ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 3,May2024,pp. 524－546
DOI:10. 16337/j. 1004-9037. 2024. 03. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

垂直领域大模型的定制化:理论基础与关键技术

陈浩泷[1,2]，陈罕之[1]，韩凯峰[3]，朱光旭[1,2]，赵奕晨[4]，杜　滢[3]

(1. 深圳市大数据研究院，深圳 518172；2. 香港中文大学(深圳)理工学院，深圳 518172；3. 中国信息通信研究院，北京 100191；4. 中国移动通信集团终端有限公司，北京 100033)

摘　要：随着 ChatGPT 等基于大模型的产品展现出强大的通用性能,学术界和工业界正积极探索如何将这些模型适配到特定行业和应用场景中,即进行垂直领域大模型的定制化。然而,现有的通用大模型可能无法完全适配特定领域数据的格式,或不足以捕捉该领域的独特需求。因此,本文旨在探讨垂直领域大模型定制化的方法论,包括大模型的定义和类别、通用架构的描述、大模型有效性背后的理论基础,以及几种可行的垂直领域大模型构建方法,期望通过这些内容为相关领域的研究者和从业者在垂直领域大模型定制化方面提供指导和参考。

关键词：人工智能;垂直领域大模型;多模态大模型;预训练大模型;大模型微调

中图分类号：TP183　　**文献标志码**：A

Domain-Specific Foundation-Model Customization: Theoretical Foundation and Key Technology

CHEN Haolong[1,2], CHEN Hanzhi[1], HAN Kaifeng[3], ZHU Guangxu[1,2], ZHAO Yichen[4], DU Ying[3]

(1. Shenzhen Research Institute of Big Data, Shenzhen 518172, China; 2. School of Science and Engineering, the Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China; 3. China Academy of Information and Communications Technology, Beijing 100191, China; 4. China Mobile Group Device Co. Ltd., Beijing 100033, China)

Abstract: As ChatGPT and other foundation-model-based products demonstrate powerful general performance, both academia and industry are actively exploring how to adapt these models to specific industries and application scenarios, a process known as the customization of domain-specific foundation models. However, the existing general-purpose foundation models may not fully accommodate the patterns of domain-specific data or fail to capture the unique needs of the field. Therefore, this paper aims to discuss the methodology for customizing domain-specific foundation models, including the definition and types of foundation models, the description of their general architecture, the theoretical foundations behind the effectiveness of foundation models, and several feasible methods for constructing domain-specific foundation models. By presenting this content, we hope to provide guidance and reference for researchers and practitioners in the customization of domain-specific foundation models.

Key words: artificial intelligence; domain-specific foundation model; multimodal large model; pre-trained foundation model; fine tuning of large models

---

基金项目:广东省基础与应用基础研究重大项目(2023B0303000001);国家自然科学基金面上项目(62371313);广东省基础与应用基础研究基金面上项目(2022A1515010109)。
收稿日期:2024-04-09;**修订日期**:2024-04-30

引　言

ChatGPT以其卓越的通用性能重塑了人们对人工智能的理解。作为ChatGPT的核心,大语言模型(Large language model)已经成为众多领域研究人员和专业人士改进工作流程的重要工具。通用大模型通常在广泛的公开数据集上进行训练,这使得它们能够学习并解决各种常见问题,但这些数据集无法完全覆盖某些特定领域的所有专业知识和技术细节,这导致尽管通用大模型具备广泛的通用知识,却缺乏足够的知识深度来满足某些特定领域的复杂需求。因此,针对特定行业的需求来构建垂直领域大模型变得尤为重要。垂直领域大模型,或称垂类大模型、行业大模型,是针对特定领域的数据和应用而开发的大模型[1]。与通用大模型相比,它们在训练过程中会使用大量特定领域的数据,从而能够更准确地理解和生成与该领域相关的专业内容。

随着类ChatGPT的产品和神经网络模型的接连推出,“大模型”概念的范围也在逐步扩张[2-4]。鉴于相关概念繁杂,为了确定本文的研究共识,需要对“大模型”概念进行定义并阐述其特点,从而奠定后文对垂直领域大模型定制化的叙述基础。本文所提及的大模型(Foundation model),是在多模态大模型(Multimodal large model)五模块框架(下文将详细介绍该框架)中,包含了能够实现其中一个或多个模块功能的神经网络模型,且该模型符合以下特点:

(1) 大数据。使用覆盖了多种场景的大量数据进行模型的训练,为模型提供充足的知识。

(2) 大参数。模型的参数量达到一定规模,足以将大量数据中隐含的知识固化到模型参数中。

(3) 通用性。模型的输入数据格式和数据处理流程能够适配多种任务场景下的输入格式和需求。

(4) 泛化性。模型拥有一定的泛化性,使其在未知数据域中依然具有良好性能。

根据大模型可处理的模态数量,可将大模型分为单模态大模型和多模态大模型:

(1) 单模态大模型。VGG[5],ResNet[6],GPT-1[7],GPT-2[8],GPT-3[9],GPT-3.5 turbo[10],BERT[11],GLM[12-13],LLaMA[14],LLaMA-2[15],iGPT[16],LVM[17],BART[18]和T5[19]。

(2) 多模态大模型。CoDi[20],CoDi-2[21],Claude-3[22],GPT-4[23],LLaVA[24],BriVL[25],ImageBind[26]和NExT-GPT[27]。

在构建垂直领域大模型的过程中将面临一系列挑战,尤其是在数据获取和预处理阶段。比如,其需要处理的垂直领域数据并不开源或难以获取,具有私密性;或是数据模态与通用大模型使用的中心模态不同,导致无法迁移现成的大模型处理该数据;又或是垂直领域数据与预训练模型的数据域有所不同,需要向预训练模型输入专业领域知识。垂直领域大模型应用方式灵活,涉及的应用领域繁杂,构建难度大、开销大,涉及的技术安全问题至关重要,期望产生的经济效益高[28-30],因此有必要对其构建方法论进行深入探索和全面梳理,并总结出相应的方法论。

以往的综述文献都更多地关注大模型本身的发展[2-4,31-36],但对于垂直领域大模型的定制化方法论方面缺乏详细的讨论。本文通过介绍垂直领域大模型定制的理论基础、垂直领域大模型的定制方法、垂直领域大模型的应用实例,以及垂直领域大模型定制化的未来发展方向,为有意构建垂直领域大模型应用的研究者及工作者提供模型定制方法论层面的参考。

1　垂直领域大模型定制的理论基础

首先从大模型的架构入手,介绍构建大模型可能涉及的所有功能模块,然后从特征提取、模态对齐、规模幂律和涌现现象4个角度,解释大模型的各模块能够提供良好性能的理论基础。

1.1　大模型的架构

参考目前的大模型相关研究,认为多模态大模型在理论上能够包含所有单模态大模型的功能和结构,即单模态大模型就是实现了多模态大模型部分功能的大模型。

文献[34]对于多模态大语言模型提出的五模块框架能够很好地囊括以语言作为中心模态的多模态大模型架构。但近期像视觉大模型[17]、图模态大模型[37]这样的非语言模态大模型主干(Backbone)的诞生,预示着大模型的主干部分将不再拘泥于语言模态。于是认为,多模态大模型的结构可以分为以下5个模块:模态编码器、输入投影器、主干运算器、输出投影器和模态解码器。图1展示了以语言作为中心模态的多模态大模型的框架。

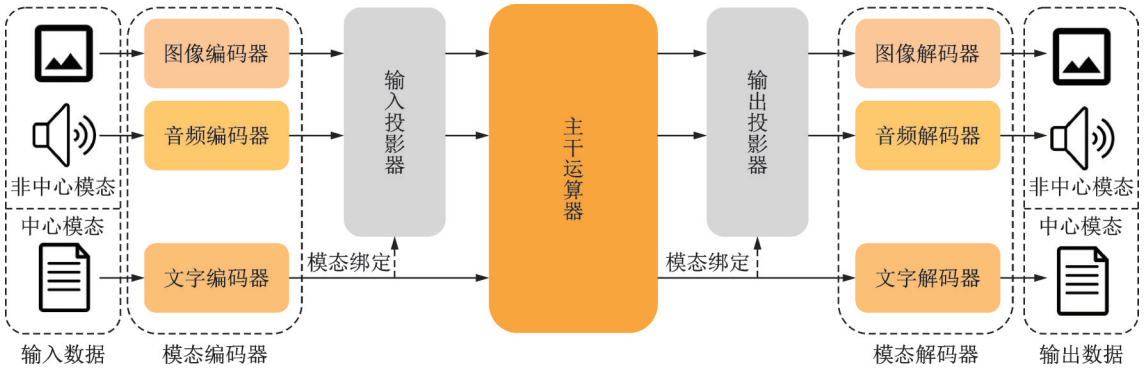

图1 多模态大模型的框架

Fig.1 Framework of multimodal foundation models

对于多模态大模型而言,定义所有输入模态的集合为M。一般而言,多模态大模型具有一个中心模态C。通过模态对齐技术,多模态大模型将其能够处理的所有模态都投影到该中心模态上。下文将给出多模态大模型的5个模块及各模块的输入、输出数据的形式定义,作为本文阐述大模型架构的理论框架。

模态编码器(Modality encoder,ME)负责将某一输入模态X的数据D_X编码成该模态域下的特征向量F_X

$$F_X = \mathrm{ME}_X(D_X) \quad X \in M \tag{1}$$

输入投影器(Input projector,IP)负责将某一模态X的特征向量F_X投影为中心模态C的特征向量F_C

$$F_C = \mathrm{IP}_{XC}(F_X) \quad X, C \in M \tag{2}$$

主干运算器(Backbone calculator,BC)负责对中心模态C的特征向量F_C进行运算,得到例如推理、生成等运算的结果$\hat{F}_C$

$$\hat{F}_C = \mathrm{BC}_C(F_C) \quad C \in M \tag{3}$$

输出投影器(Output projector,OP)负责将中心模态C的特征向量$\hat{F}_C$投影为某一模态X的特征向量$\hat{F}_X$

$$\hat{F}_X = \mathrm{OP}_{CX}\left(\hat{F}_C\right) \quad X, C \in M \tag{4}$$

模态解码器(Modality decoder,MD)负责将输出模态X的特征向量$\hat{F}_X$解码至模态X的原始数据域,解码后的数据结果为$\hat{D}_X$

$$\hat{D}_X = \mathrm{MD}_X\left(\hat{F}_X\right) \quad X \in M \tag{5}$$

垂直领域大模型的定制过程即根据业务需求选取所需要的模块(未必包括所有模块)组成业务模型,然后训练整个模型。其中个别模块可通过迁移和微调开源模型的方式部署实现。

1.2 特征提取

特征提取是从原始数据中提炼出具有代表性的特征,以助于完成特定任务的过程。在机器学习领域,特别是深度学习领域中,特征提取是至关重要的环节。由于原始数据往往包含大量冗余和噪声信息,通过特征提取能将数据转换至信息更为密集的特征空间,从而助力模型更有效地理解数据结构和模式。

