中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)01-0001-24

论文引用格式:Zhang Y J, Zhang R Q, Zhou H J, Qi J, Yu Z F and Huang T J. 2025. Research status and development trends of vision foundation models. Journal of Image and Graphics, 30(01):0001-0024(张燚钧, 张润清, 周华健, 齐骥, 余肇飞, 黄铁军. 2025. 视觉基础模型研究现状与发展趋势. 中国图象图形学报, 30(01):0001-0024)[DOI:10.11834/jig.230911]

视觉基础模型研究现状与发展趋势

张燚钧[1],张润清[1],周华健[1],齐骥[1*],余肇飞[2],黄铁军[2]

1. 中移(苏州)软件技术有限公司平台产品部,苏州　215000; 2. 北京大学计算机学院,北京　100190

摘　要: 在计算机视觉领域,尽管传统的深度学习视觉模型在特定任务上表现出色,但它们对大量标注数据的高度依赖及在新场景下性能泛化的局限性,大大增加了使用成本并限制了模型的应用范围。近年来,以Transformer为核心的新型模型结构,特别是在自监督学习领域的应用,为解决这些挑战提供了新的解决方案。这些模型通常通过大规模数据预训练,在处理复杂视觉场景中展现出强大的泛化能力,其被广泛称为视觉基础模型。本文深入探讨了视觉基础模型的研究现状与未来发展趋势,并重点关注该领域的关键技术进展及其对未来计算机视觉的潜在影响。首先回顾和梳理了视觉基础模型的背景与发展历程,然后介绍了在这一发展历程中出现的关键模型基础结构,介绍并分析了构建视觉基础模型所采用的各类预训练任务的设计思路,并根据其特性对现有的视觉基础模型进行分类。同时,对不同类型视觉基础模型中的代表性工作进行了介绍,并整理了目前可用于视觉基础模型预训练的数据集。最后,对视觉基础模型的研究现状进行总结和思考,提出了目前存在的一些挑战,并展望未来可能的研究方向。

关键词: 基础模型;计算机视觉(CV);预训练模型;自监督学习;多任务学习

Research status and development trends of vision foundation models

Zhang Yijun[1], Zhang Runqing[1], Zhou Huajian[1], Qi Ji[1*], Yu Zhaofei[2], Huang Tiejun[2]

1. Platform Product Department, China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou　215000, China;

2. School of Computer Science, Peking University, Beijing　100190, China

Abstract: In the field of computer vision, traditional deep learning vision models have exhibited remarkable performance on specific tasks. However, their substantial dependency on large amounts of annotated data and limited capability in generalization across new scenes significantly elevate usage costs and restrict the application scope of these models. Recently, novel model architectures centered around the Transformer, particularly in the domain of self-supervised learning, have emerged as solutions to these challenges. These models, typically pre-trained on extensive datasets, demonstrate robust generalization capabilities in complex visual scenarios and are widely recognized as vision foundation models. This paper delves into the current research status and future trends of vision foundation models, with a focus on key technological advancements in this field and their potential impact on future developments in computer vision. The paper begins by reviewing and organizing the background and developmental history of vision foundation models, followed by an introduction to the key model structures that have emerged in this developmental trajectory. The article further introduces and analyzes the design philosophies of various pre-training tasks employed in constructing vision foundation models, categorizing

收稿日期:2024-01-07;**修回日期**:2024-08-13;**预印本日期**:2024-08-20

*通信作者:齐骥　qiji@cmss.chinamobile.com

基金项目:国家自然科学基金项目(62088102)

Supported by:National Natural Science Foundation of China(62088102)

the existing models based on their characteristics. Additionally, the paper presents representative works in different types of vision foundation models and compiles the currently available datasets for pre-training these models. Finally, the paper summarizes the current research status of vision foundation models, reflects on existing challenges, and anticipates potential future research directions. This paper offers an expansive examination of the landscape of visual foundation models, chronicling their evolution, current achievements, and charting a course for future research. It acknowledges the Transformative impact of deep learning on computer vision, shifting the paradigm from traditional computational methods to models that excel in specialized tasks. However, it also confronts the limitations of these models, particularly their narrow applicability and reliance on extensive, meticulously annotated datasets, which have elevated deployment costs and restricted versatility. In response, the emergence of Transformer-based architectures has instigated a paradigm shift, leading to the development of vision foundation models that are redefining the capabilities and breadth of applicability of computer vision systems. This paper provides a systematic review of these models, offering historical context that underscores the transition from traditional deep learning models to the current paradigm involving Transformer-based models. It delves into the core structures of these models, such as the Transformer and vision Transformer, discussing their architectural intricacies and the principles that underpin their design, enabling them to capture the complexities of visual data with nuance and accuracy. A pivotal contribution of this paper is the thorough analysis of pre-training tasks, which is foundational to the construction of robust vision foundation models. It categorizes these tasks on the basis of their design philosophies and their effectiveness in enabling models to learn rich feature representations from large-scale datasets, thereby enhancing their generalization capabilities across a myriad of computer vision tasks. This article primarily introduces various pre-training methods such as supervised learning, image contrastive learning, image-text contrastive learning, and masked image modeling. It analyzes the characteristics of these pre-training tasks, the representative works corresponding to each, as well as their applicable scenarios and potential directions for improvement. This paper introduces existing vision foundation models based on mixture of experts according to universal visual representation backbones, aligned with language modalities, and generative multi-task models. Specifically, the article first analyzes the background of each type of foundation model, the core ideas of the models, and the pioneering work. Then, it analyzes the representative works in the development of each type of foundation model. Finally, it provides a prospect based on the strengths and weaknesses of each method. The paper also evaluates existing vision foundation models, scrutinizing their characteristics, capabilities, and the datasets utilized for their pre-training, providing an in-depth analysis of their performance on a variety of tasks, including image classification, object detection, and semantic segmentation. This paper delves into the critical component of pre-training datasets that serve as the foundational resources for the development and refinement of visual foundation models. It presents a comprehensive overview of the extensive image datasets and the burgeoning realm of image-text datasets that are instrumental in the pre-training phase of these models. The discussion commences with the seminal ImageNet dataset, which has been crucial in computer vision research and the benchmark for evaluating model performance. The paper then outlines the expansive ImageNet-21k and the colossal JFT-300M/3B datasets, highlighting their scale and the implications of such magnitude on model training and generalization capabilities. The COCO and ADE20k datasets are examined for their role in tasks such as object detection and semantic segmentation, underlining their contribution to the diversity and complexity of pre-training data. The Object365 dataset is also acknowledged for its focus on open-world target detection, offering a rich resource for model exposure to a wide array of visual categories. In addition to image datasets, the paper underscores the importance of image-text datasets such as Conceptual Captions and Flickr30k, which are becoming increasingly vital for models that integrate multimodal inputs. These datasets provide the necessary linkage between visual content and textual descriptions, enabling models to develop a deeper understanding of the semantic context. The paper anticipates research directions such as establishing comprehensive benchmark evaluation systems, enhancing cross-modal capabilities, expanding the coverage of various visual tasks, leveraging structured knowledge bases for training, and developing model compression and optimization techniques to facilitate the deployment of these models in real-world scenarios. The ultimate goal is to develop visual foundation models that are more versatile and intelligent, capable of addressing complex visual problems in the real world. Reflecting on the challenges faced by the field, the paper identifies the pressing need for more efficient training algorithms, the development of better evaluation metrics, and the integration of multimodal data. This paper also

highlights several challenges, including the need for more unified model paradigms in computer vision, the development of effective performance improvement paths similar to the "scaling law" observed in NLP, and the necessity for new evaluation metrics that can assess models' cross-modal understanding and performance across a wide range of tasks. It anticipates future research directions, such as the modularization of vision foundation models to enhance their adaptability and the exploration of weakly supervised learning techniques, which aim to diminish reliance on large annotated datasets. One of the key contributions of this paper is the in-depth discussion on the real-world applications of vision foundation models. It explores their implications for tasks such as medical image analysis, autonomous driving, and surveillance, underscoring their transformative potential and profound possible impact on these domains. In conclusion, this paper synthesizes the current state of research on vision foundation models and outlines opportunities for future advancements. It emphasizes the importance of continued interdisciplinary research to unlock the full potential of vision foundation models and address the intricate challenges in the field of computer vision. The paper's advocacy for an interdisciplinary approach is underpinned by the belief that it will foster innovation and enable the development of models that are not only more efficient and adaptable but also capable of addressing complex and multifaceted problems in the real world.

