第 3 章 Transformer 在自然语言处理中的应用

3.1 编码器

在 NLP 领域，Transformer 架构凭借其卓越的性能和高效的并行计算能力，已经成为深度学习模型的基石。Transformer 的编码器在众多 NLP 任务中扮演着至关重要的角色，以 BERT 为代表的模型，便在 NLP 领域取得了革命性的突破。

3.1.1 BERT

BERT 与基于 LSTM 或 Transformer 解码器的模型（如 ELMo[30] 和 GPT）有显著的不同，主要体现在其采用了 Transformer 编码器架构及深度双向的上下文建模方式，这种双向建模方式突破了传统单向模型的局限，显著地提升了语言表示的效果，如图 3-1 所示。与传统的单向模型只能从前向后或从后向前处理文本不同，BERT 能够同时利用前后的上下文信息。例如，在处理句子“我今天很开心”时，传统的单向模型可能只能从“我”开始逐词向后处理，或者从“很开心”开始逐词向前处理，而无法同时考虑前后词语的关联；而 BERT 通过引入 MLM 任务，实现了深度双向的上下文建模能力。在这种任务中，模型会随机掩盖输入序列中的部分词汇，要求模型根据上下文预测这些被掩盖的词汇。这使模型在预训练阶段就能够学习到前后文之间的复杂关系，从而在后续任务中更好地理解语义。

图 3-1 从 ELMo/GPT 到 BERT 的架构转变示意图

在 BERT 中，输入文本会经过 3 种不同的嵌入处理：Token Embeddings、Segment Embeddings 和 Position Embeddings，见图 3-2。这些嵌入的信息被组合成一个向量表示，作为模型的输入。Token Embeddings 用于表示文本中的每个词或子词，帮助模型理解单

个词汇的语义。Segment Embeddings用于区分不同的句子(例如,句子A和句子B),在句子对任务中尤为重要。Position Embeddings则提供了每个词在序列中的位置,帮助模型理解词的顺序。这些嵌入通过加和的方式进行组合,形成最终的输入表示。

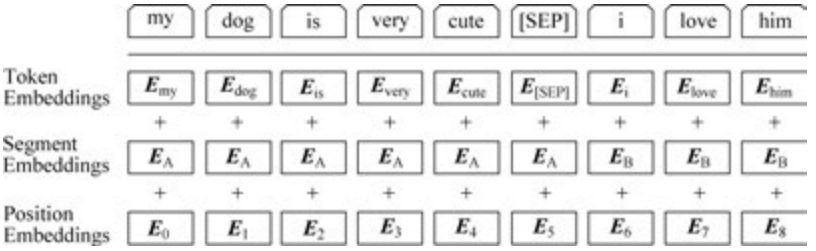

图3-2 BERT输入嵌入

BERT的每层都由多个Transformer编码器组成,每个编码器层都通过自注意力机制处理输入序列,并生成新的表示。随着层数的增加,模型能够捕捉到更高层次的语义信息,其预训练阶段采用了两项任务:MLM和NSP。MLM任务通过随机掩盖输入序列中的部分词汇,要求模型根据上下文预测这些被掩盖的词汇。具体而言,BERT会在输入文本中随机掩盖15%的词汇,并通过多头自注意力机制和FFN生成这些被掩盖词汇的预测。这种双向建模方式使BERT能够同时考虑文本中的前后文信息,从而更准确地理解词语的语义。例如,在句子"我喜欢吃[MASK]"中,模型需要根据上下文"我喜欢吃"来预测被掩盖的词汇,可能是"Pizza""水果"等,这取决于具体的上下文。

NSP任务则要求模型判断两个输入句子是否相邻,这对于捕捉句子间的关系至关重要,例如,在问答系统中,模型需要判断问题和答案是否相关。NSP任务通过在预训练阶段让模型学习句子对之间的语义关系,从而在下游任务中更好地理解文本的上下文,例如,给定两个句子:"你今天吃了什么?"和"我今天吃了Pizza。",模型需要判断这两个句子是否相邻,即第2个句子是否是第1个句子的回答。在预训练过程中,模型会接触到大量的句子对,并学习它们之间的关联性,从而在下游任务中能够更好地利用这种知识来处理文本。

BERT在下游任务中的微调策略展现出极高的灵活性,能够适应多种自然语言处理任务的需求。以文本分类任务为例,BERT通过其输出的[CLS]标记作为分类器的输入,这一设计充分利用了BERT在预训练阶段对整个句子语义的理解能力。[CLS]标记在BERT的架构中专门用于句子级别的表示,它在预训练过程中被训练来预测句子的分类,因此在微调阶段,可以直接将其输出,用于分类任务,如情感分析、新闻分类等。例如,在情感分析任务中,对于句子"这部电影真的太棒了"和"这部电影非常糟糕",通过微调BERT模型,[CLS]标记的输出可以分别表示积极和消极的情感分类。

