UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Paper Summary and Report

Attention Is All You Need

Vaswani et al., 2017

Daolong Chen

March 22, 2024

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

1 Paper Details

Title: Attention Is All You Need
Authors: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin
Publication details: Published in the 31st Conference on Neural Information Processing System (NIPS 2017)

2 Overview

This paper introduced a novel deep neural network, the Transformer, to address the limitations of existing sequence-to-sequence models for natural language processing (NLP) tasks - particularly neural machine translation [1]. The authors proposed a novel approach based solely on multi-head attention, without relying on traditional recurrent layers such as in recurrent neural networks (RNNs) [2], Long-Short Term Memory (LSTMs) [3] [4] and Gated Recurrent Units (GRUs) [5] or convolutional layers [6].

One of the key challenges with previous models was the sequential nature of recurrent layers, leading to increased computational complexity and difficulty in learning long-range dependencies [1] [3] [7]. The Transformer model overcame this through the introduction of a self-attention mechanism, allowing for parallelization of computation and avoiding the proportional increase in operations with sequence length [1].

The self-attention mechanism allows the model to focus on different parts of the input sequence during processing, capturing dependencies between the elements or tokens regardless of distance. This was achieved through the introduction of multiple attention heads, which enables the model to attend to different representation subspaces jointly and collectively contribute to the final representation of each element [1]. By eliminating the need for recurrence or convolution, the Transformer also generally has lower computational complexity. Therefore, this was a breakthrough in the field of NLP, as it allowed for more efficient training and better handling of long-range dependencies compared to traditional approaches.

3 Transformer

Similarly to the current state-of-the-art approaches to sequence-to-sequence models [5] [4], the Transformer is an encoder and decoder model (Figure 1). However, the difference is that the authors instead proposed a stack of N=6 multi-head attention and point-wise fully connected sub-layers (Figure 1). Additionally, to facilitate residual connections around each sub-layer, the outputs of each sub-layer have dimension 512 (d_{model}).

3.1 Embedding

As with all sequence-to-sequence models in NLP, the tokens need to be mapped into an embedding space. Instead of using pre-trained embeddings such as Word2Vec [8] and GloVe [9], the authors proposed to use learned embeddings that would be specific to the training dataset they used. This was incorporated into the Transformer model through embedding layers, which learn the mapping from token indices to the embedding dimension during training.

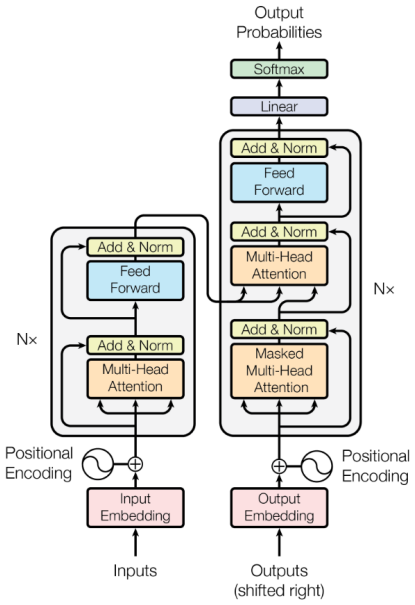

Figure 1: The Transformer architecture proposed by the authors [1]

3.2 Position Encoding

The attention mechanism proposed is order-agnostic. Therefore, the authors proposed the use of alternating sine and cosine functions of increasing wavelength (from 2π to 10000π) to map the position to a float for each element i along the length of the vector encoding (d_{model}). The position encoding is then summed with the token embedding to inject position information into the data.

Specifically the position encoding is given by the equations [1]:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

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

where $0 \leq i \leq \frac{d_{model}}{2}$ and pos is token position

3.3 Attention Mechanism

Several attention mechanisms had been proposed previously - most notably by [7] and [10]. In this paper, the authors proposed the use of dot-product attention, similar to that introduced by [10], with added scaling for stabilisation of the learning process (Figure 2 (a)). This involves 3 matrices, the query (Q), the key (K) and the values (V) which each are a collection of embedded tokens of size (d_{model}). The queries are compared with the keys to calculate attention scores. The scores are then scaled according to the square root of the K dimension and the softmax-normalised values are subsequently used to weight the values. Therefore, the resultant token vectors are a weighted average of values, which enables the model to include contextual information contained in V.

