# Transformer读书笔记

论文地址：https://arxiv.org/abs/1706.03762

## Q1 论文试图解决什么问题?

论文试图解决的问题是如何有效地应用深度学习模型来处理自然语言处理序列转换的任务。传统的序列转换模型，如循环神经网络（RNN）和卷积神经网络（CNN），传统的一系列RNN在处理序列时存在着短期记忆和长期依赖的问题，如难以捕捉长距离的依赖关系，因为是顺序计算所以并行计算能力较弱导致训练和推理速度慢等问题。因此，本文提出了一种新的序列转换模型——Transformer模型，摒弃了循环和卷积，用于在NLP任务中进行序列到序列的学习，而且训练时间更少。

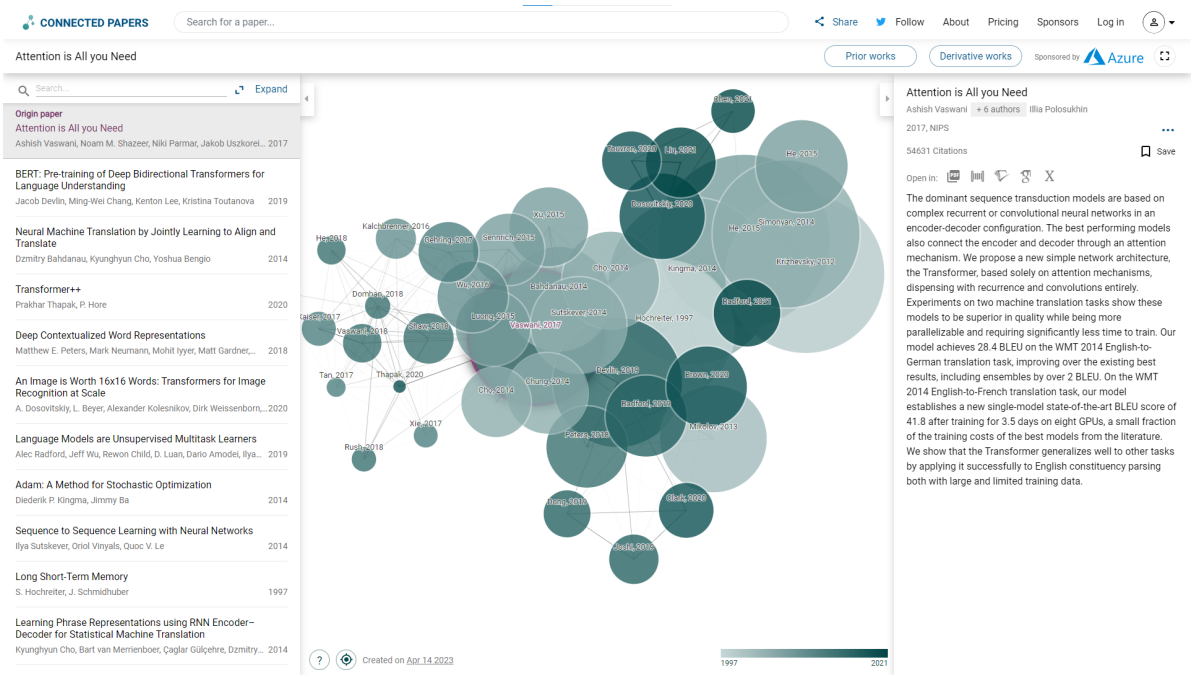
## Q2 这是否是一个新的问题?

在NLP领域中，序列到序列的学习一直是一个重要的问题，但在本文发表之前，主要的解决方案是基于RNN或CNN的模型。而在本文章之前，有不少文章也同样提出了一系列减少顺序计算的方法，但它们都是使用卷积神经网络来学习远距离位置之间的依赖关系。而本文提出的Transformer模型摒弃了之前的卷积架构，而且将操作数量降低至常数级别，并用Multi-Head Attention抵消有效分辨率下降的问题，因此可以被视为一种新的解决方案。

## Q3 这篇文章要验证一个什么科学假设?

本文的科学假设总结来说就是仅仅引入自注意力机制就已经够了（即标题Attention Is All You Need）。而实际上文章中也证明了在序列转换以及学习句法结构的任务上都取得了最新、最佳的结果，文章从编码器解码器的架构到自注意力的原理等等设计到实验结果来验证提出的Transformer模型是否可以在不使用RNN或CNN的情况下，在NLP任务中实现序列到序列学习，并且是否比传统的序列转换模型更好。

## Q4 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员?




我们结合connectedpapers.com上画出相关论文的图的结果和原论文的引用文献来总结归类之前的相关研究：

### 序列建模和转换

- [1409.3215] Sequence to Sequence Learning with Neural Networks (arxiv.org)在英法翻译任务上做的评估，这篇文章使用的架构是LSTM架构
- [1409.0473] Neural Machine Translation by Jointly Learning to Align and Translate (arxiv.org) 使用encoder-decoder 并介绍神经机器翻译（NMT）是如何进行对齐的
- [1406.1078] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (arxiv.org)使用RNN的编码器译码器来做的传统机器翻译

