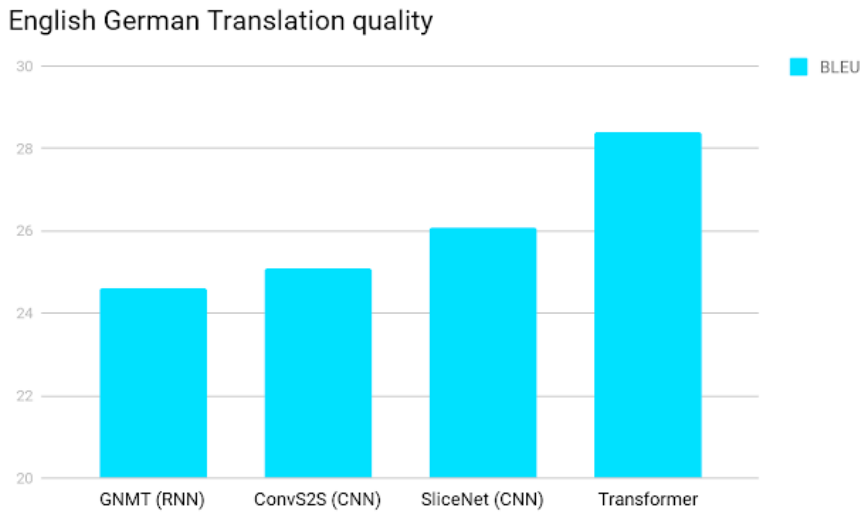

Figure 1: BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark. (*Source: Google AI Blog*)

2. **Contributions:**

The Transformer is the first model to calculate representations of its input and output using only self-attention rather than sequence aligned RNNs or convolution. The Transformer encoder reads the complete sequence of words at once, as contrast to directional models, which read the text input sequentially (left-to-right or right-to-left). As a result, it is deemed bidirectional. This feature enables the model to learn the context of a word based on its surrounds (left and right of the word).

3. **Methodology:**

- **Attention**

  In general, the **Attention mechanism** [1] examines an input sequence and determines which elements of the sequence are relevant at each stage. In other words, Attention allows the model to focus on the relevant parts of the input sequence as needed. In comparision with the classic sequence-to-sequence model [2], in the Attention model, the encoder passes a lot more data to the decoder instead of passing the last hidden state of the encoding stage, the encoder passes all the hidden states to the decoder.

- **Scaled Dot-Product Attention**

  Scaled dot-product attention is an attention mechanism that has the structure as shown in Figure (2), where the dot products are scaled down by $\sqrt{d_k}$. The Scaled dot-product attention can be described by the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

  where:
  - Q is a matrix that contains the query (vector representation of one word in the sequence)
  - K are all the keys (vector representations of all the words in the sequence)
  - V are the values (vector representations of all the words in the sequence)

  The weights $a = softmax(\frac{QK^T}{\sqrt{d_k}})$ indicates the extent to which each word of the sequence (denoted as Q) is influenced by all other words in the sequence (denoted as K).

- **Multi-Head Attention**

  Multi-head Attention is an attention mechanism module that runs an attention mechanism numerous times in parallel. Following that, the independent attention outputs are concatenated and linearly transformed into the expected dimension. Multiple attention heads, intuitively, allow for different attention to different points of the sequence, where attention heads can be thought of as distinct operations.

  The Multi-Head Attention can mathematically be expressed as:

$$\begin{aligned} &\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \ldots, \text{head}_h]\, \mathbf{W}_0, \\ &\text{where head }_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \end{aligned}$$

