注意力就是一切

Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit
Google Brain Google Brain Google Research Google Research
avaswani@google.com noam@google.com nikip@google.com usz@google.com

Llion Jones Aidan N. Gomez † ukasz Kaiser
Google Research University of Toronto Google Brain
llion@google.com aidan@cs.toronto.edu lukaszkaiser@google.com

Illia Polosukhin ‡
illia.polosukhin@gmail.com

摘要

目前主导的序列转导模型基于复杂的循环或卷积神经网络，包括编码器和解码器。表现最好的模型还通过注意力机制连接编码器和解码器。我们提出了一种新的简单网络架构，Transformer，仅基于注意力机制，完全摒弃了循环和卷积。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其他模型，同时更易并行化，并且训练时间显著缩短。我们的模型在WMT 2014年英德翻译任务上达到了28.4
BLEU，相比现有的最佳结果（包括集成模型），提高了超过2 BLEU。在WMT 2014年英法翻译任务中，我们的模型在训练了3.5天的情况下，在八个GPU上实现了新的单模型最佳BLEU得分41.8，这只是文献中最佳模型训练成本的一小部分。我们展示了Transformer在其他任务上的泛化能力，通过成功将其应用于英语成分句法分析，无论是使用大量还是有限的训练数据。

平等贡献。列表顺序是随机的。Jakob提出用自注意力替换RNN，并开始评估这个想法。Ashish和Illia设计和实现了第一个Transformer模型，并在这项工作的各个方面起到了至关重要的作用。Noam提出了缩放的点积注意力、多头注意力和无参数位置表示，并成为几乎每个细节中的另一个参与者。Niki在我们的原始代码库和tensor2tensor中设计、实现、调优和评估了无数的模型变体。Llion还尝试了新颖的模型变体，负责我们的初始代码库，以及高效的推理和可视化。Lukasz和Aidan花费了无数的长时间来设计tensor2tensor的各个部分，并替换了我们之前的代码库，大大改进了结果并大幅加速了我们的研究。

†在Google Brain工作期间完成的工作。
‡在Google Research工作期间完成的工作。
第31届神经信息处理系统会议（NIPS2017），美国加利福尼亚州长滩。

1 介绍

循环神经网络、长短期记忆网络[13]和门控循环神经网络[7]已经被确定为序列建模和转导问题（如语言建模和机器翻译[35,2,5]）中的最先进方法。自那时以来，许多工作一直在推动循环语言模型和编码器-解码器架构的边界[38,24,15]。
循环模型通常沿着输入和输出序列的符号位置进行计算。通过将位置与计算时间步骤对齐，它们生成一个隐藏状态序列h，作为前一个隐藏状态h和位置t的输入的函数。然而，这种顺序性质阻止了在训练示例内部的并行化，这在较长的序列长度下变得关键，因为内存限制限制了跨示例的批处理。最近的工作通过因式分解技巧[21]和条件计算[32]在计算效率方面取得了显著的改进，同时在后者的情况下也提高了模型性能。然而，顺序计算的基本约束仍然存在。
注意力机制已经成为引人注目的序列建模和转导模型在各种任务中的一个重要组成部分，它允许对依赖关系进行建模，而不考虑它们在输入或输出序列中的距离[2,19]。然而，在除了少数情况[27]之外，这些注意力机制通常与循环网络一起使用。
在这项工作中，我们提出了Transformer，这是一种模型架构，它摒弃了循环性，而是完全依赖注意力机制来绘制输入和输出之间的全局依赖关系。Transformer允许更高度的并行化，并且在经过仅仅12小时在八个P100
GPU上训练后，可以达到翻译质量的新水平。

2 背景

减少顺序计算的目标也是Extended Neural GPU[16]、ByteNet[18]和ConvS2S[9]的基础，它们都使用卷积神经网络作为基本构建块，为所有输入和输出位置并行计算隐藏表示。在这些模型中，从两个任意输入或输出位置关联信号所需的操作数量随着位置之间的距离增加而增加，对于ConvS2S是线性增长，对于ByteNet是对数增长。这使得学习远距离位置之间的依赖关系更加困难。在Transformer中，这被减少为一定数量的操作，尽管由于对注意力加权位置的平均化效果，导致了有效分辨率的降低，我们通过Multi-Head Attention来抵消这种效果，如第3.2节所述。
自注意力，有时称为内部注意力，是一种关联单个序列的不同位置以计算序列表示的注意力机制。自注意力已经成功地应用于各种任务，包括阅读理解、抽象摘要、文本蕴含和学习任务无关的句子表示[4,27,28,22]。
端到端记忆网络基于循环注意力机制而不是序列对齐的重复，已经在简单语言问答和语言建模任务上表现良好[34]。
据我们所知，然而，Transformer是第一个完全依赖自注意力来计算其输入和输出表示的传导模型，而不使用序列对齐的RNN或卷积。在接下来的几节中，我们将描述Transformer，解释自注意力并讨论其相对于模型[17,18]和[9]的优势。

3 模型架构

大多数竞争性的神经序列转导模型都具有编码器-解码器结构[5,2,35]。在这里，编码器将一个符号表示的输入序列(x1,...,xn)映射到一个连续表示的序列z=(z1,...,zn)。给定z，解码器然后生成一个符号的输出序列(y1,...,ym)，每次一个元素。在每个步骤中，模型是自回归的[10]，在生成下一个符号时，消耗先前生成的符号作为附加输入。