2.1.Deep Graph Learning

2.1.1 Graph Data

（1） 非欧几里德性质。与具有固定邻域区域和确定性顺序的欧几里德数据相比，图数据明确地编码了实体之间的复杂连接。

（2）各种领域：图数据在包括社交网络、生物学和交通运输等各种领域中普遍存在。不同领域的图数据可能具有不同的特征，包括不同的节点类型、边的语义和结构模式。例如，在生物网络中，节点可以代表蛋白质、基因或代谢物，而边可能表示蛋白质间相互作用或代谢反应等相互作用。这种领域特定的变化使得创建一个能够有效泛化和适应各种图结构的通用模型变得具有挑战性。

（3）各种类型：图数据具有多种类型，包括同质图[146]、异质图[104]、超图[23] 和动态图[105] 等。同质图包含相同类型的节点和边，例如由论文组成的引用网络。异质图包含多种类型的节点或边，例如包含作者和论文的引用网络。超图由连接多个节点的超边组成，可以模拟节点之间的高阶关系。动态图指的是节点和边随时间变化的图结构，例如交通流量变化形成的交通网络。

2.1.2 骨干架构

作为当前主流的骨干架构，图神经网络（GNNs）已经成为深度图学习的强大框架。大多数GNNs遵循**消息传递框架**[26]，该框架使得图中的节点能够与它们的邻居交换信息。例如，GCN[57]引入了**图卷积层**的概念，并为许多后续的GNN架构奠定了基础。GraphSAGE[32]提出了一种利用**归纳学习**为大型图中的节点生成嵌入的方法。此外，GAT[120]将**注意机制**引入到GNNs中，**允许节点在消息传递过程中对它们的邻居的重要性进行加权，增强了它们的表达能力**。这些工作极大地推动了GNNs的发展，使它们成为深度图学习的多功能工具。

尽管更深的神经网络可以获得更强的表达能力[33]，但深化GNNs并不容易。原因在于，随着GNN中层的数量增加，过多的噪声被引入到消息聚合过程中，导致所有节点的表示变得相似[65]。这也被称为过度平滑问题。此外，层数增加会导致感受野（Receptive Field）呈指数增长，导致过度压缩问题[1]的出现，其中大量信息被压缩成固定长度的节点向量。近年来，一些工作致力于解决深度图神经网络中的过度平滑问题和过度压缩问题，从而提高了下游任务的性能。例如，像DropEdge[100]这样的创新通过随机移除边来增强GCN模型的性能和可扩展性（最多达到64层）。另一个改进GNNs表达能力的研究方向是图变换器[7,58,158]。由于其完全连接的注意机制和长距离关系建模能力，图变换器架构可以缓解过度平滑问题和过度压缩问题[96]。

【**感受野**

在深度学习中，感受野（Receptive Field）是指在神经网络中某一层的输出值受到输入的影响的范围。在卷积神经网络（CNN）中，每一层的感受野表示了该层对输入图像的局部区域的感知范围。随着网络层数的增加，每一层的感受野也会随之扩大，从而使网络可以捕捉到更广泛的上下文信息。感受野的大小决定了网络对输入数据的理解能力，较大的感受野可以捕捉更多的上下文信息，但也会增加模型的计算量和复杂度。】

【**过度压缩问题**

在图神经网络中，过度压缩问题指的是在增加网络层数时，节点表示的信息被过度压缩或丢失的现象。随着网络层数的增加，每一层对节点表示的更新会受到更多层次的影响，导致节点表示逐渐丧失了原始图结构的信息，最终导致所有节点的表示趋于相似或退化为相同的表示。这种情况会降低网络的表达能力和泛化能力，从而影响模型在下游任务上的性能。过度压缩问题的出现可能是由于信息传播的衰减或信息传播的范围过大等因素引起的。】