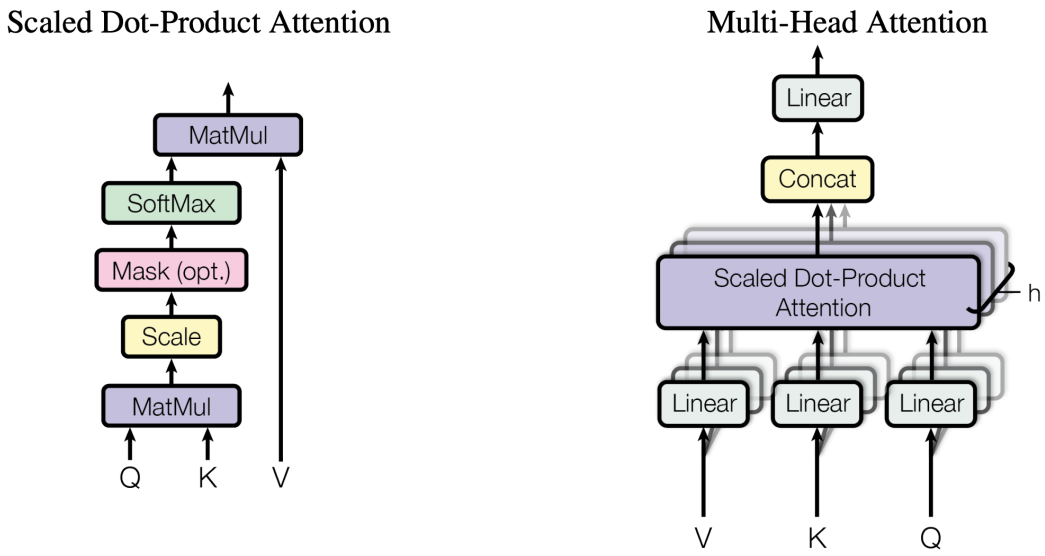

Figure 2: Scaled Dot-Product Attention (left). Multihead Attention consists of several attention layers running in parallel.

- **The Transformer Architecture**

  As shown in Figure (3), the Transformer architecture follows an encoder-decoder structure but does not rely on recurrence and convolutions in order to generate an output. Both the Encoder and the Decoder consist of modules stacking on top of each other. The encoder takes the input sequence and maps it into a higher dimensional space (n-dimnesional vector). The vector is the input of the decoder which is then turned into an output sequence.

  A decoder then constructs the output sentence word by word while consulting the encoder's representation. For each word, the Transformer begins by generating initial representations, or embeddings. Then, employing self-attention, it accumulates information from all other words, forming a new representation for each word that is informed by the full context. This phase is then repeated in parallel for all words, resulting in a series of new representations.

  The Transformer's key attribute is that each word travels through its own path in the encoder. In the self-attention layer, there exist dependencies between these paths. However, because the feed-forward layer lacks these kind of dependencies, various paths can be executed in parallel while flowing through the feed-forward layer.

- **Positional Embedding**

  In order to add position information to the model, the authors proposed the positional embedding method, which outputs a d-dimensional vector that contains information about a specific position in a sentence using **Sinusoidal function**. Let t be the desired position in an input sentence, $\overrightarrow{p_t} \in \mathbb{R}^d$ be its corresponding encoding, and d be the encoding dimension (where $d \equiv_2 0$) Then $f : \mathbb{N} \to \mathbb{R}^d$ will be the function that produces the output vector $\overrightarrow{p_t}$ and it is defined as follows:

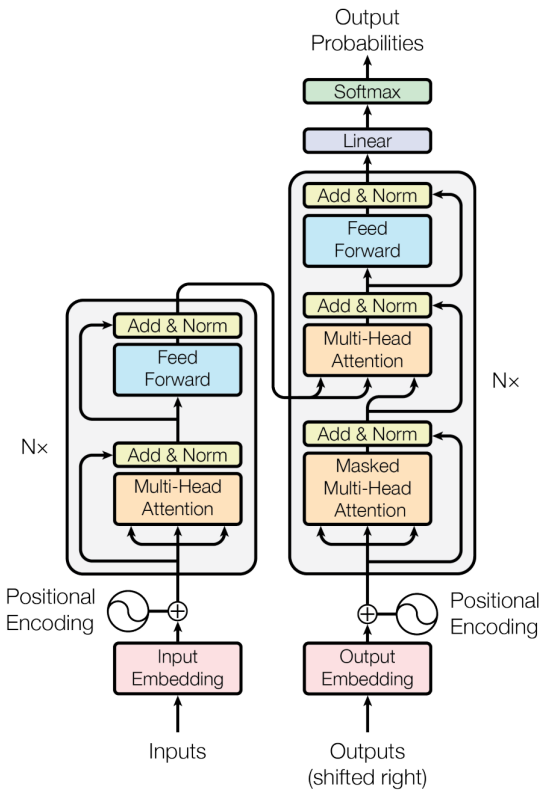

Figure 3: The Transformer Architecture

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k \cdot t), & \text{if } i = 2k \\ \cos(\omega_k \cdot t), & \text{if } i = 2k+1 \end{cases}$$

where

$$\omega_k = \frac{1}{10000^{2k/d}}$$

As it can be derived from the function definition, the frequencies are decreasing along the vector dimension. Thus it forms a geometric progression from 2π to $10000 \cdot 2\pi$ on the wavelengths.

4. **Results:**

- The transformer model (Transformer (big) in Figure 4) surpasses the best previously reported models (including ensembles) by more than 2.0 BLEU on the WMT 2014 English-to-German translation problem, creating a new state-of-the-art BLEU score of 28.4
- On the WMT 2014 English-to-French translation challenge, the Transformer model earns a BLEU score of 41.0, outperforming all previously released single models, while costing less than a quarter of the training time of the previous state-of-the-art model.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Figure 4: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

References

[1] Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[2] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks, 2014.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.