

第 5 章

注意力机制与 Transformer

Transformer 架构是深度学习发展历程中最重要的技术突破之一^①。

——杰弗里·辛顿(Geoffrey Hinton, 深度学习著名学者)

在大模型 GPT(Generative Pretrained Transformer)中,“T”就是指 Transformer,而 Transformer 的核心正是注意力机制。本章将重点讨论注意力机制的基本原理及其在 Transformer 架构中的应用。

Transformer 是一种深度学习模型架构,通过注意力机制解决了传统序列模型在长序列任务中信息压缩的缺陷。其高效性和灵活性使其成为诸如 GPT、BERT 以及 DeepSeek 等大规模预训练模型的基础设计。

以 GPT 为例,该模型利用 Transformer 架构,通过在大量文本数据上进行预训练,学会了如何生成连贯且富有逻辑的自然语言文本。而 DeepSeek 则利用类似的机制,在信息检索和语义理解方面表现出色,能够更准确地捕捉和利用数据间的复杂关联。

接下来,我们将从注意力机制入手,逐步解析 Transformer 的结构和工作原理。

5.1 注意力机制

神经网络中的注意力机制常被视为对人类选择性注意过程的一种抽象与模拟。为了更好地理解神经网络中的注意力机制,我们可以先回顾一下人类注意力机制是如何工作的。

^① 这是杰弗里·辛顿于 2021 年的访谈观点,对应的英文为: The transformer architecture is one of the most important developments in deep learning.

在人类的日常认知中,注意力是一种有限的资源。我们不可能同时关注所有的信息,而是会根据任务的需求,选择性地将注意力集中在最相关的部分。例如,当我们阅读一段文字时,虽然眼睛会扫过每一个单词,但真正被大脑处理的,是那些与当前理解或决策最相关的内容。

这种选择性关注的能力帮助人类高效地处理复杂的信息,避免信息过载,同时确保关键细节不会被忽略。同样,人工神经网络中的注意力机制旨在模拟这一特性,通过动态调整权重,选择性地关注输入数据的不同部分,从而提升模型的性能和效率。

5.1.1 关于注意力的一个著名实验

美国心理学家**丹尼尔·西蒙斯(Daniel Simons)**曾设计并进行过一个著名的实验——“**看不见的大猩猩**”(Invisible Gorilla)^①。在这段 30 秒的视频中,6 名参与者在不停地走动、换位,同时传递两只红色的球。实验要求受试者观察,并回答球在白衣服女孩手中被传递了多少次^[1]。

常见视频版本的计数答案为 15 次。受试者中,有人答对了,也有人答错了。然而,这个实验的重点并不仅在于受试者是否答对了计数题目。真正令人关注的是,在视频中,一只身穿大猩猩装的大猩猩大摇大摆地走到舞台中央,猛捶自己胸部后,又大摇大摆地走出舞台。

令人惊讶的是,大约 50% 的受试者完全没有注意到这只出现在眼前的大猩猩(图 5-1)。为什么会出现这种情况? 丹尼尔·西蒙斯将这一现象称为“选择性注意”。

“选择性注意”指出,人类在面对复杂的信息环境时,会有意识地将注意力集中在某些特定任务上,而忽略其他看似显而易见的信息。这一实验生动地说明了人类注意力的有限性以及对目标的集中性。

确实,在某些情况下,人们会出现“视而不见,听而不闻”的现象。乍看之下,这似乎是一种缺陷,但实际上却蕴含着重要的生存意义。要知道,这个世界的信息总量是无限制的,而我们的感官带宽却是有限的,正所谓“以有涯逐无涯,殆已”。因此,人类无法感知和存储“整个时间”轴上接触到的“全部空间”中的信息。

换句话说,注意力是一种有限的资源,正因为其稀缺,才显得弥足珍贵,不应随意“施

① 互联网上(如 B 站)可找到该视频,读者可自行测试一下。



图 5-1 看不见的大猩猩(注意力测试)