**2.1.3 学习范式**

深度图学习的学习范式包括三个主要类别：**监督学习、半监督学习和无监督学习**。在本节中，我们将对这些学习范式进行简要介绍。

**监督学习**

在监督设置中，**算法利用包含输入数据和相应输出标签的训练数据集**。这种范式在图分类[59]和图回归[48]等任务中找到了实际应用。例如，在分子性质预测任务[138]中，GNN 模型被训练以使用带有标签的训练数据来预测分子的特定化学性质或属性，从而为药物开发和材料研究的有价值洞察提供了可能。

【图回归是指在图数据上进行回归任务的一种方法。在图回归中，目标是预测图中节点的某些属性或值，这些属性1或值通常是连续的。例如，可以通过图回归来预测社交网络中用户的年龄、收入或兴趣程度等连续性属性。在化学分子图中，图回归可以用于预测分子的特定化学性质，如溶解度或能量。通常，图回归任务涉及利用图结构和节点特征进行预测，同时考虑节点之间的关系和连接。】

**半监督学习**

半监督学习，如最近的一项研究所强调的[106]，构成了深度图学习的主要焦点。这种方法利用**带有标签和未标记数据**来增强模型性能，其中节点分类[57]是一个突出的应用。消息传递机制[26]赋予了 GNN 在邻近节点之间迭代交换信息的能力。这种能力使得信息可以在整个图中传播，有效地将标记和未标记数据结合起来以促进预测。此外，GNN 还可以与传统方法如标签传播相结合，进一步提高其在半监督设置下的性能[151]。

【什么是标签传播算法？】

【标签传播指的是一种算法或技术，用于将已知标签的数据（有标签数据）传播到未知标签的数据（未标记数据）。在半监督学习中，通常只有一小部分数据被标记，而大部分数据是未标记的。标签传播算法可以利用已标记数据中的信息，通过在图或网络结构中传播标签信息，从而为未标记数据赋予标签，以提高模型性能。

标签传播是一种半监督机器学习算法，用于为之前未标记的数据点分配标签。在算法开始时，数据点的一个（通常是小的）子集具有标签（或分类）。这些标签在算法的执行过程中传播到未标记的数据点。在复杂网络中，真实网络往往具有社区结构。标签传播是一种用于发现社区的算法。与其他算法相比，标签传播在运行时间和对网络结构的先验信息需求量方面具有优势（无需预先知道任何参数）。缺点是它不会产生唯一解，而是多个解的总和。】

**无监督学习**

无监督学习[95]是一种更广泛的机器学习方法，从数据中学习模式和结构而无需手动标签。例如，图聚类（graph clustering）旨在仅基于节点之间的关系和连接性来发现图中固有的结构和模式。另一个例子是链接预测（link prediction），其目标是预测缺失或即将发生的连接关系。无监督学习的一个重要子类是自监督学习，其目标是利用数据本身固有的信息生成标签。基于自监督学习，GNN 可以进行端到端的训练，并应用于图聚类[125]和链接预测[168]等下游任务。

2.2. 语言基础模型 Language Foundation Models

目前，人工智能正在经历一次转型性的变革，其中一些特定的自然语言模型（例如 GPT-3）通过大规模的**自监督学习**，使用广泛而多样的数据集进行训练。这些被称为基础模型的模型能够产生各种各样的输出，使它们能够解决广泛的下游任务。与深度图学习流程相比，基础模型的方法采用了**预训练**和**适应框架**，使模型能够实现几项重大进展，包括涌现能力和同质化。基础模型主要已在自然语言处理领域确立了自己的地位，因此我们的讨论将在本节中聚焦于语言基础模型。