Key words: foundation model; computer vision(CV); pre-training model; self-supervised learning; multi-task learning

0 引　言

自从2012年AlexNet在ImageNet图像数据集分类上展现出优秀识别能力后，深度学习方法已经在计算机视觉(computer vision，CV)领域取得了飞速发展。在很多基础计算机视觉任务中，如图像分类、目标检测和语义分割等领域，基于深度学习的卷积神经网络已经取得了超过以往方法的效果，并在工业领域广泛应用(He等，2016；Krizhevsky等，2017；Redmon等，2016；Redmon和Farhadi，2018；Russakovsky等，2015)。然而，传统的CV模型在现阶段的大规模行业，如视频安防、人脸识别等应用中也面临很多问题，一些观点认为传统CV模型已经发展到了瓶颈。在真实的应用场景中需要被关注的类别可能只占样本总数非常少的比例，导致模型训练效果不佳，即长尾问题。具体来说，在大规模行业应用中存在大量细分场景的任务需求，也称为长尾场景。算法开发人员需要针对每个细分场景单独训练、部署、维护场景专用模型，导致人工成本提高。并且大部分长尾场景可供训练数据少、数据收集成本高，往往导致模型效果难以达到预期。

2017年谷歌公司提出Transformer这一模型结构，这是一种基于自注意力机制的序列到序列模型(Vaswani等，2017)。得益于在处理长序列的全局信息依赖关系方面的强大能力，Transformer模型成功应用在机器翻译等领域并取得了突破。随后几年，基于Transformer中编码器(encoder)和解码器(decoder)的各类自然语言处理(natural language processing，NLP)基础模型相继出现。其中比较有代表性的是BERT(bidirectional encoder representations from Transformers)模型和GPT(generative pretraining)模型。BERT模型是由谷歌团队在2018年推出，采用了Transformer中的encoder模块双向编码上下文信息，通过在大规模无标注语料上的预训练提高了对语义的理解能力。同年，OpenAI团队提出GPT模型。不同于BERT模型，GPT模型采用Transformer中的decoder结构，同样在大规模语料上进行预训练(Radford和Narasimhan，2018)。GPT模型是一种生成式语言基础模型，语言的生成能力较强。这两个经典基础模型推出后，在NLP领域引发了基础模型的持续研究热潮，越来越多的NLP基础模型不断涌现。GPT模型的后续版本GPT-2、GPT-3在更大规模的数据上进行训练，参数量分别达到了15亿和1 750亿(Radford等，2019；Brown等，2020)。GPT-3已经可以生成人类难以区分的文本、编程语言。谷歌提出T5基础模型采用Transformer中的encoder-decoder结构，发布了6 000万参数到110亿参数的多个版本，在翻译、文本分类和摘要等多个任务取得了最佳结果(Raffel等，2023)。这些基础模型呈现出参数规模越来越大、模型能力越来越强的发展趋势。NLP基础模型的发展在ChatGPT发布之后达到了一个重要的里程碑。现在，NLP基础模型的模型参数规模已经达到百亿甚至千亿，并且具有强大的语言处理能力。经过大规模数据预训练的NLP基础模型有望取代单任务专有模型，成为统一NLP领域任务

的基础模型(Bommasani 等, 2022)。基于Transformer结构的NLP基础模型的发展也给予了CV领域研究者启发。Dosovitskiy等人(2021)提出了视觉Transformer(vision Transformer, ViT)模型。通过将图像切分成图像块(patch)作为Transformer结构的输入序列,ViT将Transformer这种结构的应用范围从NLP领域迁移到CV领域。

类似NLP领域通过扩大模型参数规模从而提升模型性能的方式,CV领域的研究者也通过扩大ViT模型参数规模试图提高视觉基础模型的性能上限。ViT原始论文中,作者提出8 600万、3亿、6亿3种参数规模的模型(Dosovitskiy 等, 2021)。随后的工作中,ViT-G模型和ViT-e模型分别将模型参数量提高到18亿和39亿(Chen 等, 2023; Dehghani 等, 2023)。2023年,谷歌研究人员提出一种异步并行线性操作方法提升了ViT的训练效率,将ViT模型的参数提高到220亿规模(Dehghani 等, 2023)。此外,一些基于Transformer在视觉领域的改进工作将这种模型的能力扩展到各种视觉领域下游任务。例如DeiT网络(Touvron 等, 2021)通过卷积神经网络(convolutional neural network, CNN)作为教师模型,首次实现了在中等规模图像数据集上训练得到能够有效表征图像的Transformer模型。Facebook研究人员通过将目标检测问题转换成集合预测问题提出了DETR(detection Transformer)模型(Carion 等, 2020),首次将Transformer结构应用到目标检测领域。一些工作也将Transformer迁移到图像分割领域,例如SETR网络和SegFormer网络(Xie 等, 2021; Zheng 等, 2021),这类方法将ViT作为图像编码器,设计不同类型的解码器进行图像上采样,保证分割任务的预测能力。

除了能够处理单一视觉任务的视觉基础模型,研究人员也在探索具有多视觉任务处理的统一视觉基础模型。混合专家架构(mixture of expert, MoE)是多任务统一视觉基础模型的实现路径之一(Fedus 等, 2022; Lepikhin 等, 2020)。谷歌公司提出的一种基于稀疏专家混合架构的视觉基础模型V-moe是这个方向的开创性工作。混合专家架构通过路由根据视觉任务选择对应需要激活的专家网络,只需激活部分网络,提高了网络处理任务的效率(Riquelme 等, 2021)。

受NLP领域基础模型掩码语言建模(masked language modeling, MLM)预训练方法的启发,一系列适用于CV领域的掩码图像建模(masked image modeling, MIM)相继由研究者提出。例如在ViT上进行图像预训练的方法MAE(masked autoencoders)(He 等, 2022)和BeiT(bert pretraining of image Transformers)(Bao 等, 2022)。这些图像自监督学习方法使在大量图像数据集上进行视觉基础模型预训练这一路线得以实现。通过在大规模图像数据集上的预训练,视觉主干网络获得了对于各类视觉场景中图像的特征提取能力,且对不同场景图像具有良好的泛化性。通常只需要在下游细分场景上进行少量数据的微调,就可以实现较高的下游任务性能。这类方法的代表性工作有BEiT系列模型和EVA系列模型等(Bao 等, 2022; Fang 等, 2023b, c; Liu 等, 2021, 2022; Peng 等, 2022b)。

总的来说,相对于NLP领域基础模型较为快速发展和清晰的路径,视觉基础模型研究领域仍处于发展的初期阶段。CV领域的研究人员提出了各类不同的关于视觉基础模型实现路径的方法论。各类新型的视觉基础模型仍在不断问世。尽管实现方法各异,视觉基础模型的最终发展目标是一致的:提供一个可以处理各类视觉任务的通用的基础模型。具体来说,在各类下游任务上只需要少量数据微调甚至不需要微调就可以获得优秀性能,从而解决CV领域的长尾场景、数据成本高和开发成本高等问题。

