

# How to read scientific papers

6

## Attention Is All You Need

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\*<sup>†</sup>  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukasz.kaiser@google.com

Illia Polosukhin\*<sup>‡</sup>  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

### 1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [2] neural networks, in particular, have been firmly established as state of the art approaches in sequence

\*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer. Noam proposed scaled dot-product attention and the parameter-free position representation and became the other person in detail. Niki designed, implemented, tuned and evaluated countless model variants in our tensorflow2tensor. Llion also experimented with novel model variants, was responsible for our efficient inference and visualizations. Lukasz and Aidan spent countless long days designing, implementing tensorflow2tensor, replacing our earlier codebase, greatly improving results and making our research.

<sup>†</sup>Work performed while at Google Brain.

<sup>‡</sup>Work performed while at Google Research.

2

4

1

5

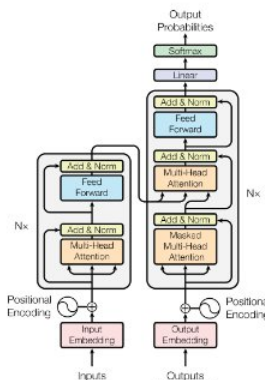


Figure 1: The Transformer - model architecture.

### Encoder Stacks

The encoder is composed of a stack of  $N = 6$  identical layers. Each layer has two sub-layers: a multi-head self-attention mechanism, and the second is a simple, position-wise feed-forward network. We employ a residual connection [11] around each of these, followed by layer normalization [10]. That is, the output of each sub-layer is  $\text{Sublayer}(x)$ , where  $\text{Sublayer}(x)$  is the function implemented by the sub-layer; these residual connections, all sub-layers in the model, as well as the embedding

inputs of dimension  $d_{\text{model}} = 512$ .

The decoder is also composed of a stack of  $N = 6$  identical layers. In addition to the encoder stack, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections, followed by layer normalization. We also modify the second sub-layer to prevent positions from attending to subsequent positions. This is done by masking the output embeddings are offset by one position, ensuring that the output at position  $i$  can depend only on the known outputs at positions less than  $i$ .

The output can be described as mapping a query and a set of key-value pairs to an output, given weights, and output are all vectors. The output is computed as a weighted sum of the weight assigned to each value is computed by a compatibility function of the corresponding key.

constraints and is significantly longer than the input. Furthermore, RNN sequence-to-sequence models have not been able to attain state-of-the-art results in small-data regimes [37].

We trained a 4-layer transformer with  $d_{\text{model}} = 1024$  on the Wall Street Journal (WSJ) portion of the Penn Treebank [25], about 40K training sentences. We also trained it in a semi-supervised setting, using the larger high-confidence and BerkeleyParser corpora from with approximately 17M sentences [27]. We used a vocabulary of 16K tokens for the WSJ only setting and a vocabulary of 32K tokens for the semi-supervised setting.

We performed only a small number of experiments to select the dropout, both attention and residual (section 5.3), learning rates and beam size on the Section 22 development set, all other parameters remained unchanged from the English-to-German base translation model. During inference, we increased the maximum output length to input length + 300. We used a beam size of 21 and  $\alpha = 0.3$  for both WSJ only and the semi-supervised setting.

Our results in Table 5 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [32], the Transformer outperforms the BerkeleyParser [29] even when training only on the WSJ training set of 40K sentences.

### 7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures.

For translation tasks, the Transformer can be trained significantly faster than the previous models on recurrent or convolutional layers. On both WMT 2014 English-to-German and English-to-French translation tasks, we achieve a new state of the art. In the future, we plan to extend the Transformer to problems involving input and output modalities such as images, audio and video. Making generation less sequential is another research goal.

We are excited about the future of attention-based models and plan to apply the Transformer to a wide range of problems. We plan to investigate local, restricted attention mechanisms to efficiently handle large inputs, such as images, audio and video. Making generation less sequential is another research goal.

The code we used to train and evaluate our models is available at <https://github.com/google-research/transformer>.

**Acknowledgements** We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

### References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06448*, 2016.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.
- [3] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V Le. Massive exploration of neural machine translation architectures. *CoRR*, abs/1703.03906, 2017.
- [4] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
- [5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.
- [6] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016.

Always:

1. Figures + captions
2. Abstract
3. Conclusion

Want to know more?

4. Introduction
5. Methods

Want to know all?

6. read full paper

Only if required:

7. Equations