舍”。人类通过注意力机制,选择性地聚焦于目标区域或当前任务,投入更多认知资源以获取细节信息,同时抑制无关信息的干扰。

这种机制并非缺陷,而是人类长期进化过程中形成的一种适应性策略。它帮助我们有效地处理环境中的关键信息,从而在复杂多变的世界中做出快速而准确的决策。这一机制不仅提高了人类的生存能力,也增强了对外界环境的适应能力,是人类智慧的重要表现之一。

5.1.2 人类视觉系统的启示

通俗来说,“注意力”的核心逻辑非常贴合其名称:从关注全部到重点关注局部的关键部分。在人脑的注意力机制中,大脑只选择一些关键信息进行处理,以提高神经网络的效率并节省认知资源。根据认知神经学的研究,注意力总体上可以分为以下两种主要类型。

1. 聚焦式注意力

聚焦式注意力(Focused Attention)是一种高度集中的注意力形式,通常用于处理复杂或具有挑战性的任务。在这种模式下,大脑会将全部或大部分注意力资源集中到单一的信息源或任务上,主动忽略无关的干扰。

聚焦式注意力需要较高的认知控制,能够帮助个体在高需求任务中保持专注。例如,当你阅读一篇难懂的学术论文时,大脑会屏蔽掉背景噪声或其他干扰,全力理解内容。

2. 显著性注意力

显著性注意力(Saliency-Based Attention)是一种自动化和无意识的注意力分配方

式。在这种模式下,大脑会优先关注环境中最显眼或最突出的刺激,而无须主动控制。例如,当你进入一个陌生房间时,你的注意力可能会立即被明亮的灯光或快速移动的物体吸引。

显著性注意力的机制在人工神经网络中也有类似的表现。例如,在卷积神经网络(CNN)中,最大池化(Max-Pooling)的过程可以被视为一种基于显著性的注意力机制,它帮助模型筛选出特征图中最显著的特征;在 LSTM 网络中,门控机制(Gating Mechanism)类似于显著性注意力,通过筛选输入、遗忘和输出的信息,优先保留最相关的内容。

比如,在一个吵闹的酒会上,一个人和朋友聊天时,尽管周围充满了噪声干扰,他仍然可以轻松听到朋友的谈话内容,而忽略其他人的声音。这是因为他利用了聚焦式注意力,主动将注意力集中在朋友的谈话上。然而,如果背景噪声中突然出现一个重要的词汇(例如,有人在人群中呼喊他的名字),他会立刻转移注意力,注意到这一刺激。这种现象则是显著性注意力的表现(图 5-2)。简而言之,聚焦式注意力通常是有意识地控制的,而显著性注意力更多是由环境中的显著刺激触发的。



图 5-2 显著性注意力

脑神经科学与人工智能科学正彼此促进、协同进化。人类的注意力机制显然启发了人工神经网络的设计。在人工神经网络中,借鉴自认知神经学的注意力机制框架,让模型在处理输入信息时不再平等地对待所有输入,而是根据每个输入的重要性分配不同的关注度(Attention)。这种选择性关注的机制显著提高了模型处理复杂任务的能力,尤其

在自然语言处理、图像识别等领域展现了卓越的效果。

5.1.3 神经网络中的注意力机制

从前面章节的讨论可以看出,在传统的 Seq2Seq 模型中(图 5-3),一旦编码器完成了对输入序列的编码,并将其最后的隐藏状态 h_m 交付给解码器,编码器的任务便告一段落,相当于“完成使命后休假”。