2.2.1 语言数据

语言数据指的是人类语言中的文本或口头内容，涵盖了自然语言的语法规则以及单词的相关语义。它可以包括书面文件、转录的音频记录以及任何其他形式的基于语言的交流。语言数据对于许多自然语言处理任务至关重要，例如机器翻译、情感分析和文本摘要。研究人员和开发者使用语言数据来训练和评估语言模型和其他自然语言处理算法。语言数据的质量和数量在自然语言处理系统的性能中起着至关重要的作用，影响其在各种语言任务中的准确性、鲁棒性和整体效果。与计算机视觉和其他领域相比，标记语言数据的规模相对较小，仅包含几千个句子。这种限制主要是由于手动注释的高成本所致。然而，有大量未标记的语言数据可从互联网、报纸和书籍等来源获取，这为利用未标记数据进行模型预训练创造了机会。此外，与图数据相比，语言数据作为欧几里得数据更容易建模，其丰富的语义信息显著增强了语言模型的知识可迁移性。

【标记的语言数据集成本高，而未标记的文本数量大，催生出模型预训练】

【相较于图数据，语言数据作为欧几里得数据跟容易建模，包含丰富的语义信息增强了语言模型的知识可迁移性】

【模型的预训练是指在将模型应用于特定任务之前，在大规模数据上进行初始训练的过程。预训练旨在让模型学习有关语言或数据的通用表示和模式，而不针对特定任务进行优化。通过在大量数据上进行预训练，模型能够捕捉到语言或数据中的各种模式、规律和特征。

在自然语言处理领域，预训练通常指的是在大型文本语料库上对语言模型进行训练，使其能够理解语言的语法、语义和语境。这种预训练模型通常被称为语言基础模型。预训练的语言模型可以用于多种下游任务，如文本分类、命名实体识别、问答等，通常通过微调或进一步训练来适应特定任务。

预训练的主要优势在于，通过在大量数据上进行训练，模型可以获得更丰富的知识和表示，从而在特定任务上表现更好。此外，预训练可以减少针对特定任务的数据需求，因为模型已经具备了一定程度的通用知识。】

【

语言模型的知识可迁移性指的是模型在完成一个任务后，所学到的知识、表示和模式能够有效地迁移到其他相关任务中，从而提高模型在这些任务上的性能。这种可迁移性意味着模型在进行预训练时学习到的通用语言知识和表示能够泛化到各种下游任务中，并为这些任务提供更好的基础。

具有良好知识可迁移性的语言模型通常能够：

1. \*\*捕捉语言的通用规律和模式\*\*：通过在大规模语料库上进行预训练，模型可以学习到语言的通用结构、语法和语义规律，从而能够应对不同领域和任务中的语言理解和生成问题。

2. \*\*提取丰富的语义信息\*\*：预训练的语言模型可以学习到单词、短语和句子之间的语义关系，以及不同语境下的词义和含义，从而能够更好地理解和生成自然语言文本。

3. \*\*适应不同的下游任务\*\*：由于预训练模型学习到的知识具有通用性，因此可以在各种自然语言处理任务中进行微调或迁移学习，而无需从头开始训练模型。这样可以大大减少训练成本，并提高模型在新任务上的性能。

总之，语言模型的知识可迁移性使得模型能够更有效地利用先前学到的知识和表示，从而在不同任务和领域中展现出更强的适应能力和泛化能力。】

2.2.2 骨干架构

基础模型的一个早期突破是**预训练语言模型**（PLMs），**旨在捕捉上下文感知的单词表示**，被证明是极为有效的通用语义特征。例如，**BERT**，基于可并行化的Transformer架构，配备了**自注意力机制**，是通过在**大量未标记数据上预训练**双向语言模型，并使用特定设计的预文本任务构思而成的。【For instance,BERT[13],grounded in the parallelizable Transformer architecture[119]with self-attention mechanisms,is conceived through the pre-training of bidirectional language models with specifically designed pretext tasks on vast unlabeled data.】这一标志性研究显著提升了自然语言处理任务的性能基准，并成为后续大量研究的催化剂，确立了当前的预训练和微调学习范式。

