Attention Is All You Need

NeurIPS 2017 Google

李智浩 2025/4/21

content

目录

引言与背景

01

传统序列建模方法

RNN与LSTM

循环神经网络(RNN)及其变体长短期记忆网络(LSTM)曾是处理序列数据的主流方法，通过隐藏状态传递信息，有效处理序列依赖。

CNN的应用

卷积神经网络(CNN)在固定长度的窗口内提取特征，适用于图像识别，也被尝试应用于序列建模，但受限于固定窗口大小。

局限性

RNN和LSTM难以并行处理，计算效率低；CNN虽能并行，但难以捕捉长距离依赖，限制了其在序列任务上的表现。

注意力机制

1.查询、键和值：在注意力机制中，输入数据被分成三部分：查询（query）、键（key）和值（value）。这些都是向量：

•查询（Q）：表示你当前关注的点或问题。

• 键（K）：表示输入数据的不同特征。

• 值（V）：表示与键相关的实际信息或答案。

1.计算相关性：首先，计算查询与所有键的相似度（相关性），这可以通过点积运算完成。相似度越高，表示这个键与查询越相关。

2.加权求和：将这些相似度通过softmax函数转换为概率权重，然后用这些权重对所有值进行加权求和。这样，模型输出一个综合了所有相关值的信息，突出重要信息而忽略不相关信息。

可以把注意力机制想象成一个学生在阅读一本书。学生在读每一页时，不会平均分配注意力，而是会根据当前问题（查询）来重点关注相关的段落或句子（键和值）。例如，如果学生在回答历史问题时，他们会特别注意书中有关历史事件的部分，而忽略其他不相关的内容。

Scaled Dot-Product Attention

给定一个输入序列 $x_1, x_2, \ldots, x_n$，注意力机制会计算每个位置之间的相关性（相似度），然后对每个位置加权求和：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

其中：

- Q 是 Query（查询）
- K 是 Key（键）
- V 是 Value（值）
- softmax 用于得到归一化权重
- d_k 是缩放因子（防止数值过大）