在命名实体识别任务中,BERT的输出则被用于标注每个词的实体类别。具体来讲,BERT为每个输入词生成一个对应的输出表示,这些表示包含了丰富的语义信息和上下文依赖关系。通过对这些输出表示进一步地进行处理,如添加任务特定的全连接层,可以实现对每个词的实体类别标注,从而完成命名实体识别任务。例如,在句子"我今天和Andy去

了香港”中，模型能够识别出“Andy”是一个人名，“香港”是一个地名，并对它们进行相应的标注。

BERT 的微调过程通常包括以下几个关键步骤。首先，选择一个预训练好的模型，如 BERT-Base 或 BERT-Large，这取决于任务的复杂性和计算资源的可用性。以中文情感分析任务为例，如果数据量较大且计算资源充足，则可以选择使用 BERT-Large 模型进行微调；如果数据量较小或计算资源有限，则可以选择使用 BERT-Base 模型。接着，准备任务特定的数据集，并使用适当的分词器对数据进行预处理，以确保输入格式与 BERT 的预训练模型兼容。例如，在中文任务中，使用 BERT 的 WordPiece 分词器对文本进行分词。假设原始文本为“我今天很开心”，分词后可能为“我”“今天”“很”“开心”；或者根据模型的分词规则生成更细粒度的子词，然后定义一个适合任务的模型架构。例如，对于文本分类任务，可以使用 BertForSequenceClassification；对于命名实体识别任务，则可以使用 BertForTokenClassification。在文本分类任务中，模型会输出一个类别概率分布。例如，在情感分析中，输出可能是积极、消极或中性的概率。

BERT 的微调策略不仅在文本分类和命名实体识别任务中表现出色，还在问答系统、文本生成等任务中得到了广泛应用。在问答系统中，BERT 可以通过微调来识别给定段落中的答案。例如，在 SQuAD 数据集上进行微调，以实现阅读理解任务。具体来讲，在问答任务中，模型的输入包括问题和相关的段落文本，通过微调后的 BERT 模型能够定位段落中与问题对应的答案。例如，给定问题“谁是 *Harry Potter* 的作者?”和段落文本“*Harry Potter* 是由 Joanne Rowling 创作的奇幻小说系列”，微调后的 BERT 模型能够识别出“Joanne Rowling”是该问题的答案。这种能力使 BERT 在问答系统中表现出色，能够准确地从给定的文本中提取出正确的答案。

在文本生成任务中，虽然 BERT 本身主要用于理解和分类任务，但结合其他模型（如 GPT 系列）可以用于生成任务，如总结文本或创作文本。例如，在文本总结任务中，可以将 BERT 与 GPT 结合使用，先利用 BERT 模型理解原始文本的语义，然后使用 GPT 模型生成简洁的总结。具体来讲，BERT 模型可以提取出文本中的关键信息和主要观点，而 GPT 模型则可以根据这些信息生成流畅的总结文本。这种结合使用的方式充分发挥了 BERT 在语义理解方面的优势和 GPT 在文本生成方面的能力，从而实现高质量的文本总结。此外，在文本创作任务中，也可以采用类似的方法，利用 BERT 模型理解用户的需求和意图，然后使用 GPT 模型生成符合要求的文本内容。这些应用展示了 BERT 在 NLP 领域的强大能力和灵活性，使其成为当前 NLP 任务中不可或缺的模型之一。

3.1.2 BERT 的改进模型

BERT 虽然通过双向编码机制实现了上下文感知的语义建模，但是其技术架构在实践层面仍面临多重挑战。首先，BERT 的生成能力受限于双向注意力机制的设计逻辑：由于采用双向注意力机制导致自回归生成时的信息泄露问题，这使其在文本生成、对话系统等需要单向自回归预测的场景中的表现显著弱于基于单向注意力机制的 GPT 系列模型，其次，

BERT 的运算复杂度呈现二次方增长特性，当处理长文本序列时，显存占用和计算耗时问题尤为突出。据实验测算，在标准 BERT-Base 模型上处理 512 长度的输入序列时，在批次(Batch)大小为 1 的 FP32 精度下，单次前向推理需要约 1.7GB 显存，这严重地制约了其在边缘计算设备的部署潜力。