【预训练语言模型，捕捉上下文感知的单词表示。】

此外，研究人员观察到，通过增加PLMs的规模，无论是通过增加模型大小还是训练数据，往往会导致模型在下游任务中的能力提升。这些规模更大的PLMs，统称为LLMs，在行为上与规模较小的对应模型（例如，拥有15亿参数的GPT-2和1750亿参数的GPT-3）呈现出独特的特征。在大规模文本数据集上训练后，它们表现出显著的能力，通常被称为新兴能力，例如上下文学习。LLMs主要利用Transformer架构，因为高度可并行化的基于Transformer的架构加速了预训练阶段，并且能够利用大规模数据集。在Transformer模型的背景下，标记（token）用作输入，代表自然语言文本中的词级单位。通常，LLMs包含数千亿（或更多）的参数，例如GPT-3、PaLM、Galactica和LLaMA等模型。

【总结：更大规模的PLM（统称为LLMs）在下游任务中解决任务的能力更强】

2.2.3学习范式

随着模型参数数量的迅速增加，对更大的数据集的需求也增长了，以有效地训练这些参数并避免过拟合。考虑到构建大规模标记数据集所需的极高成本，强调了利用大量未标记文本数据的重要性。**利用这些未标记数据集涉及到一个两步方法：首先，通过自监督学习实现通用表示，然后利用这些表示进行各种任务**。基于不同的适应方法，学习范式可以分为两种类型：**预训练和微调**以及**预训练、提示和预测**。

**预训练和微调**

在这种范式中，一个具有一致架构的模型最初被预训练为语言模型（LM），在这个过程中，它预测观察到的文本数据的概率。与端到端训练相比，预训练提供了独特的优势，并构成了基础模型能力的基石。首先，通过在庞大的文本语料库上进行预训练，使模型能够学习到通用的语言表示，这可能是解释新兴能力的一个原因。此外，预训练提供了改进的模型初始化，通常导致增强的泛化性能，使多个任务具有同质化的能力。此外，预训练还起到一种正则化的作用，有助于防止在较小数据集上的过拟合。例如，像GPT-3这样的模型被训练以语言建模为目标，为其在新兴能力和同质化方面的能力提供奖励。

**在预训练阶段之后，基础模型获得了适用于广泛任务的通用能力。**然而，**预训练模型仍然缺乏下游任务特定的信息，直接使用它们可能不会产生最佳结果。因此，我们需要为特定任务微调模型，这就是所谓的微调**。在ULMFit和BERT等模型取得成功的基础上，微调已经成为调整预训练模型的主要方法。在这种框架中，主要重点在于目标工程，包括为预训练和微调阶段设计训练目标。例如，Pegasus表明，将一个用于预测文档中重要句子的损失函数纳入模型训练，可以为文本摘要生成得到一个改进的预训练模型。微调的优势在于它可以在源任务（或领域）和目标任务之间转移知识，并提高模型的性能。由于微调数据集相对于预训练数据集较小，这个过程可以在不丢失存储的结构性语言知识的情况下有效地实现适应。

**预训练、提示词、预测**

Pre-train,Prompt and Predict.

在这种范式中，与调整预训练语言模型以适应特定的下游任务不同，这种方法涉及重新塑造下游任务，使其与原始语言模型训练期间处理的任务更加接近，通过提供文本提示【textual prompts.】来实现。通过选择合适的提示，我们可以引导语言模型的行为，使其能够预测所需的输出，有时甚至无需进行任何额外的任务特定训练。这种方法的优势在于，当配备了一组合适的提示时，它可以使单个完全无监督的语言模型处理各种任务。

从提示工程的角度来看，创建适当提示的方法可以分为手动方法和自动方法。手动方法涉及基于人类洞察力创建直观模板，这是制作提示的最直接方法。例如，有影响力的LAMA数据集提供了手动设计的填空模板，用于评估语言模型的知识。然而，手动方法在成本和精度方面面临挑战。为了解决这些问题，一些方法已经开始尝试自动提示生成。例如，**Prompt Mining**是一种模板发现方法，根据给定的一组训练输入和输出自动识别模板。