注意力机制的优势

并行处理能力

注意力机制允许模型同时关注输入序列的不同部分，显著提升训练速度和效率。

长距离依赖解决

有效捕捉远距离词汇间的关系，克服RNN和CNN在处理长序列时的局限性。

灵活性与适应性

注意力权重动态调整，使模型能根据任务需求灵活分配资源，提高模型的泛化能力。

直观可解释性

注意力权重可视化，揭示模型决策过程，增强模型的透明度和可解释性。

Transformer的提出背景

解决效率瓶颈

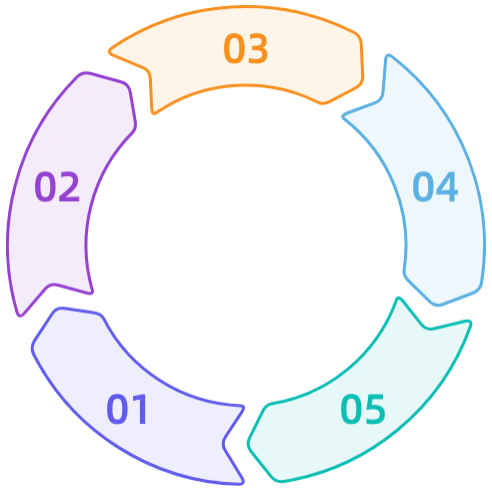
解决了RNN和LSTM处理长序列时的效率瓶颈问题。

注意力机制

采用注意力机制实现高效并行计算，提升模型性能。

Transformer推出

2017年由Google团队推出，革新了自然语言处理领域。



长距离依赖

有效处理长距离依赖问题，改善了模型的表达能力。

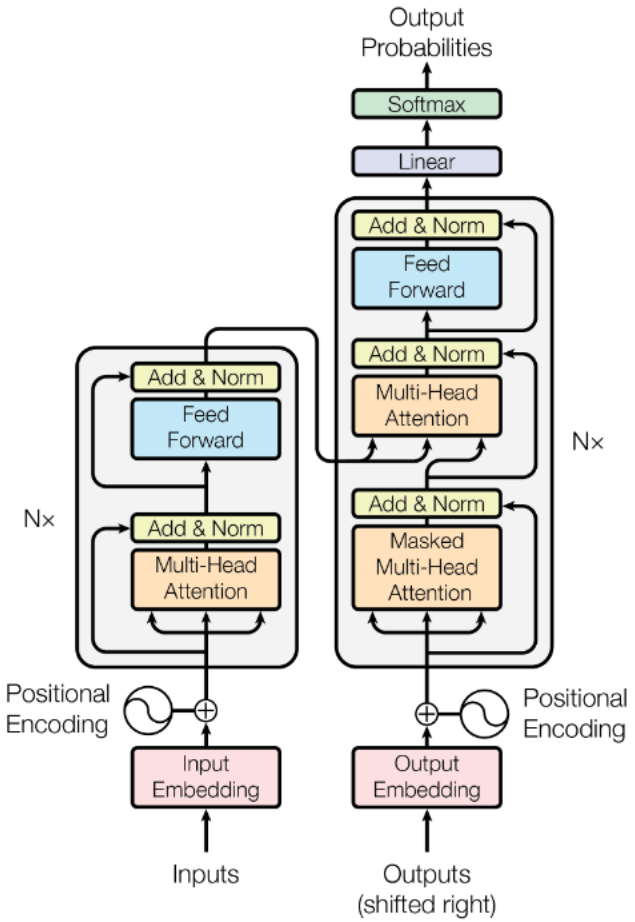
模型性能提升

显著提升了模型性能，推动了自然语言处理技术的发展。

Transformer架构概述