为突破这些技术瓶颈，学术界提出了一系列创新性改进方案。ALBERT 通过参数共享策略重构模型架构，将嵌入层的维度与隐藏层解耦，通过嵌入层矩阵分解和跨层参数共享，将参数量压缩至原始 BERT 的约 12%，同时保持约 90% 的 GLUE(General Language Understanding Evaluation)基准性能。DistilBERT[31]则开创性地应用知识蒸馏框架，通过温度调节的软标签训练策略，将教师模型(BERT-Base)的语义表征能力迁移至仅含 6 层的小型学生模型，在保持 97%语言理解能力的前提下实现推理速度提升 60%。更具突破性的改进来自 RoBERTa(Robustly Optimized BERT Approach)[32]，该模型通过系统性实证研究揭示了 BERT 原始训练范式的多个优化盲点：首先，其摒弃了效果存疑的 NSP 任务，将训练焦点集中于 MLM 任务；其次，创新性地引入了动态掩码机制，每个训练轮次(Epoch)对输入序列实施随机掩码，相较 BERT 的静态掩码策略提升模型泛化能力达 2.3%；再者，通过将批次大小扩展至 8192 并结合梯度累积技术，在 160GB 超大规模语料(包含 Common Crawl 新闻数据和 OpenWebText 开源语料)上实施 50 万步训练，使模型在 SQuAD 2.0 任务上较原始 BERT 提升 2.4 个 F1 分数(从 86.5 提升到 88.9)。

这些技术突破催生了 BERT 在工业界的深度应用革新。谷歌通过集成 BERT 模型，使特定长尾查询场景下的点击率相对提升 10%，特别是对包含介词短语和疑问词的自然语言查询理解能力得到显著增强。在智能客服领域，结合 BERT 的对话系统在客户意图识别准确率指标上达到 92.7%，相较传统 LSTM 模型提升 18.5%。更深远的影响体现在跨模态研究领域：计算机视觉研究者借鉴 BERT 的预训练范式，开发出 DINO v2 等自监督视觉模型，其通过对比学习策略在 ImageNet 数据集上达到了 81.2%的 Top-1 准确率，延续了自监督视觉表征学习的研究方向。这类模型的成功验证了预训练-微调范式在跨模态领域的普适性，为视觉-语言模型的联合训练提供了方法参考。

3.2 解码器

在 NLP 的 Transformer 架构中，尤其是在生成任务中，解码器部分在许多任务中起到了至关重要的作用。与编码器侧重于对输入进行编码不同，解码器的功能是逐步生成输出序列，这一过程需要根据给定的输入信息逐步解码并生成合理的连贯的文本。解码器被广泛地应用于机器翻译、文本生成、对话系统等任务中，尤其是 GPT 这类模型的解码器架构，已经成为生成任务中的核心技术之一。

3.2.1 GPT

GPT 与传统的 Transformer 编码器-解码器架构不同，只使用了解码器部分，采用自回

归的方式来生成文本，其架构如图 3-3 所示。在 GPT 中，解码器通过逐字生成的过程，利用已经生成的词汇来预测下一个词，并以此来完成整个文本的生成。这一自回归的方式，使生成的每个词都依赖于先前的上下文，这确保了文本的连贯性和一致性。GPT 不仅能完成短文本生成任务，还能处理更长的文本并保持较高的生成质量，尤其在生成具有创意或复杂结构的文本时，展现出其独特的优势。这种单向上下文的生成方式特别适合文本生成任务，如对话生成、文章续写、自动化内容创作等场景。

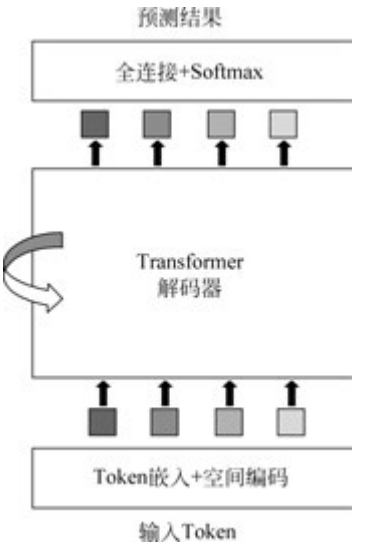

图 3-3　GPT 模型架构

在训练方式上，GPT 采用了“预训练-微调”的训练范式。首先，GPT 通过无监督学习在大规模文本数据上进行预训练，学习语言的基本特征和规律。这一预训练阶段通常依赖于大量的互联网文本数据，包括书籍、新闻文章、维基百科等，使模型能够广泛地掌握语言的语法、语义及上下文间的关系。具体的预训练任务是语言建模任务，模型的目标是通过给定前 $n-1$ 个词来预测下一个词。