当前关于基础模型的综述主要集中在NLP领域,关于基础模型在视觉领域的发展,相关梳理工作较少。本文围绕视觉领域基础模型的发展进行介绍。本文首先对视觉基础模型的背景和发展历程进行回顾和梳理。然后介绍在视觉基础模型发展历程中的关键模型基础结构,并对构建基础模型的各种预训练任务进行介绍和分析;根据模型特性对目前各种视觉基础模型进行分类,并介绍其中的代表性工作;对目前可用于视觉基础模型预训练的数据集进行了整理。最后,本文对视觉基础模型的研究现状进行总结和思考,对相关问题进行了探讨。

1 关键模型结构

在CV领域,基础模型的结构设计对于实现高效、准确的信息处理至关重要。传统设计上,这些模型依赖于CNN来捕获图像数据中的空间层次结构

和局部特征。然而，随着深度学习技术的不断发展，特别是预训练大型模型在模式识别和数据表征方面的重要性日益凸显，模型结构的要求也在不断演化。在此背景下，Transformer模型的提出标志着一个重要的转折点。最初在NLP领域取得显著成功的Transformer，通过其独特的自注意力机制，使得模型能够有效捕获长距离依赖关系，并在处理序列化数据时显示出优越的性能。它的成功不仅促进了预训练大模型在NLP领域的广泛应用，也为其他模态的基础预训练模型发展提供了新的架构思路。

ViT是将Transformer模型应用于图像处理的一个重要里程碑。ViT实现了对图像特征的全局理解。这种方法在多个计算机视觉任务中显示出了卓越的性能，尤其是在图像分类、目标检测和图像分割等领域。ViT的成功不仅证明了Transformer模型在视觉领域的巨大潜力，也为未来的视觉基础模型设计提供了新的方向和灵感。

本节将分别对Transformer和ViT这两种视觉基础模型中的关键模型结构进行介绍。

1.1 Transformer

Transformer最初应用在NLP领域的序列到序列的自回归任务中（Vaswani等，2017）。Transformer的核心设计是多头自注意力机制和位置嵌入编码。Transformer利用自注意力机制更有效地处理序列数据。与之前的一些序列转换模型不同（Cho等，2014；Hochreiter和Schmidhuber，1997），Transformer尽管也采用了编码器—解码器结构，但是完全舍弃了递归和卷积操作，转而使用多头自注意力机制和逐点前馈网络。

原始的Transformer由相同层数的编码器和解码器组成，其中编码器和解码器分别堆叠。Transformer的基本单元是注意力机制。在注意力层中，输入序列$\boldsymbol{X}, \boldsymbol{Y}$通过权重矩阵$\boldsymbol{W}$被映射为3个矩阵，即

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{Q}, \quad \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}^{K}, \quad \boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}^{V} \tag{1}$$

式中，$\boldsymbol{X} \in \mathbf{R}^{n_x \times d_x}$，$\boldsymbol{Y} \in \mathbf{R}^{n_y \times d_y}$，可以分别理解为查询$\boldsymbol{Q}$、键$\boldsymbol{K}$、值$\boldsymbol{V}$，其中

$$\boldsymbol{Q} \in \mathbf{R}^{n_x \times d_k}, \boldsymbol{K} \in \mathbf{R}^{n_y \times d_k}, \boldsymbol{V} \in \mathbf{R}^{n_y \times d_v} \tag{2}$$

Attention层的计算为

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

可以认为将$n \times d_k$的序列$\boldsymbol{Q}$编码得到一个新的序列，维度为$n \times d_v$。其中，$\sqrt{d_k}$是缩放因子，控制内积的大小。当$\boldsymbol{X} = \boldsymbol{Y}$时，式（3）描述的就是自注意力机制。

为了扩大注意力的特征表达能力，Transformer还使用“多头自注意力机制”。具体来说，单头注意力计算为

$$head_i = Attention(\boldsymbol{X}\boldsymbol{W}^{Q_i}, \boldsymbol{X}\boldsymbol{W}^{K_i}, \boldsymbol{X}\boldsymbol{W}^{V_i}) \tag{4}$$

式中，$i = 1, \cdots, h$。

多头自注意力机制计算为

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, head_2, \cdots, head_h)\boldsymbol{W}^{O} \tag{5}$$

式中，$\boldsymbol{W}^{O}$表示输出映射矩阵，具体为

$$\boldsymbol{W}^{O} \in \mathbf{R}^{hd_v \times d_{\mathrm{model}}} \tag{6}$$

多头自注意力机制将独立的注意力的计算结果拼接起来，丰富了模型特征子空间的多样性。在Transformer随后的处理流程中，多头自注意力操作后的输出将会经过两层连续的前馈层进行处理。

仅通过自注意力机制是无法表达模型输入序列的顺序的。Transformer设计了位置编码机制对输入序列中的每个位置进行有效编码。Transformer中使用的位置编码构造式为

$$PE_{(pos,i)} = \begin{cases} \sin(pos \times \omega_k) & i = 2k \\ \cos(pos \times \omega_k) & i = 2k + 1 \end{cases} \tag{7}$$

$$\omega_k = \frac{1}{10\ 000^{\frac{2k}{d}}}, \quad k = 1, \cdots, \frac{d}{2} \tag{8}$$

式中，pos和d分别代表输入向量的位置和长度，i是向量中每个元素的索引。借助正弦余弦本身的性质，这种位置编码方式使得位置索引为$pos + p$的元素可以表示为位置索引为p的向量的线性变换。这种设计便于向量中相对位置的表达。

图1展示了Transformer的整体结构。对于编码器模块，多头自注意力聚合计算了编码器内输入向量之间的关系。前馈层从自注意力层的输出中继续提取特征。相对于编码器，Transformer中的解码器通过多头交叉注意力层聚合了编码器输出和解码器输入。此外，编码器和解码器中的所有层之间都有残差连接。Transformer中也使用了归一化层增强模型的可扩展性。为了记录输入向量的次序信息，在编码器和解码器的整体输入中都附加了输入向量的位置编码嵌入。Transformer的最后一层是softmax

层，用于得到预测下一个输入的映射概率。

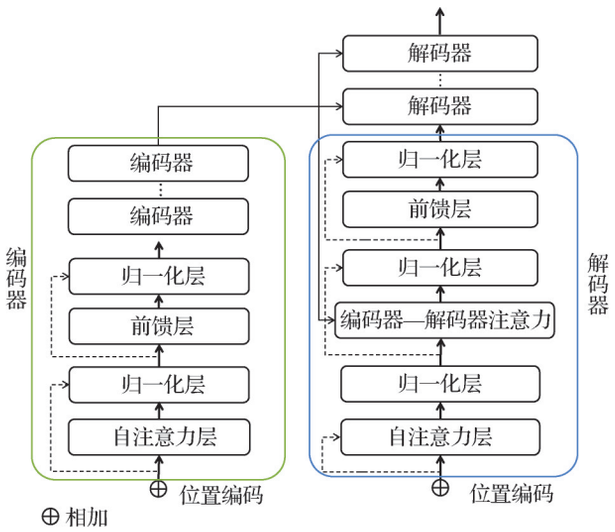

图1　Transformer结构图（Vaswani等，2017）

Fig. 1　The structure of Transformer (Vaswani et al., 2017)

1.2　Vision Transformer

Transformer结构首先在NLP领域大放异彩。BERT、GPT等一系列基础语言大模型相关工作均使用了Transformer中的编码器或解码器结构作为模型的基础结构（Brown等，2020；Devlin等，2019）。