从另一个角度来看，**从模型和提示如何组合以生成结果的角度来看**，提示策略可以分为三种方法**：无调整提示、提示调整和指令调整**。无调整提示直接基于提示生成答案，而不改变预训练LLMs的参数。提示调整在预训练模型的参数之外引入了额外的提示相关参数，并使用从下游训练样本获取的监督信号更新这些额外的参数。指令调整以类似微调过程的方式调整LM的参数，同时另外引入固定指令来指导模型的行为。这种方法在零样本情况下尤其有潜在的增强效果。

**3. 图基础模型**

本章介绍图基础模型的定义，关键特性和关键技术。然后我们将讨论图数据和图任务对图基础模型的影响。最后，我们将讨论图基础模型与语言基础模型之间的相似之处和不同之处。

**3.1 图基础模型的概念**

在本小节中，我们首先提供图基础模型的定义。随后，我们将深入探讨图基础模型的关键特征和基本技术，以及图数据和图任务对图基础模型的影响。

**3.1.1 定义和关键特征**

**定义 3.1：图基础模型（GFM）是一种期望从广泛的图数据的预训练中受益，并且可以适应各种下游图任务的模型。**

与采用端到端训练的深度图学习相比，图基础模型使用预训练来获取大量未标记图数据的知识，然后使用适应技术来适应各种下游任务。一些研究已经证明，在某些场景下，例如少样本学习，预训练和适应范式胜过深度图学习，展示了它们卓越的表达能力和泛化能力。

与旨在单个任务上取得更好性能的深度图学习不同，图基础模型预期具有两个关键特征：新能力(emergence)和同质化(homogenization)。

**新能力emergence**：emergence意味着当图基础模型具有大量参数或经过更多数据训练时，它将展示一些新的能力，也被称为新兴能力。 受到基础模型拥有的各种新兴能力的启发，我们期望图基础模型具有类似的能力，包括**上下文学习、图推理和零样本图生成等**。

* 上下文学习允许使用少量示例对各种下游任务进行预测。
* 图推理将复杂问题根据图结构分解为多个子问题，并逐步解决它们，例如解决图算法问题。
* 零样本图生成要求模型在没有任何示例的情况下根据所需条件生成图。

需要注意的是，尽管语言基础模型已经展示了各种新兴能力，但目前对图基础模型的新兴能力的研究工作仍较少。

**同质化homogenization**：homogenization意味着图基础模型可以应用于不同格式的任务，例如节点分类、链接预测和图分类。需要注意的是，由于图任务相对于自然语言处理任务具有不同的特点，实现同质化并不直接。**在实现同质化时的根本问题是决定如何统一不同类型的图任务的形式**。现有研究尝试通过**链接预测**或**图级任务**来实现同质化，但目前尚无关于哪种方法更优的共识。

**3.1.2 关键技术**

图基础模型主要包括两个关键技术：预训练和适应。本节将简要介绍这两种技术。

**预训练Pre-training：**

预训练是图基础模型发展中的一个关键概念，类似于语言模型中的作用。它涉及以自监督的方式在大规模图数据集上预训练神经网络。在预训练过程中，模型学习捕捉图中的结构信息、关系和模式。图基础模型的预训练策略有几种。对比自监督学习利用对比正样本（例如相似节点对）与负样本（例如不相似节点对）的想法来学习表示[110,181]。生成自监督学习鼓励模型重构原始图数据的结构或预测其特征[37,38]。如果将LLM作为图基础模型的一部分，我们也可以使用第2.2.3节介绍的预训练方法。这些多样化的预训练方法使图基础模型能够从原始图数据中学习有意义的表示，增强其在各种图任务中的泛化和适应能力。