给定一个无监督的语料库 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_n\}$，其目标是最大化以下似然函数：

$$L_1(\boldsymbol{U})=\sum_i \log P(u_i \mid u_{i-k},\cdots,u_{i-1};\Theta) \tag{3-1}$$

其中，k 是上下文窗口的大小，条件概率 P 由具有参数 Θ 的神经网络建模。在预训练过程中，模型通过不断调整参数 Θ，使在给定上下文的情况下，预测下一个词的概率尽可能高，这一过程通常使用梯度下降等优化算法来最小化负对数似然函数，从而使模型能够学习到语言的统计规律和语义信息。

在完成预训练后，GPT 进入微调阶段，这一阶段是 GPT 适应具体下游任务的关键过程。在微调过程中，GPT 通过在少量任务特定的数据上进行训练，能够更好地适应各种下游任务。这种预训练和微调的方法通过在大规模无监督数据上进行预训练学习通用的语言表示，然后在有监督数据上进行微调以适应特定任务，同时使用辅助目标来提高微调效果。这种方法不仅提高了模型的泛化能力，还加速了收敛。以情感分析为例，微调后的 GPT 能够准确地识别文本中的情感倾向，这对于构建智能客服系统或舆情监测工具具有重要意义。在问答系统中，GPT 可以通过微调来更好地理解问题的语义，并从给定的文本中生成准确的答案。在机器翻译任务中，GPT 的微调能够使其学习到不同语言之间的映射关系，从而达到高质量的翻译效果。

3.2.2　GPT 的演进

随着 GPT 的发展，OpenAI 陆续推出了多个版本的 GPT 模型，每个版本在前一版本的基础上进行了优化和扩展。例如，GPT-2 通过显著增加模型的参数和训练数据量，提升了

生成质量。GPT-2 的参数数量从 GPT-1 的 1.17 亿增长到约 15 亿，这是一次巨大的跳跃。这一增长使模型能够学习到更多的语言模式和语义关系，极大地提升了其生成文本的连贯性和丰富性。GPT-2 还改进了数据预处理和模型架构，在数据方面，GPT-2 使用了更加多样化和广泛的文本数据来源，覆盖了新闻、对话、书籍等不同领域，这使模型能够适应更多场景下的文本生成任务。在架构方面，尽管仍然采用的是 Transformer 架构，但 GPT-2 的深度和宽度都显著增强，从而更好地捕捉长距离依赖关系，使生成的文本上下文更加连贯、自然。更重要的是，GPT-2 不仅能生成连贯的文本，还能够创作具备创意的内容，展示了其在自动化内容创作领域的潜力。

GPT-3 拥有 1750 亿个参数，规模远远超过之前的所有版本。GPT-3 的发布标志着生成式预训练模型进入了一个新的阶段，其巨大的参数量使它在多种自然语言生成任务中展现出了卓越的性能。与之前的 GPT 模型不同，GPT-3 表现出了惊人的通用性和少样本学习能力。通过少量的任务示例，GPT-3 可以在没有经过专门微调的情况下，完成多项任务，如翻译、编程、数学推理等，这使 GPT-3 成为一个真正的通用语言生成模型。尽管 GPT-3 在文本生成任务中展现了卓越的能力，但它在某些复杂的推理任务中依然存在局限性。

随着 GPT-3 的成功，GPT 系列的后续版本，如 GPT-4 的发布，进一步地推动了解码器技术的进步。GPT-4 比 GPT-3 在规模、生成能力和推理能力上都有了显著提升。GPT-4 不仅能处理更复杂的任务，还展示了更强的上下文理解能力，尤其在多轮对话和长文本生成中，能够保持更高的连贯性和一致性。GPT-4 的训练数据更加多样化和高质量，包括更多的专业领域文本和多语言数据，这使模型在处理跨领域和多语言任务时表现更加出色。此外，GPT-4 在推理任务上比其前代模型更强大，能够进行更复杂的逻辑推理和信息综合，虽然仍面临一些挑战，但它的表现已接近人类水平。GPT-4 采用了更先进的训练技术和架构优化，如更深层次的 Transformer 网络、更高效的并行计算策略等，这些技术的结合使 GPT-4 在处理复杂任务时更加高效和准确。然而，GPT-4 和其他后续版本依然继承了某些问题。例如，虽然推理能力有所提升，但在处理需要快速适应新信息或实时知识的任务时，GPT-4 依然存在局限。GPT 模型的生成结果受到预训练数据的影响，而这些数据往往缺乏最新的信息，这限制了 GPT 在实时任务中的应用，因此研究者正在尝试将 GPT 模型与外部知识库、实时信息检索系统相结合，以提升其对最新数据的响应能力和推理灵活性。