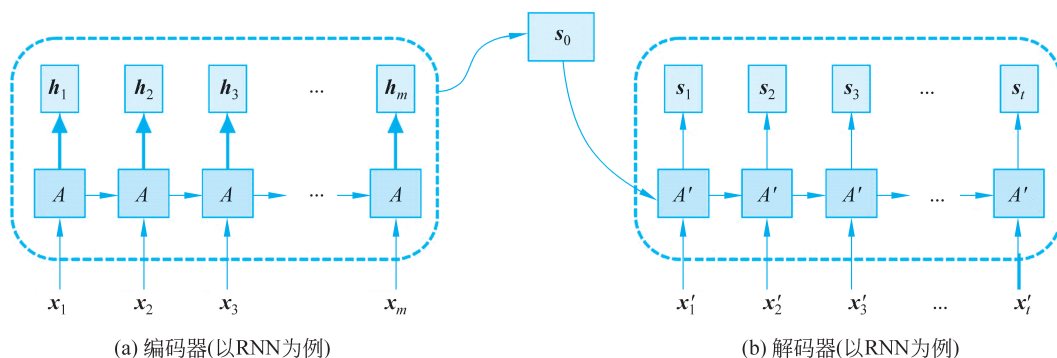


图 5-3 传统的 Seq2Seq 模型

解码器接收到编码器输出的隐藏状态 h_m , 并将其作为初始内部状态 s_0 。随后, 解码器开始逐步生成输出序列。此时, 解码器的工作模式类似于一个“文本生成器”, 基于当前的内部状态和已生成的部分输出(如先前生成的字符或单词), 逐字(或逐词)生成输出序列。

在这一阶段, 解码器的运行过程就像是一台“打字机”, 根据自身的状态通过自回归的方式, 逐一输出字符, 直到生成完整的目标序列。这种方式虽然能够实现从输入到输出的序列生成, 但由于编码器输出的隐藏状态 h_m 是一个固定长度的向量, 传统 Seq2Seq 模型在处理长序列时可能会丢失部分关键信息, 影响生成质量。

带有注意力机制的 Seq2Seq 模型与传统版本有着显著不同。在注意力机制的协调下, 解码器在生成输出序列时, 编码器不再“休假”, 而是需要持续配合工作。

解码器在输出序列的过程中, 会“不忘初心”, 总忍不住“回眸看一眼”编码器。这是因为每生成一个目标词元(Token), 解码器需要动态计算其内部状态 s 与编码器输出的隐藏状态 h 的相关性(图 5-4)。

相关性就是广义的注意力, 它使解码器在生成输出时能够有所侧重, 而不是平均对

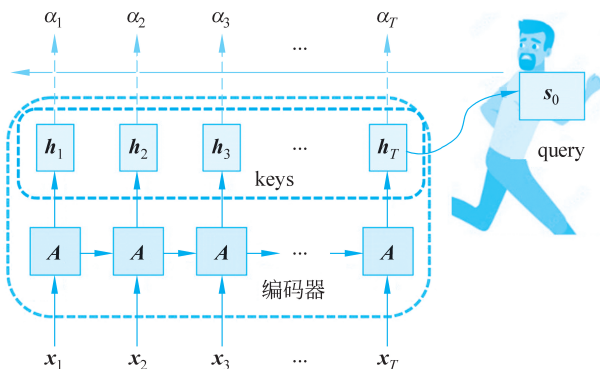


图 5-4 带有注意力机制的 Seq2Seq 模型

待所有输入状态。

假设解码器的初始状态是 s_0 (实际上,它也是编码器的最后一个隐藏状态 h_m),解码器需要根据其当前状态 s 与编码器在每个时间步的隐藏状态 $h_i (i=1,2,\dots,m)$ 之间的相关性,决定对输入序列中哪些部分应给予更多关注。

我们可以通过式(5-1)来表示这种相关性,计算解码器当前状态 s 与编码器状态 h_i 的关联程度^[2]。

$$\tilde{\alpha}_i = \text{align}(h_i, s) \quad (5-1)$$

式中,来自解码器的 s 称为查询(Query),来自编码器的一系列隐藏状态 h_i 序列可视为关键字(Keys)。align 是一个广义变换函数,它用于计算 Query 和 Keys 之间的相关性。相关性的结果可以视为一种权重分布,用于衡量输入序列中每个时间步的重要性。