这些进展激发了研究者们将Transformer应用到CV领域。例如，Transformer在视觉领域的早期应用形式一般是“CNN特征 + Transformer编码器”（Chen等，2021a；Parmar等，2018；Schlemper等，2019）。此外，也有研究人员尝试通过将原始图像缩放至低分辨率重新排列成一维序列的方法在CV领域实现类似BERT的预训练（Chen等，2020a）。

与此前一些在小规模视觉模型上应用Transformer结构不同，Dosovitskiy等人（2021）通过ViT探索了使用Transformer结构在CV领域进行大规模预训练学习的可行性。ViT这一开创性工作对大规模预训练视觉基础模型领域的研究产生了深远影响。标准的Transformer仅可接受序列输入，为了使Transformer可以适应图像输入，ViT将输入图像切分成patch，patch被展开为一维向量随后加上位置编码向量。随后，这些嵌入被送入Transformer的编码器中进行计算。

类似BERT，ViT加入了一个类别标记到图像嵌入中，最终作为图像的分类输出。在ViT中，作者使用了大规模图像数据集JFT-300M进行预训练。ViT在多个图像识别基准测试上取得了与当时最先进CNN方法媲美甚至更好的结果。尽管有实验表明，ViT在训练数据规模有限的情况下性能表现低于CNN方法，但是考虑到视觉基础模型的大规模图像数据集预训练需求，ViT具有作为模型基础结构的巨大应用潜力。

2　预训练任务

“预训练—微调”这一深度学习网络的训练范式，在NLP领域已经比较成熟，在CV基础模型领域也得到发展。

在CV基础模型领域，首先通过预训练的方法在大型数据集上进行模型的预训练。这个步骤使得模型可以学习获得在通用场景上的特征提取能力。模型随后会在下游任务的小型数据集上进行微调。这种训练范式可以使得基础模型在不需要投入大量训练成本的情况下获得在多种下游任务上较为优秀的性能。

在NLP领域，BERT、GPT等模型通过在海量文本数据上进行预训练，在多个NLP子任务上取得了超过以往单任务模型的最好性能（Devlin等，2019；Radford等，2019；Radford和Narasimhan，2018）。在视觉领域，大部分的视觉基础模型也使用预训练—微调训练范式。目前主流视觉基础模型应用的预训练方法主要有监督学习、对比学习和掩码图像建模。

2.1　监督学习

在计算机视觉领域，监督学习是一种经典的机器学习方法。监督学习利用带有标签信息的标注数据学习一个模型，并在此基础上对未标注的数据进行预测或分类。早期的一些深度学习相关工作已经通过监督学习方法学习到了可迁移的特征表示，并在多个任务上取得了不错的效果，包括图像分类、目标检测、零样本分类等任务形式（Donahue等，2014；Kolesnikov等，2020b；Kornblith等，2019；Razavian等，2014）。

当Transformer基础结构兴起后，一些研究者将监督学习应用在以Transformer为基础结构的大规模参数模型中，在大量数据上进行预训练，在下游任务上获得了显著的性能提升和鲁棒性提升（石争浩等，2023）。例如，ViT-L模型在包含了3亿幅图像的JFT数据集上进行预训练后，相比只在ImageNet训

练集上训练的情况性能提升了13%(Dosovitskiy等，2021)。微软团队之后提出一种带有分层滑动窗口设计的Swin Transformer模型，通过在千万幅图像量级的数据集上进行预训练，在分类任务、目标检测任务评测基准上取得了当时的最高水平(Liu等，2021)。谷歌研究人员将用于ViT的训练数据规模提高到了30亿的规模，并在此基础上训练出了20亿参数的ViT模型，将ImageNet上top-1准确率刷新到90.45%(Zhai等，2022b)。

虽然一系列工作表明了采用监督学习可以在模型参数规模和训练数据规模扩展的情况下，在公共评测基准和少样本迁移任务上取得性能提升，但随着模型参数量的增大，满足模型进行充分监督学习所需要的有标签数据集的规模持续增大。对大规模数据进行标注需要耗费巨大的人力物力成本。研究人员同时也在探索其他视觉基础模型可行的预训练任务。

以监督学习为主要训练方法的深度神经网络在过去数十年内取得了巨大进步，特别是在包括图像分类、目标检测、语义分割和情感分析等多种机器学习任务中。通常，监督学习使用在特定任务上有监督信息的大型数据集进行训练。视觉基础模型早期发展阶段中的一些工作如ViT、Swin Transformer均在大规模图像数据集上进行了监督学习预训练。但是，监督学习在视觉基础模型的训练上面临着瓶颈。监督学习依赖巨量的标注信息，造成训练成本的上升。此外，监督学习很容易受到泛化误差、偶然相关性和对抗攻击。研究人员希望视觉基础模型可以在更少的标签、样本下学习更多的通用知识。基于此背景，图像对比学习、图文对比学习和掩码图像学习等方法因其数据效率等优势吸引了研究人员广泛关注，很多相关工作开始使用这些范式作为预训练方法。

2.2 图像对比学习

对比学习的核心思想是学习一个或多个表征编码器。这些编码器将输入编码为表征，使得相似样本(正样本)在特征空间中的距离相近、非相似样本(负样本)在特征空间中的距离更远(Hadsell等，2006)。例如，对于同一幅图像，可以在不影响概念图像语义的情况下通过不同的变换(如裁剪、旋转和颜色变换等)来生成多个视图。这些多视图可以作为正样本，其他不同图像可以作为负样本。通过最大化正样本之间的相似度，同时最小化负样本之间的相似度，可以训练得到一个图像表征编码器。一些研究工作已经发现对比学习预训练得到的表征在特定下游任务上性能超过了监督学习方法(Chen等，2020b)。

对比学习中的一类对比方法为“仅图像对比”。图像对比学习需要设计特定的对比损失函数来实现图像实例之间的判别。一些研究者基于信息熵的损失函数是指使用交叉熵或互信息等信息论量来定义对比损失函数，例如InfoNCE(information noise contrastive estimation)(van den Oord等，2019)、SimCLR(a simple framework for contrastive learning)(Chen等，2020b)、MoCo(momentum contrast)(He等，2020)、BYOL(bootstrap your own latent)(He等，2020)和SimSiam(simple Siamese)(Chen等，2021a)等方法。这类方法通常需要构建一个大规模的负样本集合，来增加对比难度和提高特征质量，然而，这也带来了计算和内存的挑战。因此一些方法采用了动量编码器(He等，2020)、自监督预测器(Chen等，2021b；Grill等，2020)等技术来降低计算开销和提高稳定性。这类方法可以分为两个阶段：第1阶段是使用对比损失函数进行无监督训练；第2阶段是使用有监督损失函数进行微调或线性评估。其中，MoCo v3(Chen等，2021b)是目前在无监督视觉特征学习领域最先进的方法之一，它使用了Transformer网络结构和多种数据增强方式，来提高对比学习的效果。

基于距离的损失函数是指使用欧氏距离、余弦距离等距离度量来定义对比损失函数，例如NPID(non parametric instance discrimination)(Wu等，2018)、RandomCrop(Peng等，2022a)等方法。这类方法通常不需要构建负样本集合，而是直接使用全局特征分布的统计信息来进行对比学习。SwAV(swapping assignments between multiple views of the same image)提出了一种新的数据增强策略multi-crop：使用不同分辨率的多个视图来替代两个全分辨率的视图，从而提高自监督学习的效果。基于SwAV对比学习方法，Goyal等人(2021)使用10亿幅图像训练完成了13亿参数的SEER基础模型，在ImageNet评测基准上取得了当时最高水平(Goyal等，2021；Caron等，2020)。