在深度学习中,可以利用神经网络自动提取特征。神经网络模型能够端到端地从原始数据中学习特征,无需人工干预。这种特征提取方法易受泛化性问题影响,通常需要大量数据和计算资源以确保良好的性能。神经网络的每一层都将上一层的输入数据进行计算并转换到一个新的向量空间,这种设计允许灵活地定义每一层的输出维度,而无需详细说明这些转换过程。自编码器(Autoencoder)利用了这些优良特性,其目标是通过最小化原向量与重构向量之间的重构误差,学习数据的有效表示。自编码器通过将输入数据压缩成低维特征向量,再通过解码器将这些低维表示投影回原始数据空间,其结构如图2所示。本文所述的大模型架构中的模态编码器和模态解码器的一种重要的构建思路,就是将二者分别对应自编码器的编码和解码部分,配对成自编码器来进行训练。自编码器中的一种变种是变分自编码器(Variational autoencoder),其目标除了最小化重建误差外,还包括最大化输入数据的似然概率,从而学习到压缩向量的分布。例如图像模态的VQGAN[38]就是以这种方式构建的,它也被广泛应用于后来出现的图像生成模型中[39-40]。

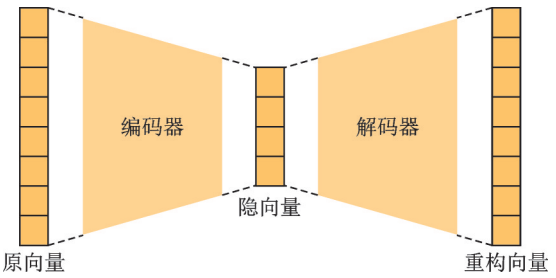

图2 自编码器架构

Fig.2 Structure of autoencoder

1.3 模态对齐

在单模态大模型的工作流程中,由于不涉及跨模态数据处理,因此其架构不包含输入投影器和输出投影器,而多模态大模型需要处理包括中心模态和非中心模态在内的多种模态数据。为了通过输入投影器和输出投影器实现模态间的数据转换,关键在于运用模态对齐(Modality alignment)技术。模态对齐的目标是将不同模态的原始数据或特征向量处理成具有相同维度的特征表示,然后通过设计损失函数来表征特征向量间的相关性,进而将各模态的特征向量投影到一个共享的特征空间中。在理想情况下,模态对齐应确保携带相同语义信息的不同模态原始数据在目标特征空间中被表示为同一点,从而便于实现跨模态信息转换。

模态对齐主要有两种架构实现方式:融合编码器架构和双编码器架构[31]:

(1) 融合编码器架构。融合编码器(Fusion encoder)架构采用Transformer模型[41]的自注意力机制(Self-attention)或交叉注意力机制(Cross-attention)来编码多模态数据。注意力机制的计算公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

式中:查询向量(Query,$\boldsymbol{Q}$)、键向量(Key,$\boldsymbol{K}$)和值向量(Value,$\boldsymbol{V}$)都是基于输入向量产生的中间向量,d_k为$\boldsymbol{Q}$和$\boldsymbol{K}$的维数。基于自注意力机制的方法需要拼接主副模态的特征向量输入Transformer中产生$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$,让模型自动关注不同模态的特征,并实现跨模态信息融合。例如,VL-BERT[42]模型将文本和图像的特征向量拼接,利用Transformer的自注意力机制实现语言-视觉特征的聚合和对齐。而基于交叉注意力机制的方法则将两个模态的特征向量分别计算$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$,进而实现跨模态信息融合。例如,DiT模型[43]采用交叉注意力机制捕捉文本与图像间的相关性,实现了文本控制的图像生成。图3展示了这两种融合编码器架构。

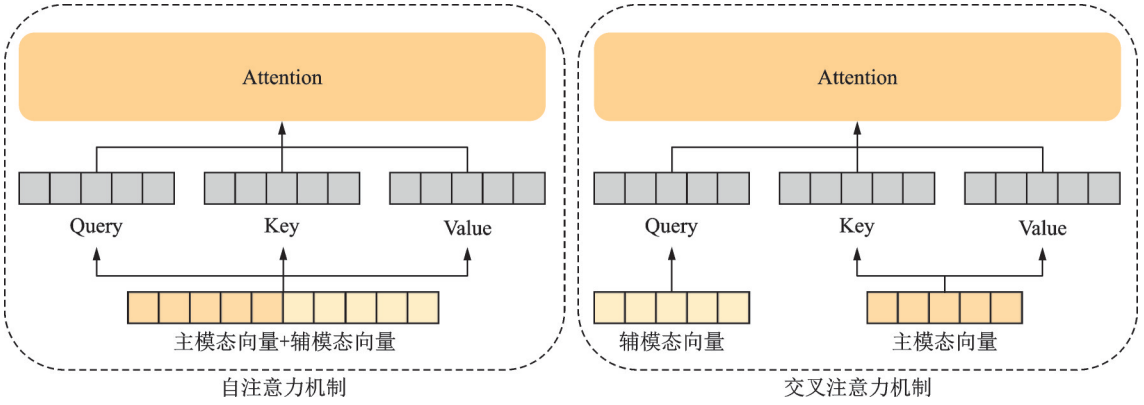

图3 融合编码器架构

Fig.3 Structure of fusion encoder

(2) 双编码器架构。双编码器(Dual encoder)架构是一种多模态学习策略,它为每种模态独立训练一个专门的编码器。该架构的核心理念在于利用对比学习的方法,通过语义相似度指标同步引导两个编码器的学习过程,以将不同编码器的输出特征向量投影到同一个向量空间中。具体而言,该模型基于一个假设:如果两个编码器输出的特征向量属于同一特征空间,那么具有配对标签的特征向量在向量空间中的距离应该较为接近,反之则相距较远。通过这种对齐方法,可以预期,描述相似对象或场景的不同模态编码器的输出结果将会足够接近,甚至在理想状态下,它们在特征空间中会汇聚于同一点。实现这一目标的关键在于构建合理的模型架构,并在大规模数据集上进行充分训练。图4展示了双编码器的处理思路。

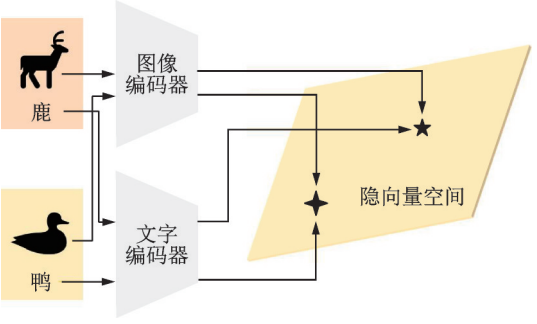

图4 双编码器架构

Fig.4 Structure of dual encoder

然而,一一对齐每对模态的成本将会非常高,要想获取每对模态都对齐的数据集也是一项挑战。为此,一些研究者提出了“桥接对齐”(Bridging alignment),或称“绑定对齐”(Binding alignment)的策略。这些方法通常通过将所有其他模态与一个中心模态进行匹配,进而在语义空间中实现所有模态的对齐。例如,ImageBind[26]将图像作为中心模态,而CoDi[20]则将文本作为中心模态,通过这种方式,它们有效地简化了多模态对齐的训练过程,并提高了模型的实用性和效率。

1.4 规模幂律

规模幂律(Scaling law),也称为规模定律或幂律定律,是指系统的某些性能随着规模扩大而变化的过程遵循一定的数学规律。在人工智能领域,特别是大模型的研究与应用中,规模幂律描述了模型性能如何随着模型规模(例如参数量、数据集规模、计算资源等)的扩展而变化的一系列规律和现象。它使用定量分析的方法揭示了大模型性能提升的内在机制。

在文献[44]中,作者探讨了不同模型的归纳偏置如何影响模型规模扩展与性能之间的关系。研究发现,模型架构确实是影响模型扩展收益的关键因素之一。该研究还指出,尽管普通Transformer架构可能并非始终能够取得最佳性能,但它却展现出了最佳的扩展能力。在计算机视觉领域[45]和自然语言处理领域[46]的研究中,基于Transformer架构的模型都显示出了模型规模与模型性能之间的指数级关系。

另一项研究[47]则考察了下游任务数量与模型规模对指令微调(Instruction-finetune)性能的影响。研究者采用了多任务联合训练的方法,在众多不同的任务上进行微调,使语言模型能够学习到更广泛的语言表示和知识,从而增强其在未见任务上的泛化能力。在联合训练过程中,通过参数共享促进了不同任务间的知识和技能迁移,显著提升了模型的泛化能力和性能。此外,联合训练还减少了单独训练每个任务所需的时间和资源,提高了训练效率。这种模型性能随任务多样性的增加而提高的现象便是一种规模幂律的体现。文献[48]的作者在其构建的大型基准测试集OPT-IML Bench上验证了模型性能随任务数量而增加的现象。另外,还有研究人员分别给出了自然语言模型[49]和各种模态的自回归生成式模型[50]在不同规模下的模型性能。

尽管规模幂律的定量表示没有一个统一的形式,但总体来说都可以表示为模型损失函数、模型可训练参数量、数据集大小,以及有计算资源等条件之间的指数关系。用模型训练的损失函数$L(\cdot)$表征模型性能,损失函数越小代表模型性能越好。式(7)描述了给定参数量的模型在足够大的数据集上训练至收敛时的性能,其中$L(N)$为损失函数,N为模型的可训练参数量,N_c为一个常数,α_N为幂律指数;式(8)给出了给定计算资源限制下,一个大小适当的模型在一个足够大的数据集上训练后的性能,其中$L(C)$为损失函数,C为给定的计算资源,C_c为一个常数,α_C为幂律指数;式(9)描述了大模型使用给定大小的数据集进行早停训练时的性能,其中$L(D)$为损失函数,D为数据集的大小(以token计),D_c为一个常数,α_D为幂律指数。

$$L(N)\propto\left(\frac{N_c}{N}\right)^{\alpha_N} \tag{7}$$

$$L(C)\propto\left(\frac{C_c}{C}\right)^{\alpha_C} \tag{8}$$

$$L(D)\propto\left(\frac{D_c}{D}\right)^{\alpha_D} \tag{9}$$

从式(7~9)可以发现,在其他条件给定的情况下,模型的损失函数随着参数量、计算资源和训练数据量的增加成指数级下降。这意味着通过增加模型参数量、加大计算资源投入和增加训练数据量,模型的性能也可以有指数级的提升。