计算 Query(解码器当前状态 s)与 Keys(编码器中的各个 h_i)之间的相关性有多种方法。在编码器中,隐藏状态 h_i 的数量共有 m 个。通过相关性计算的流程,可以得到 $\tilde{\alpha}_1 \sim \tilde{\alpha}_m$ 这 m 个权值。这些权值分数可大可小、可正可负,它们是相似度的 logits,这时我们再对它们做一个 softmax 变换,把输出结果记为 $\alpha_1 \sim \alpha_m$ 。

$$[\alpha_1, \alpha_2, \dots, \alpha_m] = \text{softmax}([\tilde{\alpha}_1, \tilde{\alpha}_2, \dots, \tilde{\alpha}_m]) \quad (5-2)$$

由于 α_i 是 softmax 的输出, $\alpha_1 \sim \alpha_m$ 都是大于 0,且相加之和等于 1。

下面我们用一个感性的例子来说明注意力机制是如何工作的。以翻译“知识就是力量”这个任务为例,解码器在准备翻译某个词时,会根据相关的词语再“回眸”一眼编码

器,从源头上获知当前的输出应对哪些词给予更多的关注(其实是在计算一种相似性)。

比如说,在翻译 Knowledge 时,解码器通过向编码器查询(Query),根据向量相似性,得知这个单词与“知识”更加密切相关,于是对“知”和“识”这两个输入向量关注权值要大一些。

类似地,在翻译 is 时,解码器更加关注“就”和“是”这两个输入向量。在图 5-5 所示的这个例子中,连接线的粗细表征了注意力的强度,直观地显示了不同单词之间的注意力分配。那么如何实现这个“更关注”这个特性呢?这自然是基于海量数据需要训练和学习。

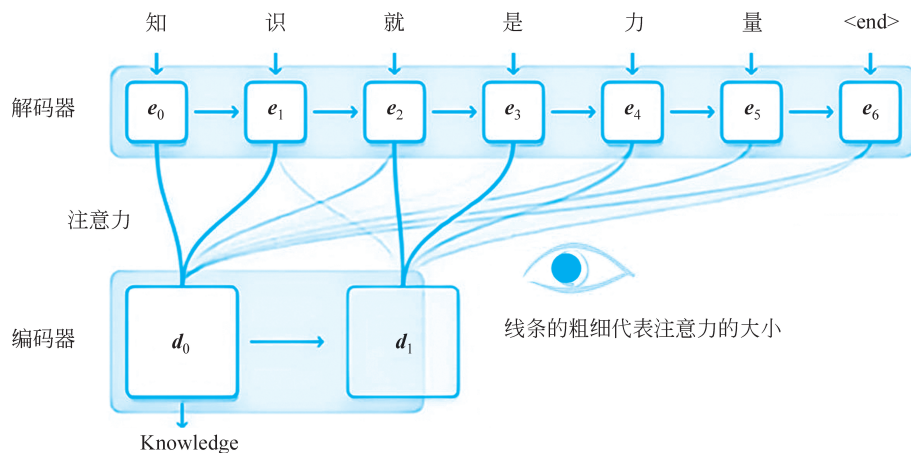


图 5-5 注意力机制的工作过程

在引入注意力机制之前,传统模型难以有效捕捉长距离依赖关系,语义信息在传递中容易丢失。这类似于记忆力较弱的人难以记起遥远的往事。而注意力机制则是一种有针对性的选择性记忆方法,即使面对很长的文本,也能准确定位和提取关键内容,从而避免丢失重要信息。

如图 5-6 所示,性能指标采用的是 BLEU 评分(Bilingual Evaluation Understudy,跨语言评估评测),这一指标由 IBM 公司提出,用于评价候选文本与参考翻译之间的相似度。当句子长度超过 20 个单词,在没有“注意力”机制的 Seq2Seq 模型中,BLEU 评分迅速下降。而加入“注意力”机制后,模型在处理长句子时,BLEU 评分能维持在较高水平。这表明,注意力机制能够有效捕捉长距离依赖,从而提升模型在长句子翻译任务中的表现。