**适应adaptation：**

图基础模型的适应涉及将这些模型调整到特定的下游任务或领域以增强其性能。这个过程包括几种技术，即普通微调、参数高效微调和提示微调。普通微调(Vanilla FT)涉及在任务特定数据上训练整个预训练模型，允许最高程度的定制化，但通常需要大量的数据和计算资源。另一方面，参数高效微调(Parameter-efficient FT)只调整模型参数的子集，在任务特定适应性和资源效率之间取得平衡[28,66]。提示微调是一种多功能方法，依赖于外部提示来指导模型的行为，使其更具适应性和有效性[111,113]。这些适应技术使图基础模型能够在各种应用中脱颖而出，利用其预训练知识，并将其能力调整到特定任务或领域，使其在各种下游应用中都有价值。需要注意的是，尽管LLMs已经发展了各种类型的提示微调方法[72]和一些其他高效的微调方法，例如Prefix Tuning[68]，但是对于图基础模型来说，相对而言提示微调方法相对较少。

3.1.3 图数据的影响

基础模型的成功取决于高质量的训练数据，并且基础模型在各种类型的测试数据上表现出显著不同的性能。在本节中，我们将从三个方面讨论图数据对图基础模型的影响：图类型、图规模和图多样性。

**Graph Type**

根据图中节点和边的类别数量，我们可以将图分类为同质图和异质图。在同质图中，所有节点和边属于同一类别。例如，在社交图中，节点表示个体（用户），边表示友谊关系，这是一个同质图，因为所有节点都是个体，所有边都表示友谊关系。另一方面，异质图具有多种类型的节点或边，代表不同类型的实体和关系。例如，电子商务图可能包含用户、产品和购买关系的节点，形成一个异质图。对于图基础模型，处理异质图会带来更大的挑战，并且通常需要设计特定的骨干架构和优化目标。尽管如此，利用基于**元路径（mata-path？）**的方法，可以将异质图映射到多个同质图中，每个元路径对应一个同质图。例如，可以将训练在同质图上的GFM应用到这些元路径诱导的同质图中的每一个上，分别得到节点嵌入。然后，这些在不同元路径下的同质图上的嵌入可以进行融合。然而，除了同质图和异质图之外，在现实世界中还有一些更复杂的图类型，例如动态图和超图，这为GFM带来了额外的挑战。

**Graph Scale.**

基于图中节点和边的数量，我们可以将图分为相对较小的图和大型图。小型图规模较小，通常包含几十到几百个节点和边。例如，化学分子图表示小分子的结构，通常包含几十到几百个原子。另一方面，大型图指的是具有大量节点和边的图，通常包含数百万甚至数十亿个节点和边。例如，阿里巴巴的电子商务图包含数十亿个节点和数百亿条边。**对于图基础模型，大型图对图基础模型的能力提出了更高的要求。首先，由于大型图具有众多的节点和通常较稀疏的边，引入了更多的噪声，并在存储和计算方面带来了更大的挑战。此外，大型图通常表现出更长的依赖关系，需要更多的神经网络层和更多的参数，这加剧了基于GNN模型的过度平滑和过度压缩问题。**

**Graph Diversity.**

基于图数据是否来自相同的领域，我们可以将图分为同领域图和跨领域图。同领域图指的是来自相似或相关领域的图数据，通常包含相似类型的节点和边。例如，Facebook和微信的社交图来自相似的领域。另一方面，跨领域图涉及来自不同领域或数据源的图数据，通常包含不同类型的节点和边，旨在解决多领域问题或跨领域任务。例如，学术网络和分子图来自不同的领域。对于图基础模型，支持跨领域图提出了更大的挑战，因为来自不同领域的图缺乏统一的基础语义。这可能导致在将模型应用于新数据集时出现弱传递性甚至负传递性。因此，**解决不同领域的异质性，并使相同的GFM能够应用于不同领域的图是GFMs面临的重要挑战**。