对比学习还受到数据增强、网络结构和优化策略等因素的影响，这些因素共同决定了视觉特征学

习的效果。一些研究探讨了不同视图生成方式对对比学习的影响(Tian等,2020),发现一些有利于语义分割或目标检测等下游任务的数据增强方式也能提高对比学习的性能。另一些研究探讨了不同网络结构对对比学习的影响(Caron等,2021;Chen等,2021b),发现使用Transformer网络结构能够进一步提升对比学习的效果,甚至可以在无监督情况下达到与有监督训练相当甚至更好的水平。此外,还有一些研究探讨了不同的优化策略对对比学习的影响,例如使用不同的学习率、批量大小以及正则化项等。

对比学习是视觉领域目前主流的自监督方法之一,可以利用大量无标注数据提高视觉任务的性能。图像对比学习直接研究不同样本实例之间的关系,这种方法天然适合于图像分类任务。与生成模型相比,应用图像对比学习方法的视觉基础模型在预训练阶段基本抛弃了解码器结构。因此,此类模型相对轻量,适合进行判别式的下游任务。然而,图像对比学习仍然存在一些问题需解决。大多数图像对比学习方法都需要对图像数据进行负样本采样,但是负样本采样这一过程通常耗费时间,且需要避免带来模型偏置。研究人员已经证明,通过数据增强,图像对比学习方法可以进一步给视觉基础模型带来性能提升,但是其中的原理解释仍然相当模糊,这阻碍了这种预训练方法在更多领域中的应用。如何进一步利用数据采样和增强的方法提升图像对比学习的能力是未来一个重要研究方向。

2.3 图文对比学习

除了仅图像对比的方法之外,还有图像文本对比学习方法。通过文本信息辅助以提升模型在视觉任务上的表现可以追溯到约20年前。Mori等人(1999)基于图像匹配的文档内容进行文档形容词、名词等描述预测,通过此预测任务提升了模型在图像检索任务的表现。Joulin等人(2016)使用了一个1亿幅含有社交媒体标签的Flickr图像数据集,通过预测图像和文本之间的相关性训练网络。作者发现这种训练方式可以获得高效的视觉特征,在图像分类、目标检测等任务上达到或者超过有监督训练的性能,同时可以在词语相似度等任务上表现出优势。Li等人(2017)提出一种基于文本注释的视觉*n*-gram模型,用于从网络图文数据中学习视觉表示。视觉*n*-gram模型借鉴了语言建模中常用的*n*-gram模型,其使用了COCO(common objects in context) caption数据集进行训练。该模型可以根据图像预测与内容相关的短语,在基于短语的图像检索等零样本任务上取得了不错的效果。

还有一些新的工作通过基于Transformer语言建模(Desai和Johnson,2021)、掩码语言建模(Sariyildiz等,2020)、对比学习(Zhang等,2022b)等方法,利用自然语言信息提升模型方法提取视觉表示的质量。然而,当时这些模型虽然在如何通过自然语言的监督信息获得有效的视觉表示方面取得了一些进展,但是在零样本图像任务上的性能表现落后于同时期的纯视觉模型(Xie等,2020)。

Mahajan等人(2018)提出了利用语言信息弱监督的预训练数据集提升神经网络在视觉任务上迁移学习效果。通过使用数十亿幅带有社交媒体标签的图像进行训练,发现在迁移到多个图像分类和目标检测评测数据集上效果优秀。Mahajan的工作证明了大规模弱监督预训练对于迁移学习性能提升的有效性。谷歌大脑团队先后通过BiT和ViT两个工作证明了在大规模图像数据集上进行的预训练可以推动模型在多个图像下游任务上取得同时期最高的性能(Dosovitskiy等,2021;Kolesnikov等,2020a)。

OpenAI团队在2021年提出了CLIP(contrastive language image pretraining)模型,首次使用了从互联网收集的4亿对图文数据进行训练(Radford等,2021)。CLIP利用对比学习的思想,将图像和文本编码后的特征在联合表示空间进行对齐匹配。CLIP这项工作对基础模型的发展具有重要影响,开启了大规模预训练基础模型的新篇章。CLIP在多个视觉任务上都能达到或者超过有监督或无监督的预训练方法,展现出强大的零样本能力。CLIP也启发了大量后续工作,如对比学习范式在其他视觉任务上的应用(Sun等,2023;Zhang等,2022b;Zhou等,2022),通过将训练数据扩展到更大规模进一步提升模型能力(Jia等,2021)等。更多相关的衍生工作在3.3节中介绍。

对比学习的高级理念是简化学习任务,要求模型从经过设计的有限选项中选出正确答案。这种任务设置鼓励模型学习图像中的高阶表征信息,而不是低阶的一般信息。包括2.2节提到的一些工作在内的研究已经从图像对比学习领域证明了这一思想的有效性(Chen等,2020b;Doersch等,2015;He等,

2020；Noroozi 和 Favaro，2016）。而图像内容对应的文本信息就是一种图像的天然高阶表征方式。传统的图像分类方法中模型会学习固定的物体类别，这使得模型无法预测和转移超出这个封闭集的概念。而自然语言在语义上比概念标签集（如物体类别）都要丰富，这种形式得以表示广泛的视觉概念，使模型的泛化能力和可用性得到了提高。图文对比学习预训练方法使得模型具有强大的零样本、少样本能力，可以高效快速地迁移到下游任务中。但是由于图文对比学习需要依赖巨量的图文匹配样本数据，数据的收集、清洗成本高，且模型的预训练需要极大的算力资源。一些研究发现尽管此类模型已经在零样本任务中表现出色，但并不能完全解决域外泛化问题。此外，计算机视觉任务需要在不同粒度级别（图像、区域、像素）进行处理，这使得跨模态语义匹配具有挑战性。构建一个可以利用不同粒度级别的图像文本数据的基础模型，寻求数据规模和语义丰富性之间的最佳权衡，对于图文匹配学习仍然是一个有吸引力的研究课题。

2.4 掩码图像建模

掩码图像建模的思路来源于自然语言处理领域的掩码语言建模。BERT 等工作证明了掩码语言建模任务对于语言表征学习的有效性（Bao 等，2020；Devlin 等，2019；Dong 等，2019）。掩码图像建模的基本思想是：对输入图像进行切片分割，然后随机掩蔽一部分切片，输入到模型中进行编码，最后通过一个解码器来重构被掩蔽的部分。这种通过部分观测来重构整体的方式，可以让模型学习到丰富的视觉表示。

现有的掩码图像建模方法根据重构目标可大致分为 3 种：重构低级图像元素，如原始像素（Fang 等，2023a；He 等，2022）、重构手工设计特征（Wei 等，2022a）和重构视觉标记（Bao 等，2022；Chen 等，2024；Dong 等，2023；El-Nouby 等，2021；Wang 等，2022c）。

在 CV 领域，BEiT 是最早应用这种思想的工作之一（Bao 等，2022）。BEiT 使用一个预训练的标记器（tokenizer）将图像编码成一系列标记（token），然后随机掩蔽部分标记，要求模型预测被掩蔽的标记。之后，掩码自动编码机（masked autoencoders，MAE）方法（He 等，2022）被提出。MAE 进一步简化了这种框架，直接在图像的切片上进行随机掩蔽，不需要引入额外的标记器，重构目标也从标记变为直接预测像素值，这使得 MAE 的训练速度比 BEiT 有了大幅提升。BEiT 的后续改进工作 BEiT 2 加入了［CLS］符号以建模图像的全局信息，并提出使用 VQ-KD（vector-quantized knowledge distillation）方法对图像进行高级语义编码（Peng 等，2022b）。BEiT 2 的掩码图像建模了图像的高级语义信息，并在多个计算机视觉任务上取得了进一步提升。还有一些研究者提出了使用知识蒸馏和自蒸馏的方法预测掩码位置的图像特征。