### RNN与编码器-解码器

- [1609.08144] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (arxiv.org)使用深度LSTM网络组成，具有8个编码器层和8个解码器层，使用注意力和残差连接做的WMT'14英法和英德任务。
- [1508.04025] Effective Approaches to Attention-based Neural Machine Translation (arxiv.org) 研究了不同注意力方法（全局源单词或局部源单词）的NMT去做WMT英德翻译任务。
- [1602.02410] Exploring the Limits of Language Modeling (arxiv.org)探索了RNNs在语言模型的边界，主要在语料库和词汇表的大小，以及语言的复杂、长期结构方面。

### 循环模型的技巧

- [1703.10722] Factorization tricks for LSTM networks (arxiv.org)介绍了lstm的分解技巧。
- [1701.06538] Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (arxiv.org)介绍了条件计算以减少计算成本。

### 注意力机制

- [1508.04025] Effective Approaches to Attention-based Neural Machine Translation (arxiv.org) 研究了不同注意力方法（全局源单词或局部源单词）的NMT去做WMT英德翻译任务。
- [1702.00887v2] Structured Attention Networks (arxiv.org)扩展了注意力机制的方法：结构化注意力网络
- [1606.01933] A Decomposable Attention Model for Natural Language Inference (arxiv.org)提出了不用RNN的自然语言推理神经架构

### 值得关注的研究人员

- Noam Shazeer 上面的Exploring the limits of language modeling和Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.都是其著作，领域在条件计算和模型的边界方面，同时也是本文的作者；
- Jakob Uszkoreit 2016年提出的不用rnn的架构A decomposable attention model，领域就是注意力方面，同时也是本文作者之一，本文中提出去除rnn的就是他；
- Łukasz Kaiser 本文作者之一，负责设计源代码，同时也引用了三篇他的论文：. Can active memory replace attention?和Neural GPUs learn algorithms和Multi-task sequence to sequence learning.，主要领域在gpu、memory、attention等方面；
- Yoshua Bengio 之前的关于RNN、GRU、NMT的一系列文章都是Bengio的学生发的；

## Q5 论文中提到的解决方案之关键是什么?

论文中提到的解决方案的关键是Transformer模型中的自注意力机制。自注意力机制允许模型在处理序列时同时关注序列中的所有位置，而不是像传统的序列转换模型那样只关注前面的位置。因此，自注意力机制能够更好地捕捉长距离的依赖关系，从而提高模型的性能。

## Q6 论文中的实验是如何设计的?

# Q7 用于定量评估的数据集是什么？代码有没有开源？

## 数据集

在前面进行翻译任务中用于定量评估的数据集采用的是WMT 2014英德和英法翻译任务的数据。

而后面用于英语成分句法分析的数据集则采用的是华尔街日报的数据，还使用了BerkleyParser语料库。

## 代码

代码已经开源，原论文中给出的代码网址是[tensorflow/tensor2tensor: Library of deep learning models and datasets designed to make deep learning more accessible and accelerate ML research. (github.com)]

但实际上哈佛的nlp团队用pytorch实现了带注释版代码的论文文章，而且里面也举了一些例子使得更容易理解：

[harvardnlp/annotated-transformer: An annotated implementation of the Transformer paper. (github.com)]

# Q8 论文中的实验及结果有没有很好地支持需要验证的科学假设？

论文中的实验结果表明，Transformer模型在机器翻译和语言建模任务中的性能都优于其他基于RNN和CNN的模型。这表明Transformer模型可以在不使用RNN和CNN的情况下，实现序列到序列的学习，并且比传统的序列转换模型更好。因此，论文中的实验及结果很好地支持了需要验证的科学假设。

# Q9 这篇论文到底有什么贡献？

本文提出了一种新的序列转换模型——Transformer模型，用于在NLP任务中进行序列到序列的学习。与传统的序列转换模型相比，Transformer模型具有以下优点：可以并行计算，能够处理长序列，且不需要使用RNN或CNN等循环结构。在实验中，Transformer模型在机器翻译和语言建模任务中的性能都优于其他基于RNN和CNN的模型。因此，本文提出的Transformer模型对于解决序列转换问题具有重要的理论和实际价值。

# Q10 下一步呢？有什么工作可以继续深入？

在本文之后，已经有很多研究工作对Transformer模型进行了进一步的改进和扩展。例如，有一些研究工作探索了如何在Transformer模型中引入额外的特征，以提高模型的性能。另外，一些研究工作还探索了如何使用Transformer模型进行多语言翻译和跨语言学习等任务。因此，未来的工作可以继续深入研究Transformer模型及其应用，在性能和应用方面进行更进一步的改进和探索。