如今，GPT 被广泛集成到实际应用中，尤其是在智能客服、内容创作、对话系统、代码生成等领域。例如，GPT 被用来生成自然流畅的对话回复，提升客服系统的效率；在内容创作方面，GPT 能够为新闻、博客、广告等文案创作提供创意和灵感；此外，GPT 还被广泛地应用于编程领域，通过代码自动补全、生成代码片段等功能，帮助开发者提高工作效率。在智能客服领域，GPT 模型被用于构建能够理解用户意图并提供准确回答的对话系统。这些系统可以处理各种类型的用户查询，包括产品信息、技术支持、常见问题解答等。通过与用户进行多轮对话，GPT 模型能够逐步深入了解用户的需求，并提供更加个性化的服务。

3.3 编码器与解码器结合

编码器-解码器(Encoder-Decoder)架构是一种常见且有效的模型结构,编码器负责处理输入序列,并将其转换为一种潜在的表示形式,而解码器则根据这一潜在表示形式生成输出序列。这种结构被广泛地应用于许多 NLP 任务,例如机器翻译、文本摘要等。举一个例子,在机器翻译任务中,输入是英语句子"Hello, how are you?",解码器则输出对应的法语句子"Bonjour, comment ça va?",然而,随着 Transformer 架构的发展,基于编码器-解码器模型的深度学习模型在处理文本生成任务时,展现出了巨大的潜力和优势。

本节的重点是介绍编码器和解码器结合的模型架构,其中最具代表性且应用广泛的一个实例便是 T5(Text-to-Text Transfer Transformer)。T5 是谷歌在 2020 年提出的一种基于 Transformer 的生成式预训练模型,它在设计上做出了许多创新,使这一架构在处理多种 NLP 任务时具备了前所未有的灵活性和通用性[40]。这种统一的框架让 T5 能够在不同任务间灵活切换,无须为每项任务设计专门的结构,只需通过微调便可适应特定任务的要求。

T5 模型架构如图 3-4 所示,编码器负责接收输入文本,并将其转换为表征,解码器则根据这些表征生成最终的输出文本。T5 的编码器与解码器通过多头自注意力机制进行信息交互,使模型能够灵活地利用输入序列的全局信息,从而生成更加连贯和有意义的文本。

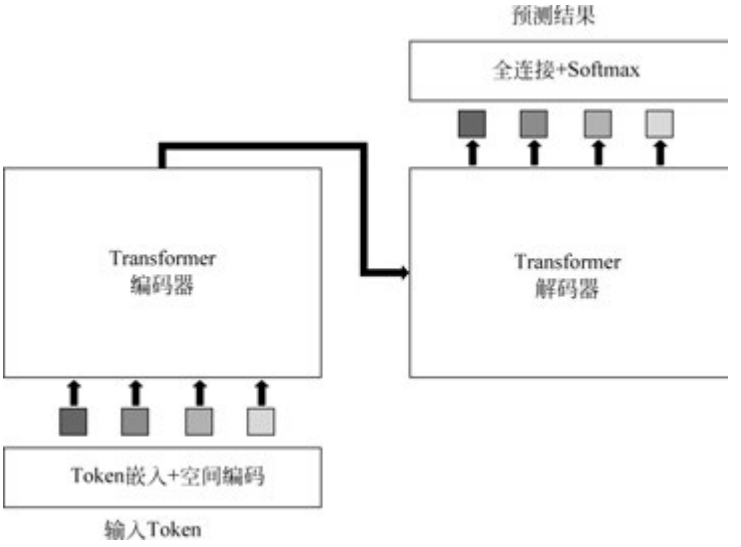

图 3-4 T5 模型架构

T5 的重要创新体现在其预训练目标上,与 BERT 的 MLM 不同,T5 采用了填空任务(Span Corruption)。在填空任务中,模型不会仅仅遮盖单个词汇,而是先随机选择一个连续的词组(Span),然后要求模型生成被遮盖的词组。这种训练方式不仅帮助 T5 学习了词汇之间的关系,还能让它捕捉到句子级别的语义信息。例如,在句子"Today, I am going to

[MASK] friend"中,模型需要生成被遮盖的词组"visit my",从而使句子变为"Today, I am going to visit my friend"。在需要生成长文本的任务中,T5 表现尤为突出。例如,在文本摘要任务中,T5 能够将一篇长文章概括成简洁的摘要,而不是仅仅选择几个关键词,生成的摘要通常更连贯且信息丰富。

