0001年1月1日 3 分钟阅读

Transformer 模型学习指南 I. 复习大纲

引言 •序列转换模型的局限性（循环神经网络和卷积神经网络）。 •Transformer 模型的提出：完全基于注意力机制，摒弃循环和卷积。 •Transformer 模型的优点：并行化能力强，训练时间短，翻译质量高。 •Transformer 模型在机器翻译和英语成分句法分析上的成功应用。

背景 •减少序列计算的必要性。 •卷积神经网络模型（Extended Neural GPU, ByteNet, ConvS2S）的并行计算方式及其局限性。 •自注意力机制的定义和应用。 •Transformer 模型与其他模型的区别和优势。

模型架构 •3.1 编码器和解码器堆栈编码器： •N=6 个相同的层堆叠而成。 •每一层包含两个子层：多头自注意力机制和位置式全连接前馈网络。 •残差连接和层归一化。 •所有子层和嵌入层的输出维度 dmodel = 512。 •解码器： •N=6 个相同的层堆叠而成。 •每一层包含三个子层：多头自注意力机制，编码器输出的多头注意力机制和位置式全连接前馈网络。 •残差连接和层归一化。 •掩码机制防止解码器关注后续位置。 •3.2 注意力机制定义：将查询（query）和键值对（key-value pairs）映射到输出的函数。 •输出是值的加权和，权重由查询与对应键的兼容性函数计算。 •3.2.1 缩放点积注意力（Scaled Dot-Product Attention）：计算查询和所有键的点积，除以 √dk，应用 softmax 函数得到权重。 •公式：Attention(Q, K, V) = softmax(QKT/√dk)V •与加性注意力（additive attention）的比较。 •3.2.2 多头注意力（Multi-Head Attention）：将查询、键和值线性投影 h 次到不同的 dk、dk 和 dv 维度。 •在每个投影版本上并行执行注意力函数。 •将输出连接并再次投影得到最终值。 •公式：MultiHead(Q, K, V) = Concat(head1, ..., headh)WO •head_i = Attention(QWQ_i, KWK_i, VWV_i) •优点：允许模型共同关注来自不同表示子空间的信息。 •3.2.3 模型中注意力的应用：编码器-解码器注意力层：查询来自先前的解码器层，键和值来自编码器的输出。 •编码器自注意力层：键、值和查询都来自同一位置，即编码器前一层的输出。 •解码器自注意力层：允许解码器中的每个位置关注到当前位置以及之前的所有位置，防止信息向左流动，保持自回归特性。 •3.3 位置式前馈网络定义：应用于每个位置的全连接前馈网络，包含两个线性变换和一个 ReLU 激活函数。 •公式：FFN(x) = max(0, xW1 + b1)W2 + b2 •3.4 嵌入和 Softmax使用学习到的嵌入将输入和输出标记转换为 dmodel 维度的向量。 •使用线性变换和 softmax 函数将解码器输出转换为预测的下一个标记概率。 •共享嵌入层和预 Softmax 线性变换的权重矩阵，嵌入层乘以 √dmodel。 •3.5 位置编码为了让模型利用序列的顺序信息，添加位

目录

文章信息

字数

阅读时间

发布时间

置编码到输入嵌入。 •位置编码与嵌入具有相同的维度 dmodel，可以相加。 •使用不同频率的正弦和余弦函数。 •公式：PE(pos, 2i) = sin(pos/10000^(2i/dmodel)) •PE(pos, 2i+1) = cos(pos/10000^(2i/dmodel)) •使用正弦函数的原因：允许模型轻松学习相对位置的注意力。

为什么使用自注意力 •比较自注意力层与循环和卷积层的各个方面。 •三个主要考虑因素： •每层的总计算复杂度。 •并行计算量（最小连续操作数）。 •网络中远距离依赖关系之间的路径长度。 •自注意力的优势： •以恒定数量的连续执行操作连接所有位置，而循环层需要 O(n) 个连续操作。 •当序列长度 n 小于表示维度 d 时，自注意力层比循环层更快。 •可以使用限制自注意力来提高计算性能，但会增加最大路径长度。 •卷积层需要堆叠多层才能连接所有输入和输出位置，增加路径长度。 •自注意力可以产生更易于解释的模型。