利用掩码图像建模方法，随着模型参数规模和训练数据规模的扩大，模型在下游视觉任务上的性能可以随之提升（Chowdhery 等，2023），这一性质对于开发视觉基础模型非常重要。EVA 是掩码图像建模应用到视觉基础模型中的一个代表性工作（Fang 等，2023c）。EVA 的研究者们探索了不同的前置任务。他们发现，简单地根据可见的图像块重构被掩蔽的 CLIP 模型的图像编码视觉特征是非常高效的。EVA 使用了 2 960 万无标签图像进行掩码图像预训练，并将模型参数规模扩展到 11 亿。在这种规模的预训练下，EVA 通过少量样本的微调在图像分类、目标检测和语义分割等任务上取得了当时的最高水平。

掩码图像建模与自然语言处理中的掩码语言建模思想类似，无需标签即可通过自监督方法学习到鲁棒的视觉表征。在掩码图像建模中，每个图像块被形式化为一个标记，每个标记都是一个独立的单元，在预训练过程中恢复被掩码的图像块标记。这一过程充分利用了 Transformer 这种基础网络结构的优势。掩码图像建模展现出了“可扩展”的重要特性，随着模型参数规模和训练数据规模的扩大，模型在下游视觉任务上的性能可以随之提升（Chowdhery 等，2023），这一性质对于开发视觉基础模型非常重要。掩码图像建模为视觉基础模型的构建提供了一个重要的预训练方法。目前研究者们正在从探索不同的重构目标（Liu 等，2022；Wei 等，2022a）、更复杂的网络设计（Tian 等，2023；Zhang 等，2022a）、从其他模态引入监督信息（Peng 等，2022b；Wang 等，2022c；Wei 等，2022b）等方向对掩码图像建模进行改进。

3 视觉基础模型介绍

3.1 基于混合专家的基础模型

随着深度学习模型以及基础模型的发展，扩大

网络参数规模以提高模型性能已经成为一种常用手段(Bommasani等,2022)。然而,这类大规模参数的模型在训练和推理阶段都需要消耗很高的算力成本(Patterson等,2021;Strubell等,2019)。一个主要原因是这类网络属于参数稠密型网络,在处理实例时需要激活所有参数。相比之下,MoE可以通过参数子集的方式提高整体模型的参数规模,同时在训练和推理时保持成本基本不变。基于MoE的基础模型在NLP领域非常流行。

Hochreiter和Schmidhuber(1997)最早在长短程记忆网络(long-short term memory, LSTM)层中直接添加一个MoE层,从而将标记通过路由分配到对应专家层,并通过实验结果验证了MoE机制的有效性(Fedus等,2022)。基于MoE的基础模型已经在NLP领域取得了很大的突破,在很多任务上表现出色(Devlin等,2019;Peters等,2018)。参数规模上,在NLP领域文本类的基础模型已经达到了千亿、万亿级别,如国外的GPT-3模型、国内的悟道2.0和阿里M6大模型等(Brown等,2020;Lin等,2021;Zeng等,2021)。其中,基于MoE的基础模型能够以更少的资源进行训练和推理,同时使模型的参数扩展到万亿规模,其中一个代表性的工作是由谷歌提出的Switch Transformer(Fedus等,2022)。Switch Transformer在保持计算成本基本恒定的同时,使用了一种简化的路由算法将模型参数扩展到1.6万亿。Switch Transformer的训练时间比T5模型提高了7倍(Raffel等,2023),并且在101种语言的测评指标上超过了mT5-base模型(Xue等,2021)。

V-moe是第一个将MoE结构设计引入视觉基础模型设计的工作(Riquelme等,2021)。V-moe以稀疏的MoE层取代了ViT种前馈层的一部分子集网络,其中每个图像切片都被路由到专家网络对应的子集。V-moe探索了各种网络设计方案,提出了一种对于MoE架构视觉基础模型的有效预训练和下游任务迁移方法。V-moe的参数量达到了150亿规模,成为当时规模最大的视觉基础模型。V-moe通过调节输入或模型权重的稀疏程度,可以平滑地调整已训练模型的性能与推理成本之间的平衡。商汤科技和上海人工智能实验室联合提出了书生视觉基础大模型INTERN(Shao等,2022)。INTERN的目标是解决当前CV领域中存在的任务通用、场景泛化和数据效率等一系列瓶颈问题。INTERN设计了3种不同等级的专家网络,分层次解决不同难度的视觉任务。INTERN模型在4类视觉核心任务的26个公开数据集上进行了评估,结果显示在只使用10%训练数据进行微调的情况下,INTERN模型在大多数情况下都超过了使用全部训练数据的对比模型。Zhu等人(2022)提出了Uni-Perceiver-MoE模型,在通用视觉感知模型中引入了条件MoE。通过条件MoE,Uni-Perceiver模型可以有效地减轻任务和模态之间的干扰,并通过对1%的下游数据的快速调整,在一系列下游任务上获得最先进的结果。同时,Uni-Perceiver模型有效减轻了推理消耗,保持了低计算和存储成本。

MoE的网络结构中部分参数由整个网络结构共享。目前主要的参数共享方式包括软参数共享(Duong等,2015;Gao等,2020;Liu等,2016;Misra等,2016)(每个专家拥有独立的模型和参数,但是通过正则化结构可以获得其他专家的内部信息)、硬参数共享(Liu等,2019;Long等,2017;Lu等,2017;Subramanian等,2018)方式(每个专家拥有共享的网络参数)。Task-MoE提出一种将两种参数共享方式结合用于多任务学习的方法(Kudugunta等,2021)。Task-MoE在任务间共享自注意力模块并且基于一个任务级的路由选择任务特定的网络。百度提出的UFO(unified feature optimization)视觉基础模型受这个工作启发,在任务级路由的基础上加入剪枝机制。UFO不仅提高了模型在下游子任务的精度,同时控制了子任务模型规模,减少了推理时的算力消耗。UFO大模型的参数规模达到了170亿,在28个公开视觉评测数据集上达到了最高水平。

除了面向多任务的MoE基础模型,微软提出的BEiT-3基础模型将特定模态专家引入基础模型设计(Wang等,2022c)。考虑到过去的一些MoE基础模型无法进行跨模态的参数共享,BEIT-3分别设计了视觉专家、语言专家和多模态专家负责不同的模态任务。BEiT-3在目标检测、语义分割和图像分类等评价基准上获得了当时的最高水平。

英伟达的研究者提出了一种数据和参数高效的基础模型Prismer(Liu等,2024)。Prismer利用一系列预训练的领域专家模型,并在专家集合的训练过程中保持参数固定,只对少量模型组件进行训练。Prismer仅使用了1 300万公开图像数据进行训练,在图像分类等任务上取得了与需要更大量数据训练

的模型相当的优秀性能。

3.2 基于通用视觉表征主干的基础模型

传统深度学习模型在处理多场景下的复杂视觉任务时具有局限性。传统的深度学习模型尤其是CNN虽然在特定的视觉任务上取得了显著的成效(Krizhevsky等,2017),但在迁移到新的未知视觉环境时,其效果表现通常较差。这种局限性根源在于CNN网络结构的特性。CNN网络架构倾向于强化输入数据中的特定偏置,这种对于特定任务的偏置可能导致CNN在面对多样化或偏差与训练数据不一致的数据集时,泛化能力受限。针对传统深度学习网络的这一缺陷,学界开始寻求更加灵活和广泛适用的模型架构,以实现泛化能力更为强大的通用视觉表征。

近年来,ViT经大规模训练后在小数据集上超越传统神经网络,但处理图像算力消耗大。微软亚研院提出的 Swin Transformer 采用分层和移动窗口设计,处理大尺寸图像计算效率高,且易集成到不同图像及视觉任务中,在视觉核心任务上表现出色。