T5 的预训练使用了名为 C4(Colossal Clean Crawled Corpus)的大规模通用语料库[41],C4 数据集是从互联网爬取并经过精心过滤清理后的文本数据,覆盖了大量不同领域和类型的内容。与 BERT 等模型使用的传统数据集相比,C4 的规模更大,内容更加多样化。这种丰富的预训练数据使 T5 在许多任务中具有更强的泛化能力,能够更好地适应不同领域的应用,例如,T5 不仅能处理新闻文本,还能够处理社交媒体上的对话、学术论文中的复杂术语等。

与其他基于 Transformer 的预训练模型相比,T5 的一个重要区别在于它的任务表示方式。BERT 是一个仅使用编码器的双向模型,通过同时考虑上下文来生成每个词汇的表示。BERT 的双向性使其在理解任务中表现出色,例如在情感分析或文本分类等任务中,能够准确地捕捉上下文信息。然而,BERT 的生成能力相对有限,因为它的设计主要侧重于理解任务而非生成任务。相比之下,GPT 系列模型是一个单向的生成模型,采用从左到右的自回归方式逐词生成文本,因此在生成任务中更具优势,但在理解任务上可能不如 BERT。T5 则采用了编码器-解码器结构,这意味着它结合了编码器和解码器的优势,既能够处理理解任务,又能处理生成任务,因而在广泛的任务类型中都表现出色,具有更强的通用性。

T5 的架构设计还具有高度的可扩展性,能够适应不同规模的计算资源和任务复杂度。谷歌发布了多个不同规模的 T5 模型,从 T5-Small(6000 万参数)到 T5-XXL(110 亿参数),用户可以根据计算资源和任务的复杂度选择合适的模型规模。例如,较小规模的 T5 模型适合资源有限的环境,如移动设备或嵌入式系统,而较大规模的模型则适用于需要高精度预测和处理大规模数据的任务。T5 的这种可伸缩性使它能够在不同的应用场景下展现出强大的适应能力,无论是在低资源环境中还是在大规模任务中都能提供可靠的性能。

在文本摘要任务中,T5 同样表现优异。它能够从长文本中提取出关键信息并生成简明扼要的摘要。由于 T5 的填空任务涉及生成长跨度的文本片段,因此它在处理长文本生成任务时具有明显的优势。相比其他模型,T5 生成的摘要不仅更加连贯,而且能够更好地捕捉文章的主旨和核心内容。在问答系统中,通过在预训练阶段学习大量的问答对,T5 能够快速地识别问题的核心,并生成相关性强且准确的回答。

尽管 T5 表现优异,但是它仍然面临一些挑战。首先,T5 的预训练和微调过程需要大量的计算资源,特别是对于大型模型(如 T5-XXL)而言,训练时间和硬件需求非常高。这个问题在资源有限的环境中尤其突出。其次,T5 在某些特定任务中的表现仍然可能受到限制,尤其是在需要复杂推理或多步逻辑推导的任务中,T5 可能会生成不一致或不合理的结果。例如,在涉及长逻辑链条的推理任务时,T5 可能会出现逻辑错误,无法产生完全准确的推理结果。此外,由于 T5 的训练数据来源于互联网,所以模型可能在生成文本时表现出偏见或不准确的现象,这是因为互联网数据本身包含了大量的噪声和潜在的偏见。

为了应对这些挑战，研究人员正在积极探索如何优化 T5 的训练方法、减少模型的偏见，并降低模型的计算成本。一些研究者提出了模型压缩和知识蒸馏技术，旨在减少模型的计算资源需求，同时保持其性能。通过这些技术，可以在不牺牲效果的情况下，减少计算资源的消耗，尤其是对于大规模模型。此外，研究人员也在探索更加精细的训练数据选择和模型调优方法，以减少 T5 在实际应用中的错误率和偏见，进一步提高其可靠性和准确性。例如，通过 fastT5 库，可以将 T5 模型转换为 ONNX[42] 格式，并进行量化，从而加快推理速度并减小模型大小。

3.4 Prompt 与 Chain of Thought

在研究大语言模型的极限能力时，Prompt 技术作为人和机器交流的关键部分，正在改变模型与任务之间的连接方式。它通过创建意思清楚的指令范围，把一个个分开的语言符号变成连续的认知对应关系，本质上是借助语言学接口来有针对性地激发模型潜在的能力。和传统监督学习直接改变参数空间不一样，Prompt 技术更像是搭建“认知脚手架”，在输入方面建立语义限制环境，引导模型在特定的解决方案空间里完成推理过程。