02

整体结构


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

整体结构

01

采用Encoder-Decoder

结构上采用编码器-解码器架构，摒弃传统循环与卷积。

02

全注意力机制

使用全注意力机制，贯穿整个模型，提高模型性能。

03

编码器模块化

编码器由多层相同模块组成，每层包括多头自注意力和前馈神经网络。

04

解码器引入掩蔽

解码器在编码器基础上，增加掩蔽多头自注意力，用于在训练过程中防止未来位置信息泄露，从而保证自回归生成的合理性。

并行性与效率

并行处理能力

摒弃RNN依赖序列处理的限制，Transformer允许所有序列元素同时计算，显著提升训练速度。

高效计算

通过多头注意力机制，Transformer能够并行处理不同子空间的信息，加速模型训练过程。

资源利用优化

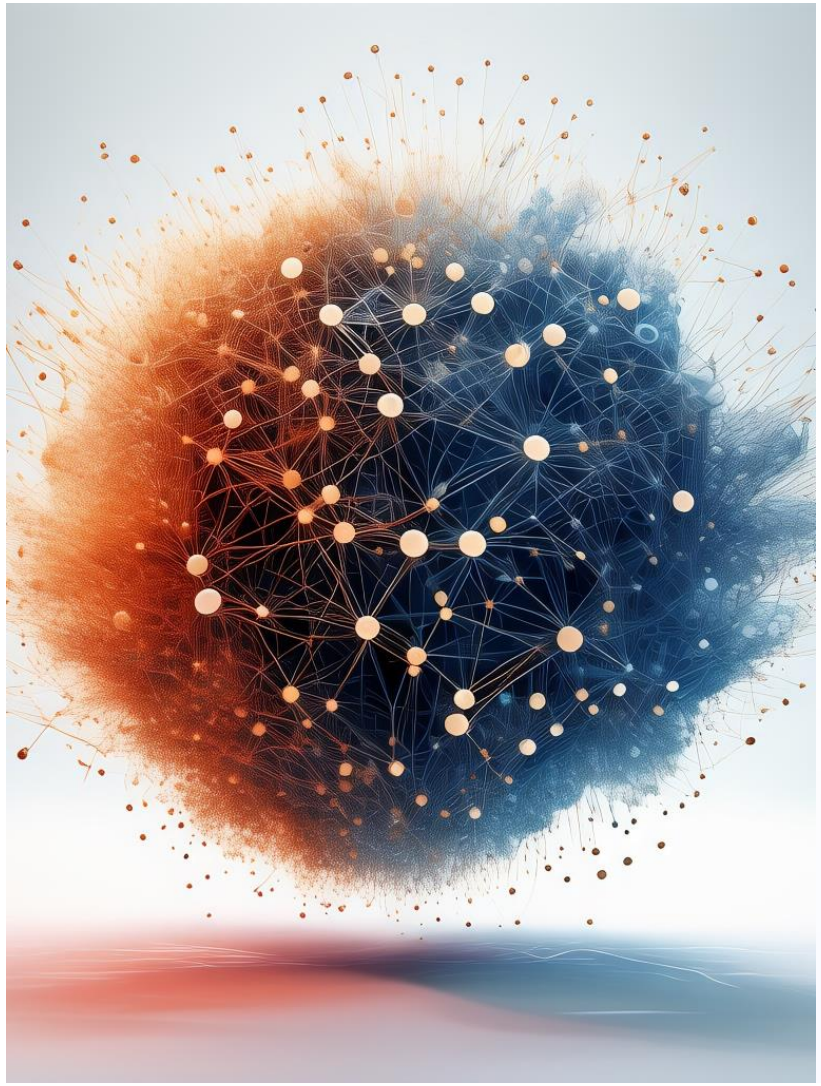
得益于其并行架构，Transformer能更有效地利用GPU资源，减少训练时间和成本。