1.5 涌现现象

规模幂律揭示了模型规模扩展可以带来模型性能的量变提升,而涌现现象是指随着模型的规模扩展达到临界点后,模型展现出新性质的现象,其中一种表现就是模型性能大幅提升[51]。涌现现象从质的维度揭示了大模型卓越性能的根源。在深度学习领域,尤其是在大语言模型领域,涌现现象被广泛观察到。例如,LLaMA等模型在多种语言任务中展现出了卓越的理解和生成能力,甚至在一定程度上具备了逻辑推理能力,但模型规模较小的语言模型却做不到这点,这就是一种模型的能力涌现现象。随着模型规模的增加,模型得以拥有更多的参数和更为复杂的结构,从而使得它能够捕捉数据中的复杂特征和模式。大模型通常展现出强大的泛化能力,即在训练集之外的数据上也能够有良好的表现,便是由于模型的大量参数能存储丰富的知识,使得它们能够在未见过的数据上也能进行精确的推断和预测,进而提供对不同任务的适应性和通用性,甚至使模型真正学习到隐藏在数据背后的原理和推理方式。在文献[51]中,作者指出不同任务和不同提示方式会影响大语言模型涌现现象的出现点。其中,采用思维链(Chain-of-thought)的提示方式能显著提升大语言模型处理复杂推理任务的能力[52],进而让涌现现象的出现点提前。

综上所述,在构建垂直领域大模型时,模型构建者需要根据可能遇到的下游任务需求和用户的提

示习惯来合理选择模型的参数规模。同时,随着模型参数量的增加,对计算资源的需求和过拟合风险也随之上升,因此不能无限制增加参数量。涌现现象不仅展示了大模型的优势,也揭示了部署模型时需要权衡的重点,对于模型设计和应用具有重要的指导意义。

2 垂直领域大模型的定制方法

将详细阐述如何从模态编码器、输入投影器、主干运算器、输出投影器以及模态解码器这5个关键模块中,根据垂直领域中的实际需求灵活选择并组合相应的模块来构建垂直领域大模型。此外还将分析具体案例,以便让读者更好地理解和应用所述的方法论。

可根据垂直领域大模型的定制化程度由低到高(换言之,借用通用大模型的程度由高到低)分为3类:基于全架构通用大模型的垂直领域增强、基于预训练模块的垂直领域大模型改造,以及无预训练模块的垂直领域大模型全架构构建。表1中对3种垂直领域大模型定制方法的特点进行了概括。

表1 垂直领域大模型的定制方法

Table 1 Customization methods of domain-specific foundation model

定制方法	定制化程度	定制难度	灵活性	算力需求
基于全架构通用大模型的垂直领域增强	低,仅定制了模型的领域知识输入方式	低	低	低
基于预训练模块的垂直领域大模型改造	中,部分模块自行构建,部分模块由迁移得来	中	中	中
无预训练模块的垂直领域大模型全架构构建	高,每个模块都可自定义构建	高	高	高

2.1 基于全架构通用大模型的垂直领域增强

通用大模型功能全面,适用于多种任务场景。若某通用大模型能够完全处理所需的数据模态,模型部署者就不必修改其架构,而只需对其进行垂直领域增强,从而实现垂直领域大模型的定制化。

根据垂直领域增强是否需要改变大模型参数,又可将其分为即插即用的垂直领域增强和基于微调的垂直领域增强两类。在表2中对基于全架构通用大模型的垂直领域增强方法进行了分类和概括。

表2 基于全架构通用大模型的垂直领域增强

Table 2 Specific-domain enhancement with the entire general-purpose foundation models

定制方法			是否修改大模型参数	是否加入新模块	领域知识提供方	具体技术
即插即用	调用已有知识	硬提示	否	否	部署者	PET[53]
		软提示	否	是	部署者	Prefix Tuning[54], P-tuning[55]
	输入新增知识	提示词	否	否	用户	LongRoPE[56], Transformer-XL[57]
		外挂知识库	否	是	部署者	RAG[58]
基于微调	基于适配器		是	是	部署者	Adapter[59], AdapterFusion[60], IA3[61]
	基于低秩矩阵分解		是	是	部署者	LoRA[62], LoHa[63], LoKr[64]
	全参数微调		是	否	部署者	PEFT[65]

2.1.1 即插即用的垂直领域增强

预训练大模型的通用性、泛化性和推理能力使其能够作为垂直领域大模型的主体。要想在不修改

大模型参数的条件下实现即插即用的垂直领域增强,可以通过调用已有知识或输入新增知识两种方式进行处理。调用已有知识的垂直领域增强旨在尽可能调用通用大模型中已经存储的垂直领域知识,在图5(a)展示了这种方式。而输入新增知识的垂直领域增强是指通过输入领域知识的方式赋予大模型对垂直领域任务的处理能力,这其中还可再分为通过提示词输入知识和通过外挂知识库输入知识两种路线,图5(b)和图5(c)中分别描述了这两种技术路线。

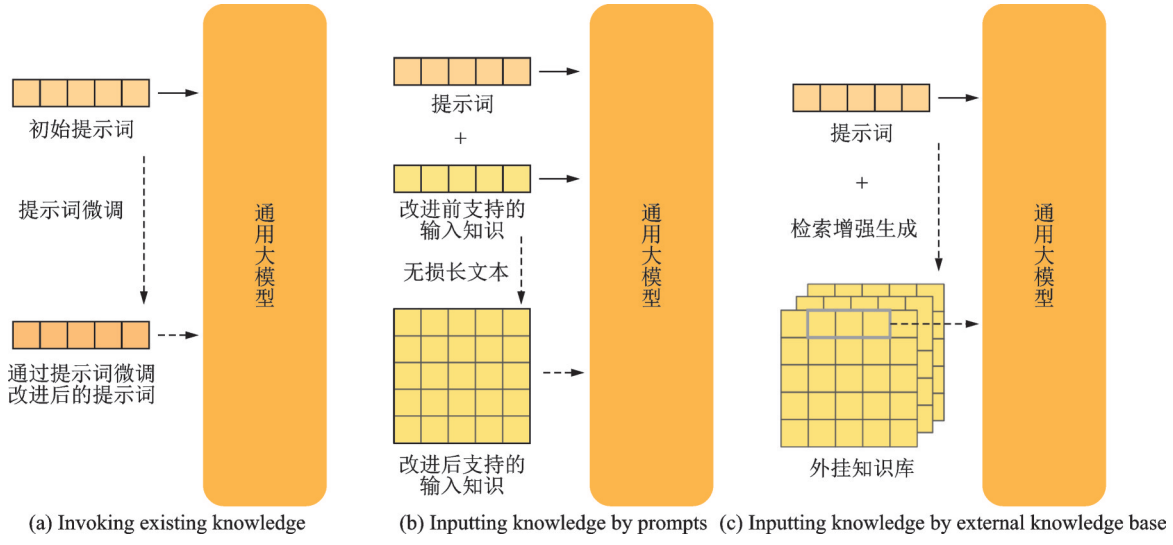

图5 即插即用的垂直领域增强

Fig.5 Plug-and-play domain-specific enhancement

(1) 调用已有知识的垂直领域增强。通用大模型在训练过程中,有可能已经接触过垂直领域的知识。提示词微调(Prompt tuning)能够通过改进提示词的方式,针对任务目标来更好地调用模型本身具有的垂直领域知识,其名字中“微调”二字的含义指的是对提示词的改进操作。具体而言,是将一段精心设置的提示词插入到输入数据前面作为模型上下文来影响生成的输出结果。这些精心设置的提示词可以是自然语言描述、示例、规则或者其他能够指导模型理解任务要求的文本或嵌入向量。模型在生成输出时会考虑这些精心设置的提示词,从而生成与任务相关的结果。提示词微调主要分为硬提示和软提示。

(a) 硬提示。硬提示(Hard prompt)方法在自然语言处理(NLP)中是一种常见的技术,它通过使用可解释和可重用的手工制作的单词和标记来指导语言模型的输出。硬提示通常是由人工设计并针对特定任务定制的,因此它们具有不易更改的特性。PET(Pattern exploiting training)[53]是一种经典的硬提示学习方法,它将问题建模成一个完形填空问题,然后优化最终的输出词。这种方法通过在少量监督数据上训练模型,并对无监督数据进行集成预测,从而实现对模型的指导。

(b) 软提示。硬提示在设计提示词时需要一定的实验探索和专业知识,并且人为设计的提示词不一定适合大模型的数据处理方式。为了简化这一过程并提高提示词微调的灵活性,研究者们提出了基于软提示的微调方法。前缀微调(Prefix tuning)[54]是软提示微调的一种形式,通过添加可学习的前缀向量(软提示词)到输入序列的开始部分来适应特定下游任务。这些前缀向量作为输入的一部分,引导模型的输出以符合任务要求。前缀微调的优势在于它只更新这些前缀向量,而不是模型的参数,从而大幅减少了计算资源和存储资源的需求,同时保留了预训练模型学习到的丰富知识。在前缀微调的基础上,研究者们又提出了P-tuning方法[55]。P-tuning使用可学习的软提示来替代固定或人工设计的单词

和标记,其核心思想是将提示词也视为模型可以学习的一部分,让模型不仅学习如何响应给定的任务,还学习如何生成最佳的提示词。这些软提示通常是一系列嵌入向量,它们在模型的输入端与实际的文本输入一起被处理。通过端到端的训练,模型自动学习到如何调整这些嵌入向量,以便更好地完成特定任务。P-tuning的优势在于它结合了前缀微调的参数效率和传统硬提示词微调的灵活性。软提示会给予模型更大的自由度来生成答案,一方面有机会产生更多样化的输出,但另一方面增加了生成不准确或不相关回答的风险。

(2) 输入新增知识的垂直领域增强。当通用大模型已有的知识不足以解决垂直领域任务时,便可以通过输入新增知识的方式引入问题背景信息,从而获得更高质量的输出结果。这种方法被称为输入新增知识的垂直领域增强。

(a) 通过提示词输入知识。提示词作为用户和大语言模型的直接接触途径,可以用来融入垂直领域的知识。然而通过提示词输入知识的做法存在一个明显的局限性:输入的领域知识量受到模型能够处理的最大提示词长度的限制。限制了大语言模型长文本输入能力的是Transformer本身的3个核心问题:

位置编码的局限性。Transformer模型通常通过正弦和余弦函数生成固定长度位置编码,这些编码对于序列中的每个位置都是唯一的。然而,当序列长度超过训练使用的最大长度时,模型将无法正确处理额外的文本,因为它无法为新位置生成有效的编码。

注意力机制的资源消耗。注意力机制是Transformer模型的核心,它允许模型计算序列中每个元素的注意力权重。但随着序列长度的增加,这种机制的计算复杂度和内存需求呈平方级增长,导致资源消耗巨大。

长距离依赖问题。Transformer在处理长序列时需要跨过大量的输入token,往往会遇到梯度消失或爆炸的问题,使得模型难以捕捉序列中相隔较远的元素之间的依赖关系。

针对上述问题,无损长文本(Lossless long text)技术应运而生。它旨在增强模型处理超出其标准输入长度限制的长文本的能力,能够支持用户将大量领域知识直接通过提示词输入到大语言模型中,作为上下文信息来实现垂直领域增强。无损长文本技术通过外推和内插两个方向拓展了大语言模型的长文本输入能力:

外推。外推(Extrapolation)是指通过扩展模型的上下文窗口,使其能够处理超出训练数据长度的新文本。这通常涉及到改进位置编码机制,以便模型能够理解和处理更长的序列。Longformer[66]通过结合局部注意力机制和全局注意力机制,有效地对外推长文本处理能力进行了扩展;BigBird[67]采用稀疏注意力机制和可逆层来外推模型的长序列数据处理能力;LongRoPE[56]则通过在自注意力中引入旋转变换来改进位置编码,使模型能处理长距离依赖,支持长达两百万tokens的输入而不影响计算效率。

内插。内插(Interpolation)是指在模型的现有序列长度能力范围内,通过调整和优化机制来提升对长文本的处理能力。这通常涉及对注意力机制的改进,以便模型能够更有效地处理长距离的信息。BERT模型[11]通过双向Transformer的预训练,增强了模型对文本的理解能力。XL-Net[68]通过置换语言模型和广义自回归预训练来增强模型的内部表示,改进了模型对长文本的处理能力。Transformer-XL[57]是一种改进后的Transformer模型,它通过引入循环机制来解决长文本处理中的梯度消失问题,从而允许模型在处理当前序列的同时,保留之前序列的信息,从而更好地理解和生成长文本内容。

(b) 通过外挂知识库输入知识。在实际应用场景中,用户可能无法提供足够的垂直领域知识来增强通用大模型。为解决这一问题,模型部署者可以外挂一个专门的垂直领域知识库来实现对通用大模型的垂直领域增强。这种方式能够让通用大模型在生成回答或执行任务时参照这个外挂的知识库,从而获得必要的领域信息和上下文,提供更加精准和有针对性的回答或解决方案。检索增强生成(Re-

trieval-augmented generation)[58]技术正是为实现这一目的而被开发的。检索增强生成技术旨在利用外挂文档库来增强语言模型的生成能力,而不需要对模型进行重新训练,特别适用于需要自定义动态知识库的任务中,如问答、文本摘要、事实核查等。检索增强生成技术的核心是在生成过程中引入一个能够在大型文档数据库中快速找到当前任务相关信息的检索组件。这个检索组件可以将模型的当前状态(例如问题的编码表示)投影到一个高维空间,并在这个空间中基于最近邻搜索算法搜索最近似的向量,从而找到最相似的文档。一旦检索到相关的文档,这些信息会被作为额外的上下文信息来辅助生成过程。检索增强生成技术的优势在于能够结合大语言模型的生成能力和外部检索系统所提供的知识,还避免了让用户自行提供领域知识。此外,由于外部知识库可以根据需要随时更换,检索增强生成技术还具有极高的灵活性和适应性。

虽然上述技术最初是为了实现大语言模型在垂直领域的增强而提出的,但它们的应用并不局限于语言模型。随着大模型领域的发展,这些技术有望被扩展到其他模态的大模型中。

2.1.2 基于微调的垂直领域增强

当即插即用的垂直领域增强技术难以实现,或是需要向通用大模型输入过多领域知识,必须对通用大模型进行深度改造时,可以转而采用基于微调的垂直领域增强策略。该策略在尽可能保留通用大模型预训练知识的同时,对其进行针对性的垂直领域增强,定制出所需的垂直领域大模型。

微调技术分为3种主要类型:基于适配器的微调、基于低秩矩阵分解的微调和全参数微调。图6分别描述了这3种技术路线。接下来,本文将按照微调所需的资源和复杂度从低到高进行排序,并对这3类技术进行详细阐述。

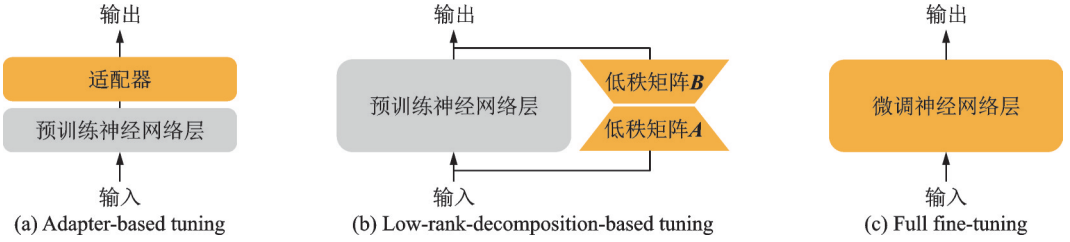

图6 基于微调的垂直领域增强

Fig.6 Fine-tuning-based domain-specific enhancement

(1) 基于适配器的微调。适配器微调(Adapter-based tuning)[59]是一种在预训练模型中插入小型可训练适配器模块的方法,旨在高效地使模型适应特定的下游任务。微调过程中,只有适配器模块的参数会被更新,而预训练模型的原有参数保持不变,从而减少计算资源和存储需求,并保留了模型在预训练阶段学习到的丰富知识。AdapterFusion[60]是一种适配器微调的扩展方法,它通过融合多个适配器模块,允许模型同时学习多个任务或适应多种不同的数据分布,而每个适配器模块可以专注于捕捉任务相关的特定特征。基于注入适配器的微调(IA3)[61]则通过在Transformer架构的注意力和前馈模块中注入学习向量来对激活层进行加权缩放。由于这些学习向量是微调过程中唯一可训练的参数,与传统的适配器微调相比,IA3显著减少了可训练参数的数量,从而降低了训练成本并提高了训练效率。此外,IA3不会增加推理延迟,因为其适配器权重可以与基础模型合并,同时保持了模型的灵活性和适用性,能够针对不同的任务和数据集进行定制化的微调。

(2) 基于低秩矩阵分解的微调。基于低秩矩阵分解的微调通过将预训练模型中的权重矩阵分解为低秩矩阵的乘积来减少需要更新的参数量。低秩矩阵分解能够捕捉权重矩阵中最重要的信息,在微调时保持原有预训练参数不变,只更新低秩分解矩阵,从而降低了微调过程中的计算和存储需求。这种

方法在提高微调效率的同时,还能保持或接近全参数微调的性能。

(a) 低秩适配。低秩适配(Low-rank adaptation, LoRA)[62]通过奇异值分解将模型参数分解为低秩矩阵的乘积,从而取得良好的微调性能。LoRA的原理可以用公式表示为

$$\boldsymbol{W}_{\text{new}} = \boldsymbol{W}_{\text{old}} + \Delta\boldsymbol{W} \tag{10}$$

式中:$\boldsymbol{W}_{\text{old}}$为原始权重矩阵,$\Delta\boldsymbol{W}$为低秩更新矩阵,它可以通过选取$\boldsymbol{W}_{\text{old}}$较小的奇异值对应的奇异向量来构造。对$\boldsymbol{W}_{\text{old}}$的奇异值分解则有

$$\Delta\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{11}$$

式中:$\boldsymbol{U}$和$\boldsymbol{V}$为从$\boldsymbol{W}_{\text{old}}$的SVD中得到的矩阵,而$\boldsymbol{\Sigma}$为一个对角矩阵,包含了$\boldsymbol{W}_{\text{old}}$的重要奇异值。通过只更新$\boldsymbol{U}$、$\boldsymbol{\Sigma}$和$\boldsymbol{V}$中的参数,LoRA实现了对模型的高效微调。

LoRA的局限性在于其通常对所有层应用相同的低秩结构,这便忽略了不同层、不同参数对下游任务的重要程度。自适应低秩适配(Adaptive low-rank adaptation, AdaLoRA)[69]是在LoRA基础上的一种改进方法,它可以自适应地决定哪些层的参数需要更新,通过自适应学习率和任务特定的参数调整策略,使得模型能够根据任务的特定需求自动调整微调的强度和范围。

有研究者还发现LoRA在某些大规模数据集上的持续预训练效果不佳,于是提出了分层重要性采样微调(Layerwise importance sampled AdamW, LISA)[70]策略,即不同层的权重范数分布呈现出不常见的偏斜性,LISA采用了重要性采样的策略,通过随机激活大模型中的不同层来进行优化。具体来说,LISA始终更新底层的embedding和顶层的linear head,同时随机更新少数中间的自注意力层。这种方法在内存消耗与LoRA相当的情况下,能够在多种下游微调任务中超越LoRA甚至全参数微调的性能。

(b) 低秩Hadamard积微调。低秩Hadamard积微调(Low-rank Hadamard product, LoHa)[63]通过引入低秩矩阵的Hadamard积来更新模型的权重。LoHa的原理可以用公式为

$$\boldsymbol{W}_{\text{new}} = \boldsymbol{W}_{\text{old}} \odot \Delta\boldsymbol{W} \tag{12}$$

式中更新矩阵$\Delta\boldsymbol{W}$可以进一步分解为两个低秩矩阵的Hadamard乘积,即

$$\Delta\boldsymbol{W} = \boldsymbol{L}_1 \odot \boldsymbol{L}_2 \tag{13}$$

式中:$\boldsymbol{L}_1$和$\boldsymbol{L}_2$为两个低秩矩阵,它们通过学习从输入数据中提取的关键信息来调整原始权重矩阵的元素。通过只更新$\boldsymbol{L}_1$和$\boldsymbol{L}_2$中的参数,LoHa实现了对模型的高效微调,同时保持了模型对新任务的适应性。

(c) 低秩Kronecker积微调。继LoHa之后出现的低秩Kronecker积微调(Low-rank Kronecker product, LoKr)[64]是另一种参数高效的微调方法。LoKr利用Kronecker积的特性来扩展权重矩阵的维度,同时保持参数数量的增加在可控范围内。Kronecker积允许模型在不同维度上学习复杂的交互,这对于捕捉输入数据中的高阶关系特别有用。LoKr的更新过程可以表示为

$$\boldsymbol{W}_{\text{new}} = \boldsymbol{W}_{\text{old}} + \Delta\boldsymbol{W} \tag{14}$$

式中:$\Delta\boldsymbol{W}$为两个低秩矩阵$\boldsymbol{L}_1$和$\boldsymbol{L}_2$的Kronecker积,即

$$\Delta\boldsymbol{W} = \boldsymbol{L}_1 \otimes \boldsymbol{L}_2 \tag{15}$$

式中:$\boldsymbol{L}_1$和$\boldsymbol{L}_2$通过Kronecker积计算出一个在微调过程中被更新的大矩阵。LoKr特别适合于那些需要增加模型维度以捕捉更复杂关系的任务,同时它还保持了与LoHa相似的参数效率。然而,LoKr可能需要更复杂的数学操作来处理Kronecker积,并且在某些情况下,它的计算成本可能会高于LoHa。

(3) 全参数微调。全参数微调(Full fine-tuning)不受限于预训练任务或数据分布,可以灵活适应各种不同的下游任务。让模型能够直接在最终任务的数据上进行端到端优化,而不需要额外的适配模块。但由于需要更新模型中的所有参数,全参数微调需要大量的计算资源和较长的训练时间。而大模型的参数量巨大,如果微调数据不足,可能出现过拟合的现象。此外,全参数微调过程中产生的中间变