Prompt 技术在多个领域展现出了广泛的应用前景和潜力。在文本生成领域，Prompt 可以用于问答系统、摘要生成和对话生成等任务。在问答系统中，通过给定适当的输入提示，Prompt 能够生成高质量的回答。在摘要生成中，Prompt 可以从长文本中提取关键信息，生成简洁明了的摘要。在对话生成中，Prompt 可以生成自然流畅的对话内容。

此外，Prompt 技术还可以用于自动化测试、智能客服、数据提取和跨语言应用等领域。在自动化测试中，通过配置相应的 Prompt 脚本，可以模拟用户在应用程序中的操作，从而验证应用程序的功能和性能。在智能客服领域，Prompt 技术能够理解用户的问题并生成相应的回答，大幅地提高了客服的效率和用户体验。在数据提取中，Prompt 可以从文本数据中提取关键信息。在跨语言应用中，Prompt 技术可以实现不同语言之间的任务处理和应用，有助于打破语言障碍，促进不同语言之间的交流和理解。

作为 Prompt 技术的高级形式，思维链（Chain-of-Thought，CoT）机制开创了符号推理与神经网络融合的新范式[33]。其创新之处在于，首次实现了端到端模型的可解释推理过程的显式化，将传统的黑箱预测转变为透明化的认知推演。具体来讲，当模型输出“首先计算运输次数：1200÷60＝20 次；接着计算总耗时：20×2＝40h”这样的推理步骤时，实际上是将分布式表征中的数学概念激活路径外化为符号操作序列。这种外化过程不仅提升了模型的计算可靠性，更重要的是建立了人机协作的认知桥梁。

CoT 机制的引入，使模型在处理复杂问题时能够逐步分解问题，通过一系列中间步骤来推导出最终答案。这种方法不仅提高了模型的推理能力，还使模型的决策过程更加透明和可解释。特别地，在解决数学问题时，模型可以通过逐步计算和推理，展示出每步的思考过程，从而让用户更好地理解模型的决策依据。此外，CoT 机制还可以与其他技术结合，进一步提升模型的推理能力。例如，CoAT（Chain-of-Associated-Thoughts）框架结合了蒙特

卡洛树搜索(Monte Carlo Tree Search,MCTS)算法和动态关联记忆机制,通过模拟人类的联想和知识更新过程,显著地扩展了模型的推理空间。这种结合不仅提高了模型的推理准确性,还使模型能够动态地整合新信息,从而在复杂的推理任务中表现出色。

目前,CoT 技术已经发展出多维进化路径。在方法论层面,从最初的启发式提示(如逐步思考)演进到结构化推理模板。早期的 CoT 主要通过简单的提示如"Let's think step by step"来激发模型的推理能力,而现在的研究开始探索更复杂的推理结构,如思维树(Tree-of-Thought,ToT)和思维图(Graph-of-Thought,GoT),这些结构通过树状或图状的推理路径,让模型在解决子问题时生成多个不同的答案选择,从而提高推理的准确性和可靠性。最新研究揭示了 CoT 的涌现特性与模型规模的非线性关系,表明千亿参数量级是实现可靠推理链的临界阈值。

在工程实践维度,CoT 技术正在重塑行业应用的范式格局。在金融领域,Morgan Stanley 部署的财富管理系统通过多级推理链实现投资策略的可解释生成。该系统利用 CoT 技术,将复杂的金融数据和市场信息分解为多个子问题,逐步推理出投资策略。这种方法不仅提高了策略的准确性,还增强了其可解释性,使投资者能够更好地理解策略的生成过程。在医疗领域,Mayo Clinic 开发的诊断辅助系统利用分层推理架构,将症状输入逐步转换为鉴别诊断。该系统通过 CoT 技术,将患者的症状和病史信息分解为多个子问题,逐步推理出可能的诊断结果。这种方法不仅提高了诊断的准确性,还增强了诊断过程的透明度,使医生能够更好地理解诊断的依据。在教育领域,可汗学院开发的数学辅导系统采用分步式 CoT,不仅可以输出最终答案,更可以展示完整的解题路径。该系统通过 CoT 技术,将数学问题分解为多个子问题,逐步推理出最终答案。这种方法不仅提高了学生的解题能力,还增强了学习过程的透明度,使学生能够更好地理解解题的步骤和逻辑。这些实践案例揭示,CoT 的价值不仅在于性能提升,更在于建立了人机协作的认知工作流,使系统从"结果生成器"进化为"思维协作者"。