扩展性增强

并行性使得Transformer易于在大规模数据集上训练，支持更复杂的模型结构和更大的模型规模。

编码器与解码器结构

03



编码器详解

多层堆叠结构

编码器由多层相同结构堆叠而成，每层包含多头自注意力机制与前馈神经网络，形成深度学习模型的基础单元。

自注意力机制

通过多头自注意力，编码器能够并行处理输入序列，有效捕捉长距离依赖关系，提高模型的并行性和效率。

前馈神经网络

每层的前馈神经网络对序列中的每个位置进行非线性变换，增强模型的表达能力，适应复杂的数据分布。

残差连接与归一化

残差连接加速训练过程，层归一化稳定梯度，共同促进深层网络的训练，提升模型性能和稳定性。

解码器结构及掩蔽机制

解码器概述

解码器采用与编码器相似但更复杂的结构，包含多头自注意力机制，特别设计用于处理序列生成任务。

掩蔽多头自注意力

通过掩蔽机制，解码器在预测序列时只能看到之前的信息，避免了未来信息的泄露，确保了生成过程的正确性。

位置信息的重要性

由于解码器需生成序列，位置编码在此过程中至关重要，确保模型理解序列中元素的顺序和位置关系。

前馈网络的角色

解码器中的前馈神经网络负责对每个位置的表示进行非线性变换，增强模型的表达能力，促进更高质量的序列生成。

残差连接与层归一化

残差连接原理

残差连接通过将输入直接加到层的输出上，解决了深层网络中的梯度消失问题，提高了模型的训练稳定性。

层归一化作用

层归一化在每一层的输出上进行，它加速了训练过程，减少了内部协变量偏移，有助于模型收敛。

结合效果

残差连接与层归一化共同作用，不仅提升了模型的深度，还保证了模型在训练过程中的高效与稳定。

实践意义

在Transformer中，这两项技术的应用极大地增强了模型的性能，使其在多种任务上取得显著成果。

核心组件详解

04

Scaled Dot-Product Attention

01

查询矩阵Q

用于表示输入序列中的查询信息。

02

键矩阵K转置

与查询矩阵相乘前需要进行转置操作。

03

除以根号d_k

标准化处理，确保数值稳定性。

04

softmax函数

将计算结果转换为概率分布形式的注意力权重。

05

值矩阵V

与注意力权重相乘，生成最终输出。

06

提高理解能力

使模型能够更有效地关注输入序列中的相关信息。

Multi-Head Attention

并行执行

通过并行地执行多个注意力机制，Multi-Head Attention能够同时捕捉输入的不同表示子空间，显著提升模型的并行性和效率。

子空间捕捉

每个注意力头关注输入的不同方面，允许模型从多个角度理解序列，增强其捕捉复杂模式的能力。

权重矩阵

每个头都有独立的权重矩阵，用于投影查询、键和值向量，从而实现对不同子空间的关注。

结果整合

所有头的结果被拼接后通过一个全连接层，以整合来自不同子空间的信息，形成最终的注意力输出。

Multi-Head Attention

注意力机制在Transformer模型中的应用

Transformer模型通过以下三种方式应用多头注意力：

编码器-解码器注意力层：在这层中，查询来自前一个解码器层，而记忆键和值则来自编码器的输出。这允许解码器中的每个位置都能关注输入序列的所有位置。
编码器自注意力层：在这层中，所有的键、值和查询都源自编码器前一层的同一位置。编码器的每个位置都能关注到前一层的所有位置。
解码器自注意力层：这层允许解码器中的每个位置只关注到该位置之前的所有位置，以保持自回归的特性。通过在缩放点积注意力中引入掩码机制，能够实现这一点，掩码会将softmax操作中非法连接的输入值设为负无穷，从而排除这些连接。