The proposed attention mechanism is given by:

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

where Q, K, and V are the input sequences into the attention sub-layer and are known as the queries,

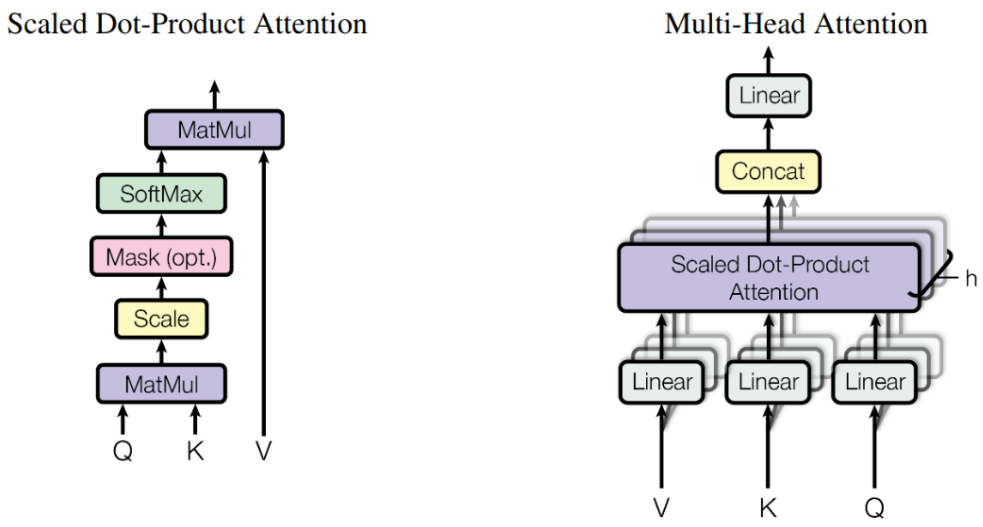

Figure 2: a) Diagram of the scaled dot-product attention mechanism b) Extension to multi-head attention [1]

Figure 3: An example of a self-attention score matrix

keys and values respectively. d_k is the dimension of the keys (d_{model}).

3.3.1 Types of Attention

The authors employed attention in 3 different ways. Self-attention is utilised in the encoder, where $Q = K = V$. For example, if the input is the embedded sequence of "I love dogs", then each token "I", "love" and "dogs" will attend to every token in that sequence and learn contextual information. An example self-attention score matrix (QK^T) is shown Figure 3.

In the decoder, the authors utilised masking of self-attention scores to prevent the Transformer from attending to positions that would be unavailable during inference (Figure 4).

In a similar way to recurrent models with attention [7], encoder-decoder attention is incorporated into

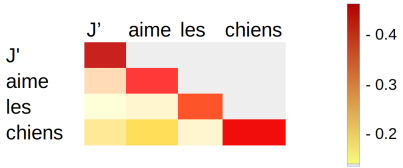

Figure 4: An example of a masked self-attention score matrix. Masked attention scores are shown in grey

Figure 5: An example of an encoder-decoder attention score matrix

the decoder. This enables the decoder to attend to all positions in the encoded sequence (Figure 5). Using contextual information from the output of the encoder the decoder predicts the next token in the translated sequence. In this example, the final translated sequence would be "J'aime les chiens". Note that a beginning-of-sentence token, <BOS>, would be added to the input sequence and an end-of-sentence token <EOS> would be output from the decoder. These were omitted for ease of understanding.

3.3.2 Multi-head Attention

Additionally, the authors hypothesised that parallel computation of attention could enable the model to jointly attend differently to different positions and hence give the model greater power to encode meaningful information from the input.

Therefore the authors proposed parallel computation of attention using h = 8 "heads". For each head, Q, K, and V are projected through 3 linear layers (W^Q, W^K, W^V) to obtain 3 matrices, each with a dimension of 64. After parallel computation of attention on the projected matrices, the output from each head is concatenated together and projected through a final linear layer (W^O) (Figure 2 (b)).

Therefore, the attention mechanism in each head i is given by [7]:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

and the output of the multi-head attention is given by [7]:

$$MultiHead(Q, K, V) = Concatenation(head_1, head_2, ..., head_h)W^O$$

3.3.3 Feed-forward Layer

A feed-forward network was proposed as a sub-layer in both encoder and decoder layers. The network simply consists of two linear layers with a ReLU activation in between. The authors proposed the dimension of the first linear layer (d_{ff}) to increase from 512 to 2048.