展望未来,CoT 技术的发展将深度耦合认知架构创新与计算范式变革。一方面,受人类工作记忆系统的动态门控机制启发,新型神经架构(如谷歌 DeepMind 的 AlphaGeometry[34])正在整合注意力机制与符号寄存器,实现推理过程的精细控制;另一方面,量子计算带来的并行性突破可能彻底解决长程推理链的计算瓶颈。通过量子比特的叠加与纠缠特性,量子计算能够同时处理多个推理步骤,从而加速长程推理链的计算过程。当这些技术趋向成熟时,人们或将见证"认知增强型人工智能"的诞生——这类系统不仅能解决复杂问题,还能以人类可理解的方式展现其思维轨迹,最终实现人工智能从工具性存在向认知性伙伴的范式跃迁。

3.5 Scaling Law

在深度学习领域,Scaling Law 揭示了模型性能与参数量、数据量和计算资源三者之间的动态关联规律[35]。这一规律不仅为模型设计提供了理论框架,而且在工程实践中成为资源分配与效率优化的核心指导原则。从计算机视觉视角来看,当模型参数规模从数百万扩

展至百亿级别时，其表征能力呈现非线性变化：ResNet-50 在 ImageNet 分类任务中达到 75.3%准确率需要 2380 万参数，而 ViT-L/16 在同等数据条件下通过引入注意力机制将参数规模提升至 8600 万时准确率达到 87.1%，当扩展至 ViT-22B(217 亿参数)时在 JFT-3B 数据集上的准确率提升至 90.45%。这种幂律关系反映了模型从局部特征提取到全局语义建模的范式转变——当参数超过特定阈值后，模型开始建立跨层特征关联，例如 CLIP 模型[36]通过 4 亿参数规模成功实现开放域图像-文本对齐，但更大规模的 Flamingo 模型(800 亿参数)[37]在少样本学习中的提升幅度相对降低。

Scaling Law 可以用数学模型来描述，其一般形式为

$$P = a \times N^{\alpha} + b \tag{3-2}$$

其中，P 是模型的性能(如准确率、损失值等)，N 是模型规模(如参数量、数据量等)，a、b 为常数，α 为缩放系数，该模型表明性能提升与规模增长呈亚线性关系($\alpha<1$)。

值得注意的是，前沿研究正突破传统 Scaling Law 收益递减的范式，MoE 通过动态激活子网络(如 V-MoE 模型仅激活每层 12.5%的参数)，在 ImageNet 任务中保持 90.35%的 Top-1 准确率同时减少 37%的计算量。脉冲神经网络(Spiking Neural Network，SNN)[38]在 Loihi 芯片上实现能效比达 658 GOPS/W，较传统 GPU 可提升 10～100 倍，但具体增益依赖于任务类型。自监督预训练结合强化学习的新范式——DeepMind 的 RoboCat 通过仿真生成 120 万训练样本，使真实世界物体抓取成功率从 61.2%提升至 75.8%。多模态扩展方面，PaLM-E[39]在语言-视觉参数比为 3∶1 时，在 OK-VQA 数据集上的推理准确率呈现 1.2 倍超线性增长。这些进展预示着下一代视觉大模型将不再单纯依赖参数堆砌，而是通过架构创新、算法优化与跨模态协同实现更高效的性能扩展。

此外，强化学习也被视为一种潜在的突破方向。强化学习是一种让智能体通过与环境进行交互来学习最优策略的方法，其核心在于智能体通过观察环境状态、执行动作并根据获得的奖励信号来调整其行为策略。强化学习通常使用马尔可夫决策过程(Markov Decision Process，MDP)进行建模，其中智能体在每个时间步先根据当前状态选择一个动作，然后环境会反馈给智能体一个奖励并转移到下一种状态。通过这种方式，智能体可以学习到一种策略，使长期累积奖励最大化，例如，在 AlphaGo 中，通过强化学习，模型可以通过自我对弈等方式提升其推理能力，从而在特定任务中取得更好的性能。在自我对弈过程中，模型会不断地与自己或其过去版本进行对弈，通过这种方式来学习和改进策略，使模型能够在复杂的环境中不断地优化其决策过程。

多智能体系统的 Scaling Law 也是一个新兴的研究方向，随着投入系统的智能体(Agent)数量的增加，其表现出来的智能越来越强。在多智能体系统中，智能体之间可以通过通信和协作来共同完成复杂的任务。随着智能体数量的增加，系统中的信息交流和协作机会也会增加，从而使系统的整体智能水平得到提升。例如，在蚁群觅食过程中，随着蚂蚁数量的增加，它们可以通过信息素的交流来更有效地找到食物源，从而表现出更强的集体智能。此外，多智能体系统中的智能体可以通过强化学习来不断地优化其行为策略，从而进一步提升系统的性能。