量会占用大量显存空间。于是研究者们提出了上文所提到的许多参数高效的微调方法(Parameter-efficient fine-tuning)[65],这些方法在保持性能的同时,可以减少资源消耗和训练时间。

2.2 基于预训练模块的垂直领域大模型改造

大模型可能包含数百万甚至数十亿的参数,通过迁移学习(Transfer learning)可以减少需要训练的模型部分,从而显著降低训练开销。这种方法被称为基于预训练模块的垂直领域大模型改造。迁移学习的本质是在构建新模型时,利用预训练模型在学习过程中被固化在模型参数中的知识。如前所述,在大模型的架构中,通常包含5个主要模块:模态编码器、主干运算器、模态解码器、输入投影器和输出投影器。其中,模态编码器、主干运算器和模态解码器3个模块承载了大量的知识,因为它们直接参与到数据的编码、运算和解码过程中。相比之下,输入投影器和输出投影器本身承载的模型知识较少,在某些情况下,它们甚至都可能没有显式的模型负责这部分功能,或者在构建新的大模型时才训练这些模块。因此,在进行垂直领域大模型改造时,一般不会选择迁移输入投影器和输出投影器这两个模块。

接下来,本文将详细介绍如何基于预训练的模态编码器、主干运算器和模态解码器实现垂直领域大模型的改造。通过这种方式,可以有效地利用预训练模型的知识,同时减少计算资源的需求,使模型更加适合特定任务和环境。

2.2.1 基于预训练的模态编码器的迁移学习

预训练大模型往往在训练过程中适应了大量数据集的分布特征,其模型参数中已经内化了充足的领域知识,非常适合作为垂直领域大模型的特征提取模块。将预训练模型的前置特征提取模块作为领域数据的模态编码器,再在该模块后对接下游任务模块即可实现任务需求。模态编码器存储了关于数据关键特征的知识。有以下两种方式迁移得到模态编码器的方式:

(1) 同一模态迁移不同数据域。这种方式将模态编码器在源数据域上预训练得到的知识迁移到目标数据域。这通常涉及对源域和目标域数据的特征分布进行对齐。通过对源域的预训练模态编码器在目标域上进行微调,能够使其适应新数据的特性。具体而言,可以通过调整或添加编码器的最后几层来实现模型的迁移。此外,可以使用领域自适应技术,如领域对抗训练技术或领域不变特征提取技术,来减少源域和目标域之间的分布差异。

(2) 迁移到不同模态。在多模态大模型的研究中,经常会遇到某些模态缺乏对应预训练编码器的情况,此时可以采用跨模态迁移的策略,即借用其他模态的编码器来处理新的数据类型。例如,ImageBind的作者们将深度和热成像数据视为单通道图像的一种,从而利用图像编码器来提取这些数据的特征。在模型初始化时利用在图像数据集上预训练的权重,相较于随机初始化可以更快收敛,并且在一定程度上提升泛化性。

2.2.2 基于预训练的主干运算器的迁移学习

在多模态大模型中,主干运算器是核心的计算组件,负责处理经过编码的特征向量,并执行分类、生成等任务。垂直领域大模型的主干运算器可以从预训练模型中迁移而来,以利用预训练模型在大规模数据集上学习到的复杂特征处理和任务执行能力。这种方法避免了从头开始训练主干运算器,但仍需要构建相应的前置模块来将数据编码成主干运算器能够处理的特征向量。例如,NExT-GPT[27]就将各种模态的原始数据都转化为语言模态的特征向量后才能输入预训练的大语言模型,让大语言模型根据任务需求对输入token进行处理。有以下两种方式迁移预训练主干运算器:

(1) 迁移单一的预训练主干运算器。利用预训练的大模型(如LLaMA)作为主干运算器,处理中心模态的数据。迁移预训练的主干运算器到垂直领域应用时,通常需要对其进行微调,以适应特定领域的数据特征和任务需求。这一步可以在有限的领域数据集上进行,通过微调模型的参数来针对化地优化模型对领域数据的处理能力。

(2) 模块化组合多个预训练主干运算器。模块化组合是一种灵活的深度学习架构设计方法,它允许根据任务需求,将多个专门化的预训练模型整合到一个统一的框架中。混合专家(Mixture of experts, MoE)模型[71]可以作为一种有效的机制来进一步优化模块化组合。MoE模型通过引入多个专家网络,并使用门控机制(Gating mechanism)和混合策略(Mixing strategy)来动态地选择和组合这些专家的输出,从而实现对不同任务或数据子集的专业化处理。

门控机制的主要作用是决定输入数据应该如何在不同的专家之间分配。它根据输入数据的特征来为每个专家生成一个权重或者分数,这些权重或分数反映了每个专家处理当前输入数据的能力或适应性。门控机制的输出通常用于指导混合策略,告诉它每个专家对于当前输入的重要性;而混合策略的作用是将多个专家的输出按照一定的规则结合起来,生成最终的模型输出。混合策略可以是简单的,如平均或加权平均,也可以是复杂的,如基于模型输出的概率分布或其他高级方法。例如,在需要同时进行图像识别和语言理解的复杂任务中,若一个专家网络擅长识别图像中的物体边缘,而另一个专家网络擅长理解自然语言中的语义关系,则MoE模型的门控机制可以根据输入数据的特点和任务需求,自动调整每个专家网络的参与程度。这使得模型在处理视觉和语言的混合输入时,能够灵活地调用最合适的专家网络,以实现最佳性能。此外,MoE模型具有良好的扩展性,能够通过添加新的专家网络和更新门控机制来适应新的任务需求或数据类型,为构建灵活的垂直领域大模型提供了可能。

2.2.3 基于预训练的模态解码器的迁移学习

模态解码器在多模态大型预训练模型中起着至关重要的作用,它负责将经过处理的特征向量转换回原始数据的形式。在生成型任务中,例如将文本转换为图像或将音频转换为文本,模态解码器不仅需要精确地解码特征向量以重建可理解的原始数据,还需要展现出一定的创造性。一些预训练的模态解码器还能够理解和处理多模态特征输入,例如,CoDi-2能够利用文本和音频共同作为条件来控制图像的生成。通过迁移这类预训练的解码器,便无需从头开始训练复杂的解码器结构,能够直接将其应用于图像生成任务。

以下是有效利用预训练模态解码器进行迁移学习的方法:

(1) 微调预训练的模态解码器。与模态编码器类似,模态解码器也可以通过在特定任务的数据集上进行微调来适应新的任务需求。这个过程通常包括对解码器的最后几层进行调整,或者增加新的层来更好地捕捉特定领域的数据特征。

(2) 迁移跨模态生成式的模态解码器。在跨模态生成任务中,预训练的模态解码器可以直接用于生成目标模态的数据。首先通过条件编码器将条件信息编码为特征向量,然后与原始数据的特征向量结合,即可实现条件生成。实现此功能的前提在于确保输入的特征向量能够被解码器正确理解,而这可能涉及到对主干运算器和输出投影器的调整。

2.3 无预训练模块的垂直领域大模型全架构构建

当模型部署者无法通过迁移预训练模型的方法构建垂直领域大模型的模块时,就需要设计和训练对应模块了。首先从整体视角分析单模态和多模态大模型的架构,为后续构建大模型的各个模块奠定基础。

单模态大模型由模态编码器、主干运算器和模态解码器3个核心模块组成。以大语言模型LLaMA 2[15]为例,模态编码器和模态解码器专门针对语言模态,采用字节对编码(BPE)算法实现编解码功能。主干运算器则是一个庞大的自回归Transformer模型。该模型通过这3个模块实现了“输入原始文本-输入文本特征向量-输出文本特征向量-输出原始文本”的完整处理流程。此外,文献[17]引入了视觉句子的概念,并提出了能够通过视觉句子来自回归生成所需图像输出的大视觉模型LVM。该文献实现了在纯视觉模态下的上下文学习(In-context learning),使模型能够直接从图像模态的提示中推断任务

并生成相应结果。这项工作不仅探索了纯视觉输入的潜力,也为构建特定领域大模型提供了新视角:中心模态的选择不必局限于语言,任何在特定领域广泛使用的模态都可成为中心模态。

多模态大模型则需要额外引入输入投影器和输出投影器来实现模态对齐。例如,CoDi-2[21]迁移使用了ImageBind[26]中提出的对齐到图像模态的多个模态编码器处理相应模态的输入数据,然后通过一个多层感知机(MLP),将图像模态的特征向量转换到语言模态的特征空间。它以大语言模型LLaMA-2-7b-chat-hf的预训练自回归Transformer作为主干运算器的基础,然后将主干运算器处理后的图像和音频特征经过MLP转换回图像域,作为控制向量输入到基于Diffusion架构的生成模型中,得到最终的图像和文本结果。训练过程结合了文本生成损失、模态转化损失和数据生成损失,端到端地训练主干运算器的多模态特征处理能力以及两个MLP的模态转换能力。此模型的模态对齐体现在两方面,一方面是通过ImageBind的预训练模态编码器将多个模态的特征向量统一对齐到了图像模态,另一方面是通过MLP实现的图像特征向量与文本特征向量间的转换。

综上所述,构建大模型首先要确定数据模态,并从中选择中心模态。接着,构建相应的模块以实现模态编码器和输入投影器的功能,将不同模态的原始数据转换为主干运算器能够处理的中心模态特征向量。随后,设计输出投影器和模态解码器模块,将主干运算器处理后的特征向量转换为各模态的原始数据形式。完成模型结构设计后,便可开始训练过程。后续内容将详细介绍各模块的实现原理以及构建方法。

2.3.1 构建模态编码器

构建特定模态的编码器就是设计一个能够从数据中提取特征向量的神经网络结构。以下是构建模态编码器的一般步骤:

(1) 预处理为合适的数据结构。根据数据模态的特性选择合适的数据结构供后续模型使用。例如,在音频处理中,常见的做法是将时域信号转换为频谱图,然后利用适用于图像的神经网络结构进行特征提取。但音频信号既可以被表征成时序向量,又可以被表征成波形图像,具体选择哪一种数据结构实际上取决于任务需求和处理难度。对于推荐系统的输入而言,常建立图结构来表征用户和物品之间的关系。在选择目标数据结构时,研究者需要在任务需求和处理难度之间做出权衡,确保数据结构既能充分表征领域知识,又适合下游模型处理。另外,由于垂直领域大模型需要具有功能上的通用性,选择目标数据结构时还需要额外考虑多种任务输入之间的适配性。

(2) 设计网络架构。根据输入数据结构的特点设计相应的网络架构。例如,文本数据可利用Transformer架构来捕捉长距离依赖关系,而图像数据则可以采用基于CNN或ViT架构的模型来提取特征。