训练 •5.1 训练数据和批处理：WMT 2014 英语-德语数据集 (4.5 百万句子对)，使用 byte-pair encoding (BPE)。 •WMT 2014 英语-法语数据集 (36 百万句子对)，使用 word-piece。 •按近似序列长度将句子对批处理在一起。 •每个训练批次包含约 25000 个源标记和 25000 个目标标记。 •5.2 硬件和时间安排：8 个 NVIDIA P100 GPU。 •基础模型：每个训练步骤约 0.4 秒，训练 100,000 步（12 小时）。 •大型模型：每个训练步骤 1.0 秒，训练 300,000 步（3.5 天）。 •5.3 优化器： Adam 优化器： β1 = 0.9, β2 = 0.98, ϵ = 10^-9。 •学习率随训练过程变化。 •公式：lrate = d_model^-0.5 *min(step_num^-0.5, step_num* warmup_steps^-1.5) •先线性增加学习率，然后按步数的反平方根成比例地降低学习率。 •warmup_steps = 4000。 •5.4 正则化：残差 Dropout：应用于每个子层的输出，在添加到子层输入和归一化之前。 •嵌入 Dropout：应用于编码器和解码器堆栈中嵌入和位置编码的总和。 •标签平滑（Label Smoothing）：ϵ_ls = 0.1，提高准确性和 BLEU 分数。

结果 •6.1 机器翻译：在 WMT 2014 英语-德语翻译任务中，大型 Transformer 模型优于以前最好的模型，创造了新的最先进 BLEU 分数 28.4。 •即使是基础模型也超过了所有先前发布的模型，并且训练成本仅为竞争模型的一小部分。 •在 WMT 2014 英语-法语翻译任务中，大型 Transformer 模型实现了 41.0 的 BLEU 分数，优于所有先前发布的单个模型。 •6.2 模型变体：改变注意力头的数量，键和值的维度，保持计算量恒定。 •减少注意力键的大小会损害模型质量。 •更大的模型更好，Dropout 非常有助于避免过拟合。 •用学习的位置嵌入代替正弦位置编码，结果几乎与基本模型相同。 •6.3 英语成分句法分析：Transformer 可以推广到其他任务。 •在 Penn Treebank 的华尔街日报（WSJ）部分（约 4 万个训练句子）上训练了一个 4 层的 Transformer。 •半监督设置：使用来自高置信度和 BerkleyParser 语料库的约 1700 万个句子的更大语料库。 •结果表明，即使缺乏特定于任务的调整，该模型也能表现出色，优于所有先前报告的模型，除了循环神经网络语法 [8] 之外。

结论 •Transformer 是第一个完全基于注意力的序列转换模型。 •对于翻译任务，Transformer 的训练速度比基于循环或卷积层的架构快得多。 •在 WMT 2014 英语-德语和 WMT 2014 英语-法语翻译任务中，都达到了新的最先进水平。 II. 小测验

Transformer模型的核心思想是什么？它与RNN和CNN模型有何不同？ 完全基于注意力机制，摒弃循环和卷积。RNN和CNN存在序列依赖，并行化能力受限，远距离依赖捕获能力较弱。Transformer通过自注意力机制实现了高度并行化，并能有效捕获长程依赖关系。

请简述Transformer模型中的编码器和解码器的结构。 编码器由N个相同的层堆叠而成，每层包含多头自注意力机制和位置式前馈网络，并采用残差连接和层归一化。解码器也由N个相同的层堆叠而成，除了编码器的两个子层外，还包含一个编码器输出的多头注意力机制，同样采用残差连接和层归一化。

请解释缩放点积注意力机制的作用，并说明为什么要进行缩放。 缩放点积注意力机制用于计算query与keys之间的相关性，并通过softmax函数得到每个value的权重，从而实现对输入序列不同部分的关注。缩放的目的是防止点积过大导致softmax梯度消失，影响模型的学习效果。

多头注意力机制的优势是什么？如何实现多头注意力？ 多头注意力机制允许模型从不同的表示子空间学习信息，从而更全面地理解输入序列。实现方法是将query、key、value线性投影到多个不同的子空间，分别进行注意力计算，然后将结果拼接并投影到输出空间。