3.3.4 Output

To generate the output of the Transformer, similar to typical sequence-to-sequence NLP models [4], the authors proposed the use of a linear layer followed by softmax normalisation over the vocabulary to obtain the predicted next token probabilities.

4 Results

The authors evaluated the performance of the Transformer on translation from English to French and English to German using the WMT 2014 dataset.

4.1 BLEU score

The primary metric used by the authors was the BLEU score, given by the equation below [11]:

$$BLEU = min(1, exp(1 - \frac{reference - length}{output - length}))(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

The second term is a measure of the precision of the output sequence with respect to the ground truth. It is the product of the precision of n-grams (a group of n successive items/tokens) from unigrams to 4-grams. The first term is a"brevity penalty" introduced due to the absence of a recall term. This penalises the output sequence exponentially for being shorter than the ground truth.

4.2 Performance

Performance was evaluated on a 'base' and 'large' Transformer. The large Transformer has the parameters $d_{model} = 1024, d_{ff} = 4096$ and $h = 16$ compared to $d_{model} = 512, d_{ff} = 2048$ and $h = 8$ for the base model.

The authors demonstrated that the large Transformer achieved a state-of-the-art BLEU score of 28.4 on the English-to-German, outperforming both state-of-the-art single models and ensemble models by more than 2.0 BLEU. For the English-to-French translation task, the big Transformer outperformed state-of-the-art single models whilst achieving a competitive BLEU score of 41.0 compared to ensemble models.

Importantly, the author's estimation of the number of floating operations highlighted that a key advantage of the Transformer is that it can achieve state-of-the-art results at a fraction of the training cost.

5 Critical Analysis and Discussion

In terms of the methodology, the motivation for the introduction of the Transformer was clearly explained. The authors highlighted the complexity per layer, the number of sequential operations and the maximum path length of the proposed self-attention mechanism and compared them to recurrent and convolutional layers to emphasise the key advantages of self-attention.

Whilst theoretically, as reasoned by the authors, multiple heads enable the Transformer to learn from different representation subspaces, only empirical figures were presented in the appendix of the paper. This was also pointed out during the discussion. Perhaps a more structured empirical investigation could have been carried out by the authors into the behaviour of the different attention heads.

The authors also mentioned the reasons for the use of sinusoidal position encoding. Intuitively, this can allow the model to extrapolate to unseen sequence lengths. Although, the authors did not explicitly investigate whether this is true, which would have been very insightful. The hypothesis that such positional encoding would enable the model to also attend to relative positions perhaps could have been expanded upon by the authors.

In terms of evaluation of the performance, the authors documented well the datasets used, the resources required, the regularisation techniques employed, and optimiser parameters which would enable the reproduction of their results. Evaluation of Transformer variants also gave important insight into the importance of the many different hyperparameters of the model.

Points that perhaps could have been included were the choice of learning rate for the Adam optimiser and the loss function used. This was later found to be cross-entropy loss from the Tensor2Tensor code. Furthermore, the paper could have perhaps expanded on the difference between how the Transformer

works during training compared to inference.

In terms of structure, the authors introduced key sub-layers and layers of the architecture in sections. This made the paper easy to read and follow. In particular, the introduction of single-head attention before multi-head attention enabled the reader to better understand the key concept behind the Transformer - multi-head attention.

Similarly to the paper, the presentation was split into sections for each sub-layer of the Transformer. Additionally, the Transformer architecture was included on all slides with the sub-layer being introduced highlighted. This was done to aid the orientation of the audience. The feedback from the cohort was that in combination with external figures, such as an example attention score heatmap, made the presentation flow well and easy to follow.

One of the main points of the discussion centred around the attention mechanism. The use of the terms query, key, and value seemed to cause a bit of confusion. This was resolved by drawing an example of the attention mechanism with the sentence "I love dogs". In the future, figures that demonstrate a relevant example will be included in presentations to aid audience understanding.