【“弱传递性”和“负传递性”是指在将模型从一个领域或数据集迁移到另一个领域或数据集时，模型的性能降低的情况。

* 弱传递性：当模型在新的数据集上的性能较之前的数据集下降，但下降幅度不是非常明显时，我们可以说模型表现出弱传递性。这意味着模型在新数据集上仍然有一定的泛化能力，但不如在原始数据集上表现得好。
* 负传递性：当模型在新的数据集上的性能明显低于在原始数据集上的性能时，我们可以说模型表现出负传递性。这意味着模型在新数据集上的表现非常差，可能无法很好地泛化到新的领域或数据集上。】

**3.1.4 Impact from Graph Tasks:**

语言基础模型可以广泛应用于各种自然语言处理（NLP）任务，而对于图基础模型，图任务的格式也非常多样，可以分为三类：节点级任务、边级任务和图级任务。

节点级任务：节点级任务是指对每个节点执行的分类、回归或预测任务。常见的节点级任务包括节点分类、节点回归和聚类系数预测。例如，在社交网络中，图节点可以代表用户，节点分类可以用于识别不同社交圈子中的用户或具有不同兴趣的用户。

边级任务：边级任务涉及对每个单独边执行的分类、回归或预测任务。常见的边级任务包括边分类、链路预测、最短路径预测、连接性预测和最大流预测。例如，在电子商务中，链路预测可以用于预测用户可能感兴趣的产品。

图级任务：图级任务关注整个图。常见的图级任务包括图分类、图回归、图生成、图聚类、图凝聚和平均聚类系数预测。例如，在生物信息学中，图属性预测可以用于预测分子化合物的生物活性或毒性，从而加速药物发现过程。

总之，图中任务的格式非常多样，可以分为三种类型：节点级、边级和图级，每种类型都有广泛的应用。这无疑增加了图基础模型的同质化挑战。例如，在（synthetic）合成数据集上的图分类和节点分类任务中，建模结构信息通常更为关键。另一方面，在具有丰富节点特征的图上处理节点分类任务时，建模特征信息变得更为重要。此外，更相似的任务也将具有更低的转移难度，这意味着这些任务更有可能使用相同的GFM解决。尽管增强表达能力有望提高许多节点级、边级和图级任务的性能，但也有一些研究表明，对于图生成任务来说，过强的表达能力可能并非必要。

【Transformer架构是一种用于自然语言处理（NLP）任务的深度学习模型架构，由Vaswani等人在2017年提出。它通过使用**自注意力机制**（self-attention mechanism）来处理输入序列，不再需要循环神经网络（RNN）或卷积神经网络（CNN）等传统序列模型。Transformer模型的主要组成部分包括编码器（encoder）和解码器（decoder）。

1. 编码器（Encoder）：编码器接受输入序列并将其转换为表示序列，其中每个表示对应于输入序列中的一个位置。编码器由多个相同的层堆叠而成，每个层包含两个子层：

- 多头自注意力机制（Multi-Head Self-Attention）：它允许编码器同时关注输入序列的不同位置，从而更好地捕捉序列中的长距离依赖关系。

- 前馈神经网络（Feed-Forward Neural Network）：它对每个位置的表示进行全连接层的处理，以便进行特征提取和转换。

2. \*\*解码器（Decoder）\*\*：解码器接受编码器产生的表示序列，并生成目标序列的预测。解码器也由多个相同的层堆叠而成，每个层包含三个子层：

- 多头自注意力机制：类似于编码器中的自注意力机制，但在这里，解码器还要执行掩码操作，以确保在生成目标序列时不会使用未来位置的信息。

- 编码器-解码器注意力机制（Encoder-Decoder Attention）：这一层使得解码器可以关注编码器产生的表示序列，以便了解输入序列的信息。

- 前馈神经网络：与编码器中的相同。

Transformer架构的关键创新在于自注意力机制，它使得模型可以并行处理序列中的所有位置，并且在捕捉长距离依赖关系方面表现出色。Transformer模型在NLP领域中取得了巨大成功，并成为了许多任务的基础模型，如机器翻译、文本生成和情感分析等。】