(3) 训练模态编码器。使用样本数量和种类都充足的数据集对模态编码器进行预训练,使其学习到模态数据的一般特征和分布。预训练是向模型灌输知识的过程,如果数据集的大小或多样性不足,模型都可能无法学习到完整的模态数据表示。一种训练模态编码器的方法是,将模态编码器和模态解码器组合成自编码器,以最小化重构误差为目标进行无监督训练。另一种训练方法是,针对特定任务设计模型,使用该任务的损失函数进行有监督训练,训练完成后将模型的上游部分迁移作为模态编码器。然而,这种训练方式无法得到与之配套的模态解码器,这可能影响后续模块的设计和功能。因此,在设计模态编码器时还需要考虑到整个大模型架构的一致性。

2.3.2 构建输入投影器

输入投影器的作用是将来自不同模态的数据投影到一个共同的特征空间中。正如1.3节中所讨论的,模态对齐可以通过融合编码器或双编码器两种架构实现。构建输入投影器时,关键在于选择是采用桥接器策略来整合不同模态的输入向量,还是通过微调方法使不同模态的投影器相互靠近,这两种

策略分别对应融合编码器和双编码器的理念。在训练过程中使用多模态理解任务的损失函数,如多模态分类或生成任务的损失,来训练模型的跨模态投影能力。此外,也可以采用端到端的训练方式,在优化大模型整体的性能的同时学习跨模态投影。如前文所述,CoDi-2模型[21]利用了ImageBind[26]中通过CLIP对齐的编码器作为图像、音频模态编码器和部分输入投影器,在其后结合了一个MLP作为另一部分的输入投影器,在端到端训练大模型的过程中优化了MLP,从而实现了从图像、音频对齐到文本的效果。

2.3.3 构建主干运算器

主干运算器负责对中心模态的特征向量进行理解和生成。要想构建一个针对垂直领域的主干运算器,首先需要选定该领域中最普遍和最能承载领域信息的数据模态,以此作为主干运算器处理的模态,并基于此设计模型架构。目前主流的模型架构都基于Transformer,而完整的Transformer模型由编码器和解码器两部分组成,其中编码器负责分析输入数据,提取出紧凑的特征表示,解码器利用这些特征表示来得到输出内容。由于二者结构不同,一般而言,编码器的理解能力更为强大,而解码器则拥有更强的生成能力。基于Transformer的大模型主干运算器可以采用不同的架构形式,包括编码器(encoder-only)架构,解码器(decoder-only)架构,以及编解码器(encoder-decoder)架构。表3中对以上3种架构的特点进行了概括。

表3 编码器架构、解码器架构和编解码器架构的对比

Table 3 Comparison among encoder-only, decoder-only and encoder-decoder structures

模型架构	生成能力	理解能力	计算量	模型示例
编码器架构	弱	强	低	BERT[11]
解码器架构	强	弱	低	GPT系列[7-10,23] LLaMA系列[14-15]
编解码器架构	强	强	高	BART[18] T5[19]

(1) 基于编码器架构的主干运算器。编码器架构只包含Transformer的编码器部分,通常用于需要理解输入文本,而不是生成新的文本序列的任务,如文本分类、情感分析等。由于只包含编码器部分,编码器模型结构相对简单,但也只能产生固定长度的输出,生成能力较弱。在生成任务方面,基于编码器架构的主干运算器仅能处理缺失序列补全这种广义上的生成任务。BERT[11]就是一个著名的编码器架构的例子。

(2) 基于解码器架构的主干运算器。在解码器架构中,解码器直接处理输入序列并生成输出序列,而没有专门的编码器来将输入序列加工成紧凑的特征表示。这一方面减少了参数量和计算开销,但另一方面也导致其对输入序列的理解难度会更大,从而限制了模型的长序列处理能力。这种架构在生成输出序列时不需要显式的上下文表示,而是通过自注意力机制在序列内部自动捕捉信息。基于解码器架构的主干运算器通常通过自回归生成的方式,即根据先前的生成内容逐个生成词或字符,来完成序列文本生成任务。像GPT系列[7-10,23]和LLaMA系列[14-15]的大语言模型都属于解码器架构。

(3) 基于编解码器架构的主干运算器。编解码器架构能够同时拥有编码器的理解能力和解码器的生成能力,但这也导致模型的参数量和计算成本较高。如Meta的BART[18],Google的T5[19]模型都采用了这种架构。

2.3.4 构建输出投影器及模态解码器

模态解码器有生成式和判决式两种类型。生成式模态解码器能够在满足条件信息(Conditioning)

的前提下,生成高质量的数据样本。判决式模态解码器则拥有更精确的恢复能力,其侧重点在于根据输入向量准确地重建数据样本。当模态解码器使用基于生成式模型或判决式模型的构建方式时,也会影响输出投影器的设计,因此需要联合考虑这两个模块。图7(a)和图7(b)分别展示了生成式模态解码器和判决式模态解码器的运行过程。

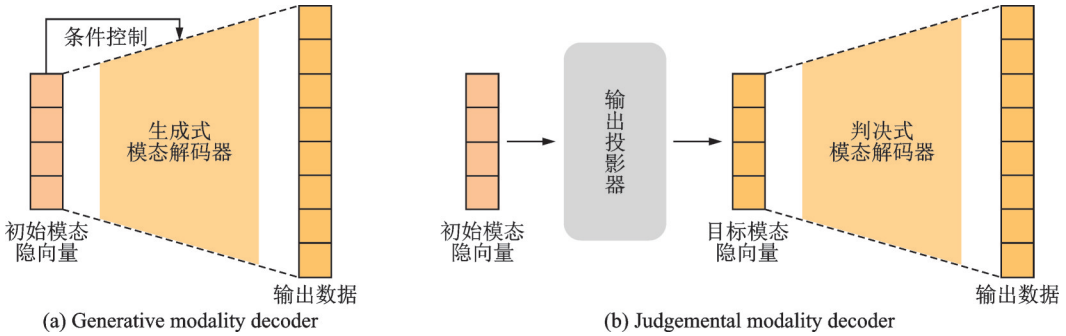

图7 输出投影器及模态解码器的运行过程

Fig.7 Running process of output projector and modality decoder

(1) 生成式模态解码器。生成式模态解码器采用生成式模型作为神经网络实现,能够利用条件信息控制生成过程。当将其他模态的特征向量作为条件信息,将生成的数据作为解码输出时,就可以称这样的生成式模型为生成式模态解码器。此类模态解码器不强制要求构建显式的输出投影器。如基于扩散模型的图像生成模型——DiT[40]能够根据前一步的结果和条件向量逐步生成图像。在该模型中,交叉注意力机制实现了输出投影器的功能。另一个用自注意机制实现输出投影器功能的例子是VAR[72]模型。通过在上下文输入中加入模态标签,使用自回归生成机制的VAR就知道了需要将哪些内容作为生成的控制向量,剩余的部分就是之前生成的内容,从而实现了自回归式图像生成。

生成式模态解码器通常采用端到端的训练策略。在训练过程中,模型的生成部分和模态交互部分同时进行优化。训练目标通常是最小化生成数据与真实数据之间的差异,并确保生成数据满足给定的条件。例如,如果模型的目标是根据文本描述生成图像,训练时会使用大量的文本-图像对,并通过比较生成图像和真实图像的相似度来优化模型参数。这种相似度可以通过像素级的损失函数(如均方误差)或更高级的感知损失(如VGG损失)来衡量。此外,还可以使用对抗性训练来提高生成质量。

(2) 判决式模态解码器。判决式模态解码器只负责直接将特征向量恢复成原始数据的形式,因此需要配合显式的输出投影器将其他模态域上的特征向量转换到目标模态来使用。例如,将VQGAN[38]中的解码器部分作为多模态大模型的判决式模态解码器时,就需要先显式地构建一个输出投影器将其他模态域上的特征向量转到图像模态上,再通过解码器部分将特征向量解码为原始数据形式。此显式输出投影器通常采用监督学习的训练方式,即模型接收来自其他模态的特征向量作为输入,并学习将这些特征投影为目标模态的特征向量,通过最小化两个特征向量之间的误差来提升投影准确度。模态解码器则与模态编码器一起作为自编码器进行训练,通过最小化重构误差来提升重建性能。

3 垂直领域大模型的应用实例

随着人工智能技术的迅速发展,大模型作为一种强大的工具,已经在各种垂直领域展现出了广泛的应用前景。大模型不仅具备处理海量数据和复杂任务的能力,还能够通过深度学习和模式识别技术,为各行各业带来新的突破和创新。在通信、自动驾驶、数学、医疗、法律、艺术、金融等各个领域,大模型正逐渐成为推动行业进步和创新的重要引擎,为人类社会的发展注入了新的活力。

(1) 在通信领域,大模型有望被广泛应用于网络规划、网络性能优化、故障检测和预测、资源调度等方面[73]。例如,可以使用针对通信领域微调后的大语言模型来处理网络日志数据,对特定的网络问题进行建模和解决。此外,通信网络大模型通过利用时空关联和知识推理,有望识别并预防导致服务质量下降的体验问题,从而提升服务水平和缩短故障响应时间,为精细化的智能化实时网络优化奠定基础。在工业场景中,大模型的业务理解能力也有望助力优化信号传输和调度策略,提高业务效率[74]。在网络大模型的研究中,文献[75]提出的NetGPT架构有望成为实现通信网络内生智能的有效途径,而文献[76]则探讨了在构建通信大模型过程中可能遇到的挑战和问题。

(2) 在自动驾驶领域,大模型能在车辆的感知、决策和规划等多个关键环节中发挥着核心作用[77]。具体来说,自动驾驶中的感知任务涉及到对车辆周围环境的实时监测,包括其他车辆、行人、交通标志和道路状况等。大模型通过分析摄像头、雷达和激光雷达(LiDAR)等传感器收集的数据,能够识别和分类各种物体,并构建车辆周围的详细地图。这种高级的感知能力是实现安全自动驾驶的基础。在决策层面,大模型需要根据感知到的信息做出快速而准确的判断,如何避让障碍物、选择合适的行驶路径、以及在复杂的交通情况下做出最优的驾驶策略。例如,DriveGPT[78]这样的多模态大模型不仅能够处理视觉数据,还能够理解和回答语言模态的指令,如根据语音输入的目的地进行路径规划。另外,文献[79]提出的pFedLVM能够利用预训练大视觉模型的强大性能进行图像特征提取,作为后续任务的基础。

(3) 在数学推理领域,大模型可以被用于解决数学问题、证明定理和发现模式等任务。例如,可以利用预训练的语言模型来理解和解释数学公式,并利用微调技术对特定的数学推理问题进行求解。文献[80]提出的MAmmoTH模型将思维链(chain-of-thought)与思维编程(program-of-thought)进行混合,充分发挥了大语言模型的理解能力和编程语言的计算能力,实现了良好的数学推理表现。

