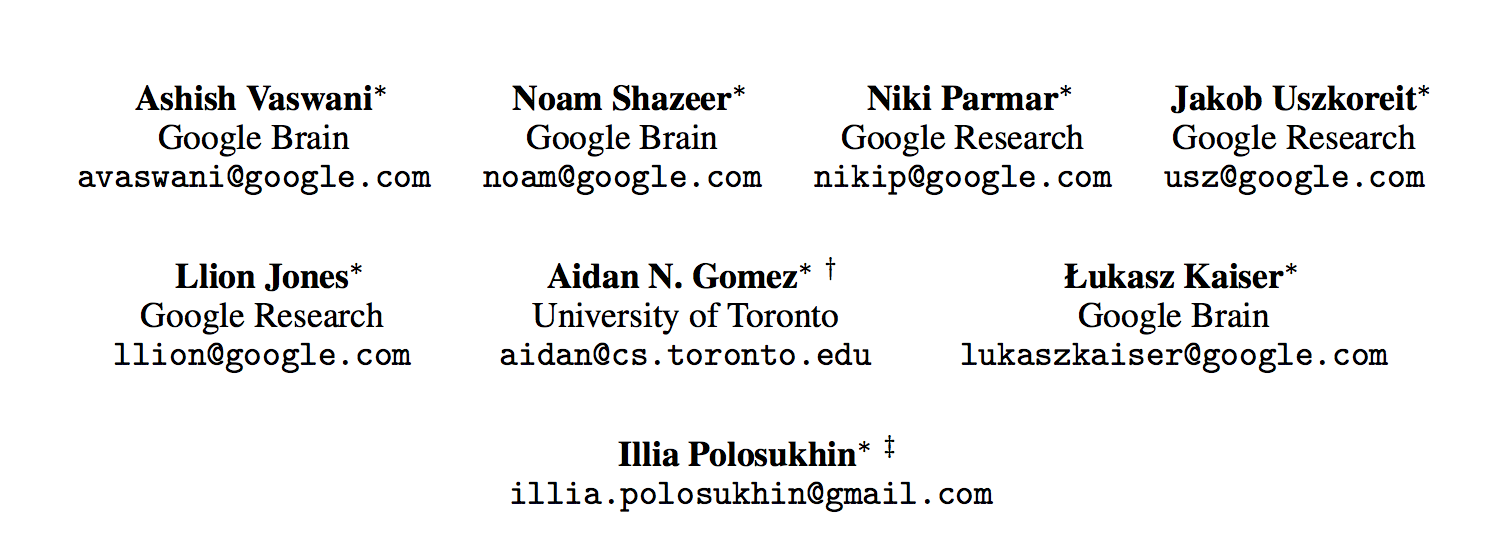
**Attention Is All You Need**



**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.

**摘要**

主流的序列传导模型基于包括编码器和解码器的复杂递归或卷积神经网络。最优模型还通过一个注意力机制连接编码器和解码器。我们提出一个新的简单的网络架构, Transformer, 其完全基于注意力机制, 并重新分配递归与卷积。在两个机器翻译任务上的实验表明, 这些模型在质量上优于并行, 且训练时间显著减少。我们的模型在 WMT 2014 的英语-德语翻译任务中实现了28.4 BLEU, 提高了超过 2 BLEU 的现有最佳效果, 包括总效果。在 WMT 2014 英语-法语的翻译任务中, 在8个GPU上训练了3.5天后，我们的模型达到了一个新单一模式的、先进的 BLEU 评分41.8 ，这是最佳模型训练成本的一小部分。可以看出，Transformer通过对大量、有限的训练数据进行英语成分分析, 可以成功地推广到其他任务中。