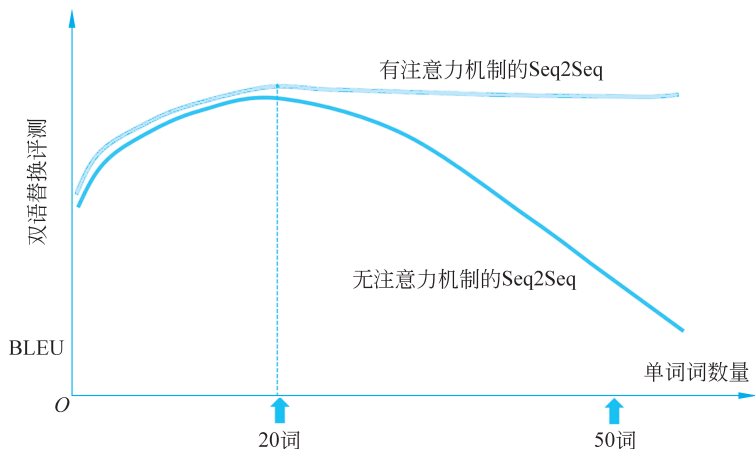


图 5-6 有无 Attention 机制的性能对比

5.2 Transformer 模型

注意力机制是 Transformer 的重要理论基础。如今,Transformer 模型基本已全面取代了传统的循环神经网络(RNN)和卷积神经网络(CNN),成为自然语言处理(NLP)的标配架构,也是众多大模型(如 GPT、Llama 和 DeepSeek 等)的核心架构。

在深入探讨 Transformer 的创新点之前,有必要先了解这一模型诞生的背景,以及它为何能够成为深度学习领域的一次革命性突破。

5.2.1 注意力机制的二次发明

著名经济学家布莱恩·阿瑟(Brian Arthur)在其著作《技术的本质》中指出^[3]: 所谓新技术实际上只是对之前已有技术的重新组合。发明的核心在于“挪用”,一种自觉或不自觉的借鉴行为。熊彼特(Joseph Schumpeter) 也持有类似观点,他认为创新就是“新的组合”,即通过将看似毫不相关的事物联系起来,以出人意料的方式重新呈现给人们。

在前面的章节中,我们已简单讨论了注意力机制。然而,真正让注意力机制“出圈”的,是后起之秀——由阿西什·瓦斯瓦尼(Ashish Vaswani)等在 2017 年发表的论文《注意力就是你所需要的一切》(“Attention is All You Need”)^[4]。这篇论文的发表不仅奠定了 Transformer 模型的理论基础,更开启了以“××× is all you need”为标题的研究热

潮。这种标题风格或许也成了 AI 研究领域特有的幽默和文化符号。

一篇论文之所以能够成为经典，并不一定在于它原创性地提出了什么全新的概念，更可能在于它巧妙地整合了已有技术，从而解决了当下许多人关注的棘手问题。这种层次的突破往往令原始贡献者的名字逐渐被人淡忘，但解决问题的模型和方法却被世人铭记。

例如，在人工智能的发展历程中，有许多这样的里程碑：杰弗里·辛顿 (Geoffrey Hinton) 发扬光大的 BP 反向传播算法，亚历克斯·克里泽夫斯基 (Alex Krizhevsky) 等在 AlexNet 中使用的 Dropout 技术，何恺明 (Kaiming He) 等在 ResNet 中提出的“残差” (Residual) 思想，以及本节重点讨论的 Transformer 模型中的注意力机制。这些成果无一不是站在前人研究的基础上，通过改进与组合优化，实现了关键的创新与突破。

Transformer 的作者瓦斯瓦尼等也坦言，自注意力机制 (Self-Attention) 的思想并非他们的首创。他们的贡献在于将这一机制应用于序列建模任务，并通过 Transformer 架构对其进行系统化的优化。这种模式的意义，就如同爱迪生发明电灯。电灯的第一个发明人并非爱迪生，但爱迪生的贡献在于让电灯走出实验室，“飞入寻常百姓家”。