(4) 在医疗领域,大模型在医疗领域的应用涵盖了疾病诊断与预测、个性化治疗、药物研发、医疗资源管理以及健康监测与预警等多个方面[81]。例如,通过分析医疗数据和模式,提高诊断准确性、优化治疗方案、加速药物研发过程,并且帮助医疗机构合理规划资源、提高服务效率,同时实现了患者健康状态的实时监测与预警。文献[82]中提到的华佗GPT模型,能够通过模拟医生的诊疗过程,为患者提供初步的医疗咨询和建议。该模型不仅可以减轻医生的工作负担,还能让患者在偏远地区或资源匮乏的环境中获得及时的医疗服务。

(5) 在法律领域,大模型能够对法律文书进行深入分析,识别出文本中的关键信息和法律概念,从而辅助律师和法律顾问进行更为精确的案件分析和法律咨询。例如,大模型可以自动识别合同中的条款,提取重要的法律元素,如义务、权利、条件和期限等,帮助律师快速理解文档内容并识别潜在的法律风险。此外,大模型还可以用于案件预测,通过分析历史案例和相关法律条文,预测案件的可能结果,为律师制定辩护策略提供数据支持。文献[83]提出的ChatLaw大模型则可以提供实时的法律咨询和解答服务,帮助非专业人士理解复杂的法律问题,甚至可以自动生成法律文书草稿,减轻律师的工作负担。此外,大模型还可以辅助进行法律研究,快速检索相关法律文献和判例,为法律论证提供坚实的依据。

(6) 在艺术领域,大模型的应用正在探索和改变着创意表达的方式和艺术生产的过程。大模型可以通过学习大量的艺术作品和创意概念,生成新颖的艺术作品、音乐、文学作品等,为艺术家提供创作灵感和创意支持。例如,可以使用生成式模型来生成艺术作品,并利用微调技术对生成的作品进行风格和内容的调整[84-85]。目前,视频生成领域的Sora[86]已经能够让用户使用文本控制生成的内容,生成栩栩如生的视频作品。

(7) 在金融领域,大模型能够涵盖风险管理、投资策略、市场预测、欺诈检测等多种任务,为金融机构和投资者提供了强大的工具来优化决策、降低风险和提高效率[87-88]。例如,使用大模型构建信用评分系统,来评估借款人的信用风险。这些模型通过分析借款人的历史信用记录、财务状况、债务水平等因素,预测借款人未来偿还贷款的能力,并据此决定是否批准贷款申请以及贷款利率。或者,使用大模型来考虑各种资产类别的历史表现、相关性、风险和预期收益率,以及历史市场数据、宏观经济指标、政治事件等因素,为投资者做出最佳的决策。

综上所述,大模型作为一种强大的人工智能工具,已经在各种垂直领域展现出了巨大的潜力和价值。随着技术的不断进步和应用场景的不断扩展,大模型将继续发挥重要作用,推动各行业的数字化转型和智能化发展。这些应用实例展示了大模型在不同领域的广泛应用,也强调了为实际需求选择合适的垂直领域大模型定制方法的重要性。

4 垂直领域大模型定制化的未来发展方向

大模型技术的发展已经取得了显著的成就,但随着技术的不断进步,新的挑战和问题也逐渐浮现。

4.1 数据方面的挑战

首先,垂直领域数据的获取和数据结构的建模是一个重要的挑战。大模型通常需要大量的高质量数据来进行训练,而在特定领域获取这些数据可能既昂贵又耗时。此外,隐私保护法规的加强使得数据的收集和使用受到更多限制。为了解决这一问题,未来的研究可以集中在开发新的数据采集和标注技术,以及利用合成数据和弱监督学习方法来减少对大量标注数据的依赖。这不仅能够降低成本,还能在保护隐私的前提下,有效地利用数据资源。另一方面,要想做好垂直领域的数据结构建模,则需要研究者和工作者们深入理解垂直领域的业务流程,提取关键业务数据,构建完善的数据预处理流程。

其次,多模态数据理解是另一个关键挑战。尽管现有的大模型在处理文本数据方面表现出色,但对图像、声音等其他模态的理解能力仍有待提高,更遑论垂直领域中可能出现的各种新数据模态。构建能够综合处理多种模态数据的统一模型,对于提高模型在多模态任务上的性能和泛化能力至关重要。这要求研究者和工作者不仅要深入理解不同模态数据的特点,还要探索有效的多模态融合和交互理解机制。

4.2 模型架构方面的挑战

在垂直领域大模型的架构设计方面,一个核心的挑战是如何构建能够有效捕捉和表达垂直领域深层语义的模型。这要求模型不仅要具备广泛的知识基础,还要能够理解和适应特定领域的知识和输入模态。针对垂直领域的大模型需要在架构上实现高度的模块化和可定制性,能根据特定应用场景进行调整和优化,以适应不同领域的数据特性和任务需求。另外,模型的可解释性在垂直领域也尤为重要。在设计架构时,研究者需要考虑如何构建模型以便使其决策过程和输出结果能够被领域专家和最终用户所理解和信任。这可能涉及开发新的模型机制、引入可解释的模型中间表示,或者设计可视化工具来展示模型的内部工作机制。

4.3 算力方面的挑战

算力资源也是大模型技术面临的一个重要挑战。训练和运行大模型需要巨大的计算资源,这不仅增加了经济成本,还可能对环境造成影响。因此,研究如何提高模型训练和推理的效率,以及如何减少能源消耗,成为了一个迫切需要解决的问题。未来的研究方向可能包括开发更高效的模型压缩和加速技术,如知识蒸馏、模型剪枝、量化等,以及探索更高效的训练算法和专用硬件设计。

轻量化部署是大模型技术的另一个重要方向。大模型的体积和计算需求往往使得它们难以在移

动设备和边缘计算场景中部署。为了使大模型能够在资源受限的场景中运行,需要开发轻量级的模型架构和部署策略。这可能涉及到模型的简化、蒸馏和优化,以减少模型的大小和计算需求,同时保持或提高其性能。或者采用云边端协同的方法,将大模型的训练和推理过程拆分到不同层级的服务器上进行协同部署。

4.4 安全方面的挑战

最后,安全问题也是大模型技术必须面对的挑战。大模型可能被用于生成虚假信息、侵犯隐私或被恶意利用,同时模型本身也可能受到对抗性攻击。要想确保模型的安全性和可靠性,相关的工作包括但不限于以下几个关键点:首先,加强模型的鲁棒性,以抵御潜在的对抗性攻击,这可能涉及开发先进的对抗性训练技术,以及实施更为严格的数据清洗和预处理步骤;其次,开发和部署高效的恶意输入检测机制,利用异常检测算法和实时监控系统来识别和阻止恶意行为;再者,注重隐私保护,采用如差分隐私等技术减少模型对敏感数据的依赖,同时确保用户数据的安全和隐私。除了技术方面的努力,在工作流程和政策方面也需要增强安全性。比如,实施安全审计和认证流程,对模型的安全性进行全面评估,并确保模型的开发和部署符合伦理和法律标准,持续更新安全策略,以应对新兴的安全威胁和挑战。通过这些措施,可以为垂直领域大模型提供一个更加安全的运行环境,确保其在各个垂直领域的应用既高效又可靠。

总之,大模型技术虽然取得了显著的进步,但仍面临着多方面的挑战。解决这些挑战需要跨学科的合作和创新,包括数据科学、信息安全、伦理法律等领域的专业知识。通过这些努力,相信能够促进大模型技术的健康发展,推动其在更多领域的应用。

参考文献:

[1] 创新世界周刊编辑部. 垂直领域大模型[J]. 创新世界周刊, 2023(2): 110-111.
Innovation World Weekly Editorial Department. Vertical domain large model[J]. Innovation World Weekly, 2023(2): 110-111.

[2] 罗锦钊, 孙玉龙, 钱增志, 等. 人工智能大模型综述及展望[J]. 无线电工程, 2023, 53: 2461-2472.
LUO Jinzhao, SUN Yulong, QIAN Zengzhi, et al. Overview and prospects of artificial intelligence large models[J]. Radio Engineering, 2023, 53: 2461-2472.

[3] 孙柏林. 大模型评述[J]. 计算机仿真, 2024, 41: 1-7,24.
SUN Bailin. Review of large models[J]. Computer Simulation, 2024, 41: 1-7,24.

[4] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学: 信息科学, 2023, 53: 1645-1687.
CHE Wanxiang, DOU Zhicheng, FENG Yansong, et al. Natural language processing in the age of big models: Challenges, opportunities, and developments[J]. Chinese Science: Information Science, 2023, 53: 1645-1687.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J/OL]. (2015). https://arxiv.org/abs/1409.1556.

[6] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.

[7] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre- training[EB/OL]. (2018-04-03). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[8] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[9] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[10] BLOG O. Introducing ChatGPT[EB/OL]. (2022-11-30). https://openai.com/blog/chatgpt.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding

[J/OL]. (2018). https://arxiv.org/abs/1810.04805.

[12] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling[J/OL]. (2021). https://arxiv.org/abs/2103.10360.

[13] ZENG A, LIU X, DU Z, et al. GLM-130b: An open bilingual pre-trained model[J/OL]. (2022). https://arxiv.org/abs/2210.02414.

[14] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[J/OL]. (2023). https://arxiv.org/abs/2302.13971.

[15] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: Open foundation and fine-tuned chat models[J/OL]. (2023). https://arxiv.org/abs/2307.09288.

[16] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2020: 1691-1703.

[17] BAI Y, GENG X, MANGALAM K, et al. Sequential modeling enables scalable learning for large vision models[J/OL]. (2023). http://arxiv.org/abs/2312.00785.

[18] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J/OL]. (2019). https://arxiv.org/abs/1910.13461.

[19] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21: 1-67.

[20] TANG Z, YANG Z, ZHU C, et al. Any-to-any generation via composable diffusion[J/OL]. (2023). http://arxiv.org/abs/2305.11846.

[21] TANG Z, YANG Z, KHADEMI M, et al. CoDi-2: In-context, interleaved, and interactive any-to-any generation[J/OL]. (2023). http://arxiv.org/abs/2311.18775.

[22] ANTHROPIC. Introducing the next generation of Claude[EB/OL]. (2024-03-04). https://www.anthropic.com/news/claude-3-family.

[23] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[J/OL]. (2023). https://arxiv.org/abs/2303.08774.

[24] LIU H, LI C, WU Q, et al. Visual instruction tuning[J/OL]. (2023). http://arxiv.org/abs/2304.08485.

[25] FEI N, LU Z, GAO Y, et al. Towards artificial general intelligence via a multimodal foundation model[J]. Nature Communications, 2022, 13(1): 3094.

[26] GIRDHAR R, EL-NOUBY A, LIU Z, et al. ImageBind: One embedding space to bind them all[J/OL]. (2023). http://arxiv.org/abs/2305.05665.

[27] WU S, FEI H, QU L, et al. NExT-GPT: Any-to-any multimodal LLM[J/OL]. (2023). http://arxiv.org/abs/2309.05519.