Swin Transformer V2对Swin Transformer进行了进一步升级(Liu等,2022)。Swin Transformer V2成功地训练了一个30亿参数的通用视觉表征主干模型,能够处理更高分辨率的图像。

BEiT模型借鉴NLP基础模型BERT的思想,提出了一种新的自监督视觉表征模型,通过掩码图像建模任务对ViT进行预训练。这一方法不仅在图像分类和语义分割等下游任务上展现出当时的最优能力,还验证了BEiT这种通用视觉表征主干在区分图像语义区域和视觉对象边界方面具有出色的能力。

EVA融合了对比学习和掩码图像建模的思想,目前被广泛应用作为通用视觉表示主干网络之一,对于视觉基础模型的研究领域具有重要的影响力(Fang等,2023c)。EVA在具有挑战性的大词汇表实例分割任务上取得了巨大进步。

除了基于Transformer结构的通用视觉表征主干网络之外,一些研究工作也提出了基于其他网络结构的视觉主干网络。由商汤科技研究者提出的Internimage模型展示出了基于可变形卷积神经网络的大型视觉基础模型(Wang等,2023)。与传统的卷积神经网络相比,可变形卷积神经网络增加了对卷积核形态调整能力,可以根据输入数据动态偏移。Internimage首次在4亿预训练图像上将基于CNN架构的通用视觉主干网络扩展到10亿参数规模,并在目标检测任务上达到了当时的最优水平,证明了卷积网络也具有成为大规模视觉基础模型组件的潜力。

一些研究工作尝试将CNN的高效计算优势与Transformer的全局注意力优势结合起来(Dai等,2021;Wu等,2021;Zhang等, 2022a),HiViT是其中的一个代表性工作(Zhang等,2022a),其解决了掩码图像建模预训练任务中计算量大和内存消耗多的问题,使得视觉主干具有对于图像分层建模的能力。面对ImageNet分类任务,HiViT在同等参数量的情况下获得了超过ViT模型0.6%精确度以及相对Swin Transformer模型1.9倍的加速比。

3.3 对齐语言模态的视觉基础模型

计算机视觉中的核心任务包括图像分类、物体检测等。这些任务的主要目标是将有意义的语义概念分配到对应的视觉实例中,例如整幅图像或者是图像部分区域。传统的计算机视觉系统需要在一个预先设定好的视觉概念集合中进行训练并且预测。这也就是计算机视觉中所说的闭集问题(close set),例如在ImageNet (Deng等,2009)和JFT300M中的图像类别标签以及COCO数据集(Lin等,2014)中的物体类别标签。

随着计算机视觉技术的发展,在很多特定数据集任务上深度学习模型已经获得了接近人类甚至超越人类的性能。尽管如此,由于缺乏对不在训练集中未知类别的识别能力,这种在视觉概念闭集中进行特定任务的范式限制了模型的泛化性和任务通用性。例如,传统图像分类任务是从预先定义范围的图像集中学习特征,很难将分类能力迁移到语义概念在预定义范围外的图像集中。针对这些问题,研究人员提出利用文本作为监督信息以增强视觉模型应对开集问题(open set)能力的方法。相对于传统视觉网络中将语义概念在分类头结构中进行区分这种将语义概念过于简化的处理方式,最近提出的新方法通常将语义信息通过文本编码器进行处理。这类处理方法能够使得编码后的语义概念具有远超简单物体标签的语义丰富度,使得语义概念的覆盖范围获得极大拓展。

由OpenAI提出的CLIP模型是图像文本融合对齐方面的开创性工作(Radford等,2021)。研究人员利用算力优势,使用网络爬取的4亿个图像文本对

进行CLIP模型的预训练。此外，作者使用对比学习方法构建CLIP模型的跨模态对齐和零样本学习能力。CLIP模型的图像文本对齐能力具体构建步骤为：将图像和文本分别进行特征提取编码后计算图像文本对的余弦距离，通过对比损失来拉近匹配的图像文本对的距离、拉远不匹配的图像文本对的距离，以此在特征空间建立起图像和文本之间的联系。

如CLIP等此类模型将图像模态和语言模态在同一特征空间进行对齐，其训练目标函数为

$$\min_{\{\theta,\phi\}} L_{\mathrm{CLIP}} = L_{i2t} + L_{t2i} \tag{9}$$

式中，L_{i2t}表示图像到文本的损失，L_{t2i}表示文本到图像的损失。具体为

$$L_{i2t} = -\sum_{i \in \boldsymbol{B}} \log \frac{\exp(\tau \boldsymbol{u}_i \boldsymbol{v}_i)}{\sum_{j \in \boldsymbol{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)} \tag{10}$$

$$L_{t2i} = -\sum_{j \in \boldsymbol{B}} \log \frac{\exp(\tau \boldsymbol{u}_j \boldsymbol{v}_j)}{\sum_{i \in \boldsymbol{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)} \tag{11}$$

CLIP的整体损失由L_{i2t}和L_{t2i}求和。L_{i2t}计算了在一个样本批次$\boldsymbol{B}$中，正确匹配图像嵌入$\boldsymbol{u}_i$和$\boldsymbol{v}_i$的对数概率。表达式$\exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)$计算了图像和文本嵌入的点积经过温度参数$\tau$缩放后的指数，表示它们的相似度。分母对批次中所有文本嵌入的这些值进行求和，作为一个归一化因子。L_{t2i}的表达与L_{i2t}类似，但专注于将文本对齐到图像。

CLIP模型在零样本分类任务上取得了超过有监督学习的经典网络ResNet50（residual neural network 50）（He等，2016）的性能。得益于整体网络结构简单、兼容多种图像、文本编码器以及其开源的特性，CLIP对后续的基础模型方面的研究影响很大。很多基础模型的工作基于CLIP的研究展开，利用CLIP强大的跨模态对齐能力提升模型在下游任务上的表现（姜定和叶茫，2023；张浩宇 等，2022）。CLIP开启了在基础视觉任务中借助图像文本模态匹配、以自然语言作为监督辅助提升图像任务效果这种新范式的先河，研究者们在此范式基础上提出一系列改进的方法。

如图2所示，CLIP模型使用网络收集的4亿对图文匹配数据进行训练。这种规模巨大的网络数据集的收集和清洗会占用大量人力物力资源，从而限制了数据集规模以及模型规模的进一步扩大。鉴于此，谷歌的研究者提出了ALIGN（a large scale image and noise text embedding）基础模型（Jia等，2021）。