位置编码

位置信息的重要性

位置编码赋予模型理解序列中词序的能力，弥补了自注意力机制缺乏位置信息的不足。

正弦函数编码

采用正弦和余弦函数，根据位置和维度动态生成编码，确保模型能区分不同位置的词汇。

频率特性

不同维度的编码具有不同的频率，低频编码关注长距离依赖，高频编码捕捉局部特征。

可加性与线性变换

位置编码设计为可加到词嵌入上，并能通过线性变换层进一步处理，无缝融入Transformer架构。

前馈神经网络

01

非线性变换

前馈神经网络对每个位置的表示进行非线性变换，增强模型的表达能力。

02

两层结构

由两个全连接层组成，第一层宽度较大，第二层恢复到输入维度，形成瓶颈结构。

03

激活函数

使用ReLU或GELU作为激活函数，增加模型的非线性特性，促进特征学习。

04

并行处理

前馈网络可以并行处理所有位置的输入，提高模型的训练和推理效率。

模型训练与优化

05

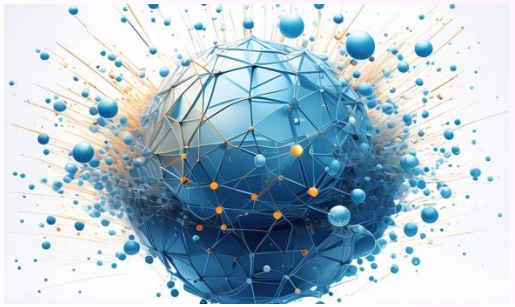
训练数据集

数据集选择

采用WMT 2014英德与英法翻译任务数据集，涵盖大量平行语料库，为模型训练提供坚实基础。

数据规模

英德翻译任务包含约450万句子对，英法翻译任务则拥有约3600万句子对，确保模型充分学习语言规律。



数据质量

高质量的双语对照文本，有效促进Transformer模型在翻译准确性上的突破，实现卓越性能。

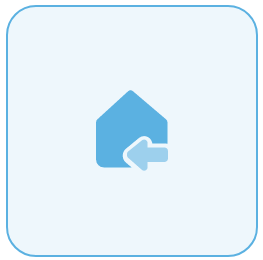
优化器与学习率调度

优化器选择

采用Adam优化器，结合梯度累积，有效提升模型收敛速度与稳定性。

动态调整

根据模型训练状态动态调整学习率，平衡探索与利用，加速收敛过程。



学习率策略

实施warm-up与衰减策略，初始阶段快速学习，随后逐步降低学习率以精细调整。

正则化技术

批量归一化技术

在每个小批次上对数据进行标准化处理，加快训练速度，同时稳定梯度更新过程。

权重衰减应用

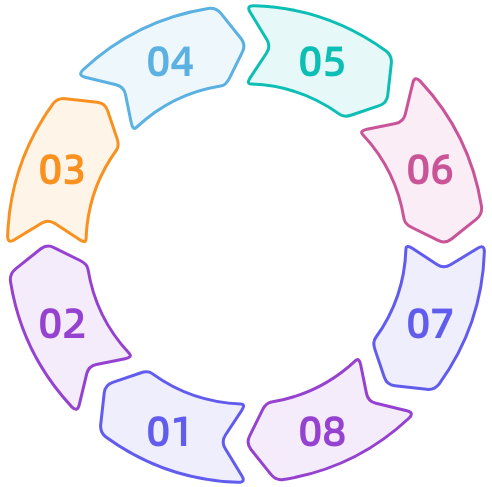
通过正则化项限制权重值的大小，控制模型复杂度，有效降低过拟合的风险。

标签平滑处理

将硬性的one-hot标签转换为软标签，增加模型对不确定性的容忍度，提高鲁棒性。

Dropout随机失活

在训练过程中随机失活部分神经元，减少模型对特定特征的依赖，提高泛化能力。



减少过拟合现象

结合多种方法综合减少过拟合，包括Dropout、权重衰减等，确保模型具有良好的泛化性能。

提升模型鲁棒性

利用标签平滑等技术增强模型对外界变化的适应能力，避免模型过度自信导致的错误预测。

改善训练稳定性

通过批量归一化解决内部协变量偏移问题，使训练过程更加平稳，提高最终模型的质量。

增强泛化能力

综合运用上述技术手段，全面提升模型的泛化能力和实际应用价值，确保模型在未知数据上的表现。

实验结果与影响

06

翻译任务表现

英德翻译

Transformer在英德翻译任务中，取得28.4 BLEU分数，超越同期最佳模型。

英法翻译

英法翻译中，BLEU分数高达41.8，同时显著提升训练速度。

综合表现

不仅在翻译精度上领先，还在语法分析等任务中展现卓越性能。

Transformer的影响

01

NLP领域革新

Transformer架构彻底改变了NLP领域，其高效并行处理能力加速了模型训练，推动了深度学习在语言理解上的进展。

02

后续模型涌现

基于Transformer，诞生了BERT、GPT系列、T5等模型，它们在问答、摘要、情感分析等任务上取得了显著成果。

03

跨领域应用

Transformer不仅限于NLP，还在计算机视觉、语音识别等领域展现出强大潜力，促进了多模态学习的发展。

04

研究新方向

Transformer激发了对模型可解释性、效率优化等开放问题的深入探索，引领了AI研究的新趋势。

后续发展与应用


Transformer影响
NLP领域革新
启发BERT模型，提升文本理解能力。
推动GPT系列发展，实现高质量文本生成。
促进T5模型创新，增强多任务处理能力。
CV领域潜力
在图像分类任务中展现高效性能。
应用于目标检测，提高识别精度。
语音识别
改善语音转文字的准确率。
支持多语言语音识别，拓宽应用范围。
可解释性研究
探索模型内部机制，提高透明度。
开发可视化工具，辅助理解模型决策过程。
效率优化
减少计算资源消耗，加快训练速度。
优化模型架构，降低推理延迟。
研究热点
探索新的应用场景，拓展技术边界。
结合其他技术，如强化学习，创造更多可能性。

THANKS

李智浩 2025/4/21