[28] 蔡磊, 孟宪波, 韩冬梅, 等. 大模型在军事垂直领域的应用[J]. 舰船科学技术, 2024, 46(5): 171-175.
CAI Lei, MENG Xianbo, HAN Dongmei, et al. The application of large models in the military vertical field[J]. Ship Science and Technology, 2024, 46(5): 171-175.

[29] 张俊, 徐箭, 许沛东, 等. 人工智能大模型在电力系统运行控制中的应用综述及展望[J]. 武汉大学学报(工学版), 2023, 56(11): 1368-1379.
ZHANG Jun, XU Jian, XU Peidong, et al. A review and outlook on the application of artificial intelligence large models in power system operation control[J]. Journal of Wuhan University (Engineering Edition), 2023, 56(11): 1368-1379.

[30] 洪伟权, 朱莹莹. 高算力挑战下运营商应优先发展垂直行业 AIGC 模型[J]. 通信企业管理, 2023: 18-21.
HONG Weiquan, ZHU Yingying. Under the challenge of high computing power, operators should prioritize the development of vertical industry AIGC models[J]. Communication Enterprise Management, 2023: 18-21.

[31] DU Y, LIU Z, LI J, et al. A survey of vision-language pre-trained models[J/OL]. (2022). http://arxiv.org/abs/2202.10936.

[32] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[J/OL]. (2023). http://arxiv.org/abs/2306.13549.

[33] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J/OL]. (2023). http://arxiv.org/abs/2303.18223.

[34] ZHANG D, YU Y, LI C, et al. MM-LLMs: Recent advances in multimodal large language models[J/OL]. (2024). http://

arxiv.org/abs/2401.13601.

[35] CAO Y, LI S, LIU Y, et al. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT[J/OL]. (2023). http://arxiv.org/abs/2303.04226.

[36] ZHOU C, LI Q, LI C, et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT[J/OL]. (2023). http://arxiv.org/abs/2302.09419.

[37] MAO H, CHEN Z, TANG W, et al. Graph foundation models[J/OL]. (2024). http://arxiv.org/ abs/2402.02216.

[38] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 12868-12878.

[39] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 10674-10685.

[40] PEEBLES W, XIE S. Scalable diffusion models with transformers[J/OL]. (2023). http://arxiv. org/abs/2212.09748.

[41] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 1.

[42] SU W, ZHU X, CAO Y, et al. VL-Bert: Pre-training of generic visual-linguistic representations[J/OL]. (2019). https://arxiv.org/abs/1908.08530.

[43] PEEBLES W, XIE S. Scalable diffusion models with transformers[J/OL]. (2023). http://arxiv. org/abs/2212.09748.

[44] TAY Y, DEHGHANI M, ABNAR S, et al. Scaling laws vs model architectures: How does inductive bias influence scaling? [J/OL]. (2022). http://arxiv.org/abs/2207.10551.

[45] ZHAI X, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[J/OL]. (2022). http://arxiv. org/abs/2106.04560.

[46] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J/OL]. (2020). http://arxiv.org/abs/2001.08361.

[47] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

[48] IYER S, LIN X V, PASUNURU R, et al. Opt-IML: Scaling language model instruction meta learning through the lens of generalization[J/OL]. (2022). https://arxiv.org/abs/2212.12017.

[49] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J/OL]. (2020). https://arxiv.org/abs/2001.08361.

[50] HENIGHAN T, KAPLAN J, KATZ M, et al. Scaling laws for autoregressive generative modeling[J/OL]. (2020). https://arxiv.org/abs/2010.14701.

[51] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J/OL]. (2022). http://arxiv.org/abs/2206.07682.

[52] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of Conference and Workshop on Neural Information Processing Systems. [S.l.]: [s.n.], 2022, 35: 24824-24837.

[53] SCHICK T, SCHÜTZE H. Exploiting cloze questions for few shot text classification and natural language inference[J/OL]. (2020). https://arxiv.org/abs/2001.07676.

[54] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[J/OL]. (2021). https://arxiv. org/abs/2101.00190.

[55] LIU X, ZHENG Y, DU Z, et al. GPT understands, too[J/OL]. (2023-02-23). https://www.sciencedirect.com/science/article/pii/S2666651023000141.

[56] DING Y, ZHANG L L, ZHANG C, et al. Longrope: Extending LLM context window beyond 2 million tokens[J/OL]. (2024). https://arxiv.org/abs/2402.13753.

[57] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context[J/OL]. (2019). https://arxiv.org/abs/1901.02860.

[58] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[59] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of International Conference on Machine Learning. [S.l.]: [s.n.], 2019: 2790-2799.

[60] PFEIFFER J, KAMATH A, RÜCKLÉ A, et al. Adapterfusion: Non-destructive task composition for transfer learning[J/OL]. (2020). https://arxiv.org/abs/2005.00247.

[61] LIU H, TAM D, MUQEETH M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965.

[62] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[J/OL]. (2021). https://arxiv.org/abs/2106.09685.

[63] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[J/OL]. (2016). https://arxiv.org/abs/1610.04325.

[64] HACKBUSCH W, KHOROMSKIJ B N. Low-rank kronecker-product approximation to multi-dimensional nonlocal operators. Part I. Separable approximation of multi-variate functions[J]. Computing, 2006, 76: 177-202.

[65] DING N, QIN Y, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.

[66] BELTAGY I, PETERS M E, COHAN A. Longformer: The long-document transformer[J/OL]. (2020). https://arxiv.org/abs/2004.05150.

[67] ZAHEER M, GURUGANESH G, DUBEY K A, et al. BigBird: Transformers for longer sequences[J]. Advances in Neural Information Processing Systems, 2020, 33: 17283-17297.

[68] YANG Z, DAI Z, YANG Y, et al. XL-Net: Generalized autoregressive pretraining for language understanding[J]. Advances in Neural Information Processing Systems, 2019, 50: 3250.

[69] ZHANG Q, CHEN M, BUKHARIN A, et al. Adaptive budget allocation for parameter-efficient fine-tuning[C]//Proceedings of the Eleventh International Conference on Learning Representations. [S.l.]: [s.n.], 2022: 10512.

[70] PAN R, LIU X, DIAO S, et al. LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning [J/OL]. (2024). https://arxiv.org/abs/2403.17919.

[71] MA J, ZHAO Z, YI X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [S.l.]: ACM, 2018: 1930-1939.

[72] TIAN K, JIANG Y, YUAN Z, et al. Visual autoregressive modeling: Scalable image generation via next-scale prediction[J/OL]. (2024). http://arxiv.org/abs/2404.02905.

[73] 中国电信. 网络大模型白皮书 (2023)[R]. 北京: 中国电信, 2023.
China Telecom. White paper on large network models (2023)[R]. Beijing: China Telecom, 2023.

[74] 徐东兵, 刘瑞宏. 通信网络大模型构建思路[J]. 通信世界, 2024. DOI: 10.13571/j.cnki.cww.2024.02.014.
XU Dongbing, LIU Ruihong. Construction ideas for large-scale communication network Models[J]. Communications World, 2024. DOI: 10.13571/j.cnki.cww.2024.02.014.

[75] CHEN Y, LI R, ZHAO Z, et al. NetGPT: An AI-native network architecture for provisioning beyond personalized generative services[J]. IEEE Network, 2024, 3: 1.

[76] TONG W, PENG C, YANG T, et al. Ten issues of NetGPT[J/OL]. (2023). https://arxiv.org/abs/2311.13106.

[77] 朱冰, 贾士政, 赵健, 等. 自动驾驶车辆决策与规划研究综述[J]. 中国公路学报, 2024, 37: 215-240.
ZHU Bing, JIA Shizheng, ZHAO Jian, et al. A survey on decision-making and planning for autonomous vehicles[J]. China Journal of Highway and Transport, 2024, 37: 215-240.

[78] XU Z, ZHANG Y, XIE E, et al. DriveGPT4: Interpretable end-to-end autonomous driving via large language model[J/OL]. (2023). https://arxiv.org/abs/2310.01412.

[79] KOU W B, LIN Q, TANG M, et al. PFedLVM: A large vision model (LVM)-driven and latent feature-based personalized

federated learning framework in autonomous driving[J/OL]. (2024). https://arxiv.2405.04146V1.

[80] YUE X, QU X, ZHANG G, et al. MAmmoTH: Building math generalist models through hybrid instruction tuning[J/OL]. (2023). https://arxiv.org/abs/2309.05653.

[81] 郭华源, 刘盼, 卢若谷, 等. 人工智能大模型医学应用研究[J]. 中国科学: 生命科学, 2024, 54: 482-506.
GUO Huayuan, LIU Pan, LU Ruogu, et al. Research on the medical application of large-scale artificial intelligence models[J]. Science China: Life Sciences, 2024, 54: 482-506.

[82] ZHANG H, CHEN J, JIANG F, et al. HuatuoGPT, towards taming language model to be a doctor[J/OL]. (2023). https://arxiv.org/abs/2305.15075.

[83] CUI J, LI Z, YAN Y, et al. ChatLaw: Open-source legal large language model with integrated external knowledge bases[J/OL]. (2023). https://arxiv.org/abs/2306.16092.

[84] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J/OL]. (2020). http://arxiv.org/abs/2006.11239.

[85] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 3813-3824.

[86] BROOKS T, PEEBLES B, HOLMES C, et al. Video generation models as world simulators[EB/OL]. (2024-02-15). https://openai.com/research/video-generation-models-as-world-simulators.

[87] 曾晨光, 杨蕊菱, 王宇鹏, 等. 金融行业中的大语言模型[J]. 数字经济, 2023(11): 64-67.
ZENG Chenguang, YANG Ruiling, WANG Yupeng, et al. Large language models in the financial industry[J]. Digital Economy, 2023(11): 64-67.

[88] WU S, IRSOY O, LU S, et al. BloombergGPT: A large language model for finance[J/OL]. (2023). https://arxiv.org/abs/2303.17564.

作者简介:

陈浩泷(2000-),男,硕士研究生,研究方向:无线通信中的机器学习、多模态大模型,E-mail:haolongchen1@link.cuhk.edu.cn。

陈罕之(1999-),男,博士研究生,研究方向:机器学习、多模态、数字孪生,E-mail: chen. hanzhi@outlook.com。

韩凯峰(1993-),男,高级工程师,研究方向:6G无线人工智能、通信感知一体化技术,E-mail:hankaifeng@caict.ac.cn。

朱光旭(1989-),**通信作者**,男,博士生导师,研究方向:边缘智能、无线通信中的机器学习、联邦学习、通感一体,E-mail:gxzhu@sribd.cn。

赵奕晨(1989-),男,工程师,研究方向:无线通信技术、标准,E-mail:zhaoyichen@cmdc.chinamobile.com。

杜滢(1978-),女,教授级高级工程师,研究方向:6G关键技术研发及国际标准化,E-mail:duying1@caict.ac.cn。

(编辑:夏道家)