Transformer模型中位置编码的作用是什么？有哪些常用的位置编码方式？ 位置编码用于向模型提供序列中token的位置信息，因为自注意力机制本身不具备序列顺序感知能力。常用的位置编码方式包括正弦和余弦函数编码，以及学习到的位置嵌入。

Transformer模型的自注意力机制在编码器和解码器中分别是如何应用的？ 编码器中的自注意力机制允许每个位置关注到编码器前一层的任何位置，从而学习序列内部的依赖关系。解码器中的自注意力机制在训练时需要进行掩码，防止每个位置关注到未来的信息，保证自回归特性。

请解释Transformer模型中残差连接和层归一化的作用。 残差连接用于缓解深度网络中的梯度消失问题，提高模型的训练效果。层归一化用于稳定网络的训练过程，加快收敛速度，提高模型的泛化能力。

Transformer模型中的前馈网络是如何设计的？它有什么作用？ Transformer模型中的前馈网络是一个两层的全连接网络，中间采用ReLU激活函数。它对每个位置的向量进行独立的非线性变换，增强模型的表达能力。

在Transformer模型的训练过程中，使用了哪些正则化技术？它们的作用是什么？ 在Transformer模型的训练过程中，使用了残差Dropout、嵌入Dropout和标签平滑等正则化技术。残差Dropout和嵌入Dropout用于防止模型过拟合，标签平滑用于提高模型的泛化能力和准确性。

Transformer模型在机器翻译任务中取得了哪些成果？ Transformer模型在WMT 2014英语-德语和英语-法语翻译任务中都取得了当时的领先成果。 III. 答案 Key 1.Transformer模型的核心思想是什么？它与RNN和CNN模型有何不同？ Transformer 模型的核心思想是完全依赖于注意力机制来处理序列转换任务，摒弃了传统的循环神经网络 (RNN) 和卷积神经网络 (CNN)。与 RNN 的序列依赖和 CNN 的局部感受野不同，Transformer 通过自注意力机制实现了高度并行化，并且能够有效地捕获长程依赖关系。 2.请简述 Transformer 模型中的编码器和解码器的结构。 编码器由 N 个相同的层堆叠而成，每层包含一个多头自注意力子层和一个位置式全连接前馈网络子层，每个子层后都跟随着残差连接和层归一化。解码器与编码器类似，也由 N 个相同的层堆叠而成，但每个层额外增加了一个多头注意力子层，用于关注编码器的输出，同样也使用了残差连接和层归一化。 3.请解释缩放点积注意力机制的作用，并说明为什么要进行缩放。 缩放点积注意力机制用于计算输入序列中不同位置之间的关系，从而为每个位置生成一个加权表示。缩放（除以 √dk）是为了防止点积过大，导致 softmax 函数的梯度过小，从而影响模型的学习效果，尤其是在键的维度 dk 较大时。 4.多头注意力机制的优势是什么？如何实现多头注意力？ 多头注意力机制的优势在于它允许模型从不同的表示子空间学习信息，从而更全面地理解输入序列。实现方法是将 query、key、value 通过不同的线性变换投影到多个不同的子空间，分别计算注意力，然后将各个头的结果拼接并再次线性变换到输出空间。 5.Transformer 模型中位置编码的作用是什么？有哪些常用的位置编码方式？ 位置编码的作用是向模型提供序列中 token 的位置信息，因为自注意力机制本身不具备序列顺序感知能力。常用的位置编码方式包括正弦和余弦函数编码，以及学习到的位置嵌入。 6.Transformer 模型的自注意力机制在编码器和解码器中分别是如何应用的？ 在编码器中，自注意力机制允许每个位置关注到编码器前一层的任何位置，从而学习序列内部的依赖关系。在解码器中，自注意力机制在训练时需要进行掩码（masking），防止每个位置关注到未来的信息，以保证自回归特性，从而进行正确的序列生成。 7.请解释 Transformer 模型中残差连接和层归一化的作用。 残差连接的作用是缓解深度网络中的梯度消失问题，使得更深的网络更容易训练。层归一化的作用是稳定网络的训练过程，加快收敛速度，并提高模型的泛化能力，使其在未见过的数据上也能表现良好。 8.Transformer 模型中的前馈网络是如何设计的？它有什么作用？ Transformer 模型中的前馈网络是一个两层的全连接网络，中间采用 ReLU 激活函数。它的作用是对每个位置的向量进行独立的非线性变换，增强模型的表达能力，从而更好地捕捉输入序列中的复杂模式。 9.在 Transformer 模型的训练