类似地，Transformer 的出现不仅为自然语言处理 (NLP) 领域带来了革命性的变化，更让全球 NLP 社群将目光聚焦到注意力机制的潜力上，从而开启了新的研究热潮，产生了巨大的“注意力”。

5.2.2 Transformer 的诞生背景

在没有 Transformer 模型之前，机器翻译工作主要依赖于基于循环神经网络 (RNN) 的“编码器-解码器”模型。然而，RNN 存在固有的缺陷，其根本原因在于其循环连接的方式，所有的信息都作用在同一个权值矩阵上。如果权值矩阵最大特征值小于 1，那么这个矩阵就容易出现梯度消失问题，而梯度是神经网络训练过程中至关重要的“指南针”，一旦梯度消失，训练过程将失去方向，模型无法有效学习。

尽管后来引入了 RNN 的改良版本，例如长短期记忆网络 (LSTM) 和门控循环单元 (GRU)，这些变体通过记忆单元和门控机制在一定程度上缓解了梯度消失问题，但它们并未从根本上解决这一难题。RNN 及其变体在捕捉长距离依赖关系时，始终表现得“踉踉跄跄，勉为其难”。

Transformer 模型诞生的第二个背景是 RNN 及其变体的计算效率问题。RNN 为

了记住过往信息,不得不依赖其循环结构。在这种结构中,计算第 t 时刻的输出需要依赖第 $t-1$ 时刻的数据,而第 $t-1$ 时刻的数据又需要第 $t-2$ 时刻的数据支撑。环环相扣的依赖关系导致 RNN 无法实现并行计算,从而极大地限制了模型的训练速度和可扩展性。

为了解决 RNN 的这些局限性,人们开始尝试使用 CNN 进行序列处理。CNN 擅长捕捉局部依赖关系,并且具有较高的并行计算效率。然而,CNN 的劣势在于其对长距离依赖关系的建模能力较弱,而这种能力对于许多自然语言处理任务至关重要。

正是由于梯度消失问题对长距离依赖建模的限制,研究者们逐渐意识到,依赖循环连接的网络架构存在先天不足。这种困境成为 Transformer 模型诞生的关键背景之一。Transformer 的设计通过完全抛弃循环结构,依赖注意力机制直接建模序列中任意位置之间的关系,彻底摆脱了传统 RNN 的限制,从而实现了对长距离依赖的高效建模。

为进一步提升神经网络在捕捉长距离依赖关系上的性能,人们提出了注意力机制(Attention Mechanism)。注意力机制通过动态分配权重的方式,使网络能够灵活地关注序列中的任意部分,从而克服了 RNN 和 CNN 的局限。然而,在处理复杂任务时,单一的注意力机制往往不足以捕捉多种类型或层次的依赖关系。多头注意力机制(Multi-Head Attention)的提出,解决了这一问题。

因此,多头注意力机制的引入,成为 Transformer 模型诞生的第三个背景。这一机制通过并行计算多个不同的注意力分布,使模型能够同时关注到序列中的不同特征和层次,极大地增强了模型的表达能力,并奠定了 Transformer 成为 NLP 领域革命性模型的基础。

5.2.3 Transformer 架构

Transformer 模型能够捕捉序列中任意位置之间的关系,而无须像传统 RNN 那样逐个处理序列中的每一个元素。这种机制使 Transformer 能够并行处理整个序列,大幅提升了训练效率和计算性能。Transformer 的结构主要由两个核心部分组成:编码器(Encoder)和解码器(Decoder),如图 5-7 所示。

Transformer 模型采用编码器-解码器(Encoder-Decoder)结构,其中编码器(左侧)负责将输入序列转换为连续的上下文表示,解码器(右侧)依据编码器输出生成最终结果。