There were also questions during the discussion regarding the output of the Transformer during inference time, as mentioned before, the paper did not include much detail regarding the Transformer during inference. During inference, the Transformer is auto-regressive - it outputs one token at each time step and takes in all the tokens output as input into the decoder stack. This was demonstrated again by drawing a relevant example that showed the decoder input and output at each time step.

The cohort was also curious as to whether the Transformer has since superseded recurrent neural networks. It can depend on the task, however generally, Transformer-based architectures outperform recurrent neural networks particularly in NLP tasks [12] [13].

Some constructive cohort feedback included that the presentation focused heavily on the architecture of the Transformer, and could have included references to healthcare applications of Transformer-based architectures. The text on the presentation slides could have also been more succinct or laid out more neatly. This useful feedback will be incorporated into future presentations.

6 Impact

This seminal paper had a huge impact and was ground-breaking not only in the field of NLP but machine learning in general. The success of Transformers in neural machine translation prompted research into alternative tasks across NLP. Transformer-based architectures have achieved state-of-the-art results on benchmarks for tasks including text summarisation, question-answering and sentiment analysis, demonstrating the effectiveness of Transformers in capturing the intricate patterns in natural language. Transformer-based architectures have since dominated the field of NLP, with the introduction of powerful cutting-edge large language models such as Generative Pre-trained Transformers (GPT) [14] and Bidirectional Encoder Representations from Transformers (BERT)[15].

GPT is a large language model, developed by OpenAI, based on the Transformer decoder. It is pre-trained on a large corpus of text to predict the next token given the previous tokens in the sequence [16]. Hence GPT is unidirectional. On the other hand, BERT is a large language model, developed by Google, based on the Transformer encoder. In contrast to GPT, BERT is bidirectional and is trained to predict randomly masked words in the sequence. This enables BERT to learn bidirectional context. [15]. Researchers have also since investigated the potential of Transformer-based architectures outside of NLP, such as computer vision [17] and time-series applications [18].

In terms of healthcare applications, particularly with the large amounts of data from electronic health records, there has been research into the analysis of clinical time-series data. The authors of [19] were among the first to introduce a Transformer-based model for multivariate clinical time-series classification. Clinical time series are particularly challenging due to sparsity and irregularity of measurements. By leveraging triplets of observations, [20] demonstrated that the Self-supervised Transformer for Time-Series (STraTS) achieved state-of-the-art performance on clinical time-series benchmarks.

7 Summary

In summary, the introduction of the Transformer architecture [7] marked a paradigm shift in sequence-to-sequence modelling for NLP, providing an effective alternative to recurrent and convolutional approaches through the use of self-attention mechanisms. The authors demonstrated the model's ability to parallelise computations and capture long-range dependencies, achieving state-of-the-art BLEU scores for English-to-German and English-to-French translation at a fraction of the computation cost. Transformer-based architectures have since become widely used and state-of-the-art across NLP tasks. The huge success of Transformers led to research outside of NLP applications in computer vision for example and exciting research is ongoing.

8 Acknowledgements

This work was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017.

[2] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 4 1990.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[4] Ilya Sutskever Google, Oriol Vinyals Google, and Quoc V Le Google. Sequence to Sequence Learning with Neural Networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[6] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *34th International Conference on Machine Learning, ICML*, 70:1243–1252, 2017.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. *International Conference on Learning Representations*, 2015.

[8] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*, 2013.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[10] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. *Conference on Empirical Methods in Natural Language Processing 2015*, pages 1412–1421, 2015.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311, 2002.

[12] Sandeep Kumar and Arun Solanki. An abstractive text summarization technique using transformer model with self-attention mechanism. *Neural Computing and Applications*, 35(25):18603–18622, 9 2023.

[13] Surafel M Lakew, Mauro Cettolo Fondazione, and Marcello Federico. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, 2018.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI*, 2019.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Google AI*, 2019.

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *OpenAI*, 2018.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, S. Gelly, Jakob Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*, 2020.

[18] Azad Deihim, Eduardo Alonso, and Dimitra Apostolopoulou. STTRE: A Spatio-Temporal Transformer with Relative Embeddings for multivariate time series forecasting. *Neural Networks*, 168:549–559, 11 2023.

[19] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and Diagnose: Clinical Time Series Analysis Using Attention Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):4091–4098, 4 2018.

[20] Sindhu Tipirneni and Chandan K. Reddy. Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6), 7 2022.