过程中，使用了哪些正则化技术？它们的作用是什么？ Transformer 模型在训练过程中使用了残差 Dropout、嵌入 Dropout 和标签平滑等正则化技术。Dropout 用于防止模型过拟合，通过随机丢弃一部分神经元来减少模型对特定训练样本的依赖。标签平滑则通过对目标概率分布进行平滑来提高模型的泛化能力。 10.Transformer 模型在机器翻译任务中取得了哪些成果？ Transformer 模型在机器翻译任务中取得了突破性进展，在 WMT 2014 英语-德语和英语-法语翻译任务中都取得了当时最先进的 (state-of-the-art) 成果。它不仅在翻译质量上超越了之前的模型，而且在训练速度上也大幅提升。 IV. 论文格式问题 1.请讨论Transformer架构相较于循环神经网络（RNN）或卷积神经网络（CNN）在并行计算能力方面的优势。 2.请阐述Transformer模型中的多头注意力机制如何提升模型性能，并分析其与单头注意力机制的差异。 3.请分析位置编码在Transformer模型中的作用，并比较正弦位置编码与学习型位置编码的优缺点。 4.Transformer模型在英语成分句法分析中的应用体现了其怎样的泛化能力？ 5.请讨论Transformer模型在机器翻译任务中取得成功的关键因素，并展望基于注意力机制的模型在未来的发展方向。 V. 关键术语词汇表 •Sequence Transduction Model (序列转换模型): 一种将一个序列转换为另一个序列的模型，如机器翻译、语音识别等。 •Recurrent Neural Network (RNN, 循环神经网络): 一种处理序列数据的神经网络，通过循环连接处理时序信息。 •Convolutional Neural Network (CNN, 卷积神经网络): 一种主要用于处理图像数据的神经网络，通过卷积操作提取特征。 •Attention Mechanism (注意力机制): 一种使模型能够关注输入序列不同部分的技术，通过权重分配突出重要信息。 •Self-Attention (自注意力): 一种注意力机制，允许序列中的每个位置关注到序列中的所有其他位置。 •Multi-Head Attention (多头注意力): 一种使用多个注意力头并行计算的注意力机制，每个头关注不同的表示子空间。 •Scaled Dot-Product Attention (缩放点积注意力): 一种计算注意力权重的方法，通过点积计算 query 和 key 之间的相似度，并进行缩放。 •Encoder (编码器): 一种将输入序列转换为中间表示的神经网络。 •Decoder (解码器): 一种将中间表示转换为输出序列的神经网络。 •Residual Connection (残差连接): 一种将层的输入直接添加到输出的技术，用于缓解梯度消失问题。 •Layer Normalization (层归一化): 一种对层的输入进行归一化的技术，用于加速训练和提高泛化能力。 •Positional Encoding (位置编码): 一种向模型提供序列中位置信息的技术，因为注意力机制本身不感知顺序。 •Byte-Pair Encoding (BPE, 字节对编码): 一种用于将文本分割成子词单元的技术，用于处理未知词。 •Word-Piece: 与 Byte-Pair Encoding 相似的 subword 分词算法。 •Adam Optimizer (Adam 优化器): 一种自适应学习率的优化算法。 •Dropout (丢弃法): 一种通过随机丢弃神经元来防止过拟合的正则化技术。 •Label Smoothing (标签平滑): 一种通过平滑目标概率分布来提高泛化能力的正则化技术。 •BLEU Score: 双语评估辅助工具 (Bilingual Evaluation Understudy)，一种用于评估机器翻译质量的

指标。 •Constituency Parsing: 成分句法分析。 •Warmup Steps: 在训练初期线性增加学习率的步数。 •Beam Search: 一种在序列生成任务中使用的搜索算法，用于找到最优的输出序列。

分享这篇文章