提供适当的归属，Google特此授予权限，仅可在新闻或学术作品中使用本论文中的表格和图表。

《注意力就是一切》

AshishVaswani NoamShazeer NikiParmar JakobUszkoreit
GoogleBrain GoogleBrain GoogleResearch GoogleResearch
avaswani@google.com noam@google.com nikip@google.com usz@google.com

LlionJones AidanN.Gomez † ukaszKaiser
GoogleResearch UniversityofToronto GoogleBrain
llion@google.com aidan@cs.toronto.edu lukaszkaiser@google.com

IlliaPolosukhin ‡
illia.polosukhin@gmail.com

摘要

主导的序列转导模型是基于复杂的循环或卷积神经网络，包括编码器和解码器。性能最好的模型也通过注意力机制将编码器和解码器连接起来。我们提出了一种新的简单网络架构，Transformer，完全基于注意力机制，完全放弃了循环和卷积。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以往的模型，同时更易于并行化，并且训练时间显著缩短。我们的模型在WMT 2014英德翻译任务上达到了28.4
BLEU，相较于现有的最佳结果（包括集成模型），提高了超过2 BLEU。在WMT 2014英法翻译任务中，我们的模型在训练了3.5天、使用了八个GPU后，得到了新的单模型最优BLEU分数41.8，这只占了从文献中最佳模型的一小部分训练成本。我们展示了Transformer在其他任务上的泛化能力，通过成功应用于英文成分句法分析，并处理大量和有限的训练数据。

*等贡献，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与Illia一起设计和实现了第一个Transformer模型，并在这项工作的每个方面都起到了至关重要的作用。Noam提出了可缩放的点积注意力、多头注意力和无参数的位置表示，并成为几乎每个细节中的另一个重要人物。Niki在我们的原始代码库和tensor2tensor中设计、实现、调优和评估了无数个模型变种。Llion也尝试了新颖的模型变体，负责我们的初始代码库，并进行高效的推理和可视化。Lukasz和Aidan花费了很多时间设计tensor2tensor的各个部分，并替换了我们之前的代码库，大大提高了结果并大幅加快了我们的研究。

†在GoogleBrain期间进行的工作。
‡在GoogleResearch期间进行的工作。

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 简介

递归神经网络，长短期记忆[13]和门控循环神经网络[7]在序列建模和转换问题（如语言建模和机器翻译）中，被确定为最先进的方法[35,2,5]。多个研究努力继续推动递归语言模型和编码器-解码器架构的边界[38,24,15]。
递归模型通常将计算沿着输入和输出序列的符号位置分解。通过将位置对齐到计算时间的步骤，它们生成一个隐藏状态序列h，作为位置t的前一个隐藏状态h和输入的函数。然而，这种顺序性质在训练示例内部阻止了并行化，这在较长的序列长度上变得关键，因为内存限制限制了跨示例的批处理。最近的工作通过因子化技巧[21]和条件计算[32]在计算效率上取得了显著的改进，同时也在后一种情况下改善了模型的性能。然而，顺序计算的基本约束仍然存在。

注意机制已经成为引人注目的序列建模和转换模型在各种任务中的积极部分，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，这种注意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，这是一种模型体系结构，它摒弃了递归性，而是完全依赖于注意机制来绘制输入和输出之间的全局依赖关系。Transformer允许更高程度的并行化，并可以在仅在八个P100
GPU上训练了仅12小时后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，在所有输入和输出位置上并行计算隐藏表示。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，这被减少到了一定数量的操作次数，尽管由于对注意权重位置的平均，导致了有效分辨率的减少，我们通过多头注意力来抵消这种效果，如3.2节中所述。

自注意力，有时称为内部注意力，是一个注意机制，它涉及到同一个序列中不同位置之间的关系，以计算序列的表示。自注意力已经在多种任务中成功使用，包括阅读理解、抽象总结、文本蕴涵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机制而不是序列对齐循环，并且已经在简单语言问答和语言建模任务中表现良好[34]。

据我们所知，然而，Transformer是第一个完全依赖于自注意力来计算其输入和输出的表示的转换模型，而不使用序列对齐的循环神经网络或卷积。在接下来的几节中，我们将描述Transformer，论证自注意力及其相比于[17,18]和[9]等模型的优势。

3 模型架构

大多数具有竞争力的神经序列转导模型具有编码器-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，将先前生成的符号作为额外的输入。