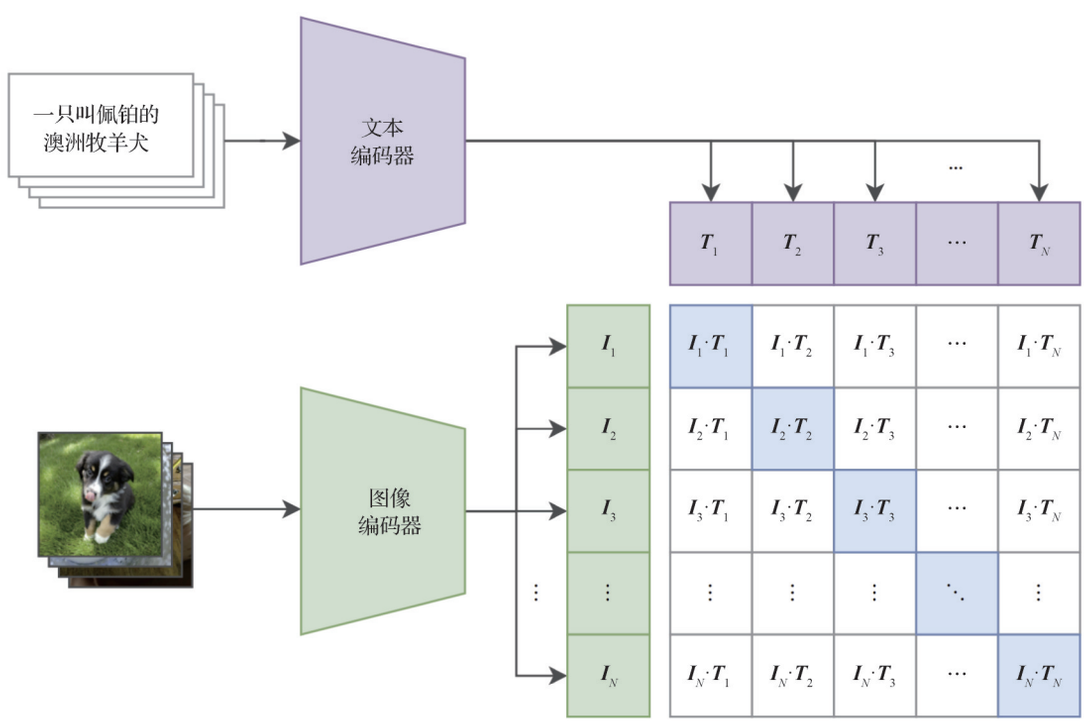

图2　CLIP示意图（Radford等，2021）

Fig. 2　An illustration of CLIP（Radford et al.，2021）

ALIGN使用简单的dual encoder结构在一个共享隐空间匹配视觉、语言两种表征，并且在18亿规模的含噪声图文数据集上进行训练。ALIGN模型在零样本任务上取得了当时最优性能，说明通过图文数据集规模的扩张，即使数据含噪声较大也能促进模型性能的提升。

CLIP和ALIGN模型奠定了视觉语言特征对齐的视觉基础模型的双塔结构，即广义上的图像编码器和文本编码器。双塔结构中的图像编码器和文本编码器分别对数据集中已经配对的图像和文本进行编码，同时通过对比学习算法最小化对比损失函数，从而将编码后的图像特征和文本特征进行匹配对齐。

一系列工作对如何更好地对齐图像和文本模态进行了探讨。随着大规模预训练技术的发展，在大型数据集上完成预训练的视觉模型和语言模型已经具备了在对应模态上强大的特征提取能力。然而，在双塔结构中引入大规模预训练模型也给视觉语言模态对齐带来一些新的问题。例如，如何充分利用预训练的模态编码器，如何协同不同模态的编码特征进行对齐，同时不损失编码器的特征提取能力。Zhai等人（2022b）对基于视觉语言模型双塔结构的多种设置策略进行了实验，提出“LiT：Locked image tuning”策略能够让对齐了语言模态的视觉模型取得最佳性能。LiT策略为如何充分利用单一模态预训练模型提高视觉基础模型在下游任务上的性能表现提供了有价值的参考。

CLIP和ALIGN这类模型是将视觉和语言模态在一个稠密嵌入空间（dense embedding space）进行建模。Jia等人（2021）在此基础上提出可以将文本模态映射到每个维度，表示一个子单词的稀疏嵌入空间（sparse embedding space）的STAIR（sparse text and image representation）模型。通过将视觉特征与稀疏的语言特征进行匹配，基础模型的可解释性相比之前的模型得到了进一步增强。

Yao等人（2022）提出FILIP（fine-grained interactive language-image pre-training）模型，FILIP模型具有从图像中学习区域细粒度视觉特征的能力。FILIP模型通过设计针对图像patch和文本token交叉模态的隐交互（late interaction）机制，让模型在预训练过程中自主学习到图像patch和文本token之间的细粒度语义对应的关系。

FILIP是一种不依赖细粒度的人工标注的融合语言模态的视觉基础模型。由于模型中的patch由图像等分切割而成，FILIP的细粒度对齐准确性依赖于图像中patch的划分，这种情况下对图像中物体的定位精确度不高。

微软团队提出了在具有细粒度标注的图像文本数据集上训练的GLIP（grounded language image pre-training）模型（Li等，2022）。GLIP模型提出了短语定位（phrase grounding）概念，具体含义为，将图像中的各个对象与整幅图像文本描述中的短语进行匹配对齐。

3.4 生成式多任务基础模型

随着NLP领域基础模型的不断发展，以往需要针对不同NLP任务专门设计模型的做法正逐渐演变为通过一个统一的多任务大型基础模型来处理各类NLP任务。

在NLP领域，研究人员已经探索了多种方式来赋予基础模型多任务能力。这些方式包括增加任务特定的头部（Devlin等，2019）、为下游任务添加适配器模块（Houlsby等，2019），或者使用软提示来匹配模型的功能与下游任务（Lester等，2021）。这些工作启发CV领域的研究者们通过生成式的方法统一视觉领域的多种任务。“统一多任务”具体体现在以下几个方面：任务无关性（支持不同的任务类型）、任务完备性（需要有足够的任务丰富度以此让模型获得鲁棒的零样本泛化能力）。

在最近的视觉基础模型领域中，有一系列工作围绕着生成式多任务基础模型展开。Wang等人（2022b）提出了一个任务模态均不可知的OFA（one for all）模型。该模型通过序列到序列的学习方法，实现了跨模态任务和单模态任务的统一。OFA在多个视觉任务测评基准上取得了领先的性能。CoCa（contrastive captioners）模型则利用对比损失和图像描述损失对图像和文本编码器进行预训练，在零样本的ImageNet分类任务上取得了领先的准确率（Yu等，2022）。Flamingo模型桥接了处理视觉任务和文本任务的预训练模型以充分利用两种模态的优势，其可以处理交错的视觉文本序列，并且从中获得了强大的少样本学习能力（Alayrac等，2022）。

此外，GIT（generative image-to-text Transformer）模型采用简化的图像编码器的文本编码器架构，不仅在多个视觉任务上取得了高性能表现，更是在

TextCaps场景文字理解评测基础上首次超过了人类表现(Wang等,2022a)。Lu等人(2023)提出的Unified-IO(unified input output)模型将输入和输出同质化成一系列离散的词汇标记来实现任务间的统一,进而在一个基于Transformer的架构上联合训练超过90个不同的视觉和语言数据集。

UniTAB模型(Yang等,2022)提出了一个用于基于视觉建模任务的统一文本和框输出模型,通过自回归方式生成文本和框标记。受Pix2Seq(Chen等,2022)工作的启发,研究者提出了一种全新的任务不可知视觉语言架构GPV-1(general purpose vision)模型(Gupta等,2022)。GPV-1模型能够接收图像和文本描述作为输入,并输出与任务相关的文本或边界框。GPV-1旨在减少开发新应用程序所需的时间和专业知识,可以学习并执行分类、定位和描述生成等任务,而无需对架构或学习过程进行修改。GPV-1为通用视觉系统的研究提供了新的方向和方法论。

然而,生成式多任务模型由于模型体量通常较大,在任务推理时速度较慢、算力消耗大;同时,生成式多任务模型在单个视觉任务的性能表现上往往难以超过传统的特定任务模型;此外,生成式多任务模型由于其模型结构的整体性特点,在学习过程中容易受到其他任务训练的干扰。

4 预训练数据集

视觉基础模型需要大量数据进行预训练。本节介绍目前视觉基础模型中常用的大规模图像数据集以及图像文本数据集。表1展示了相关主流数据集的基本概况。

4.1 图像数据集

1)ImageNet。ImageNet数据集是在2010年左右作为ImageNet大规模视觉识别挑战赛的一部分发布的(Deng等,2009)。该数据集包括约1 000个类别,每个类别大约有1 000多幅图像,总计约100万幅图像。这些类别涵盖了从动物到日常物体等多种类别。数据集中的图像都经过人工标注以确保精确匹配。ImageNet-1k是CV研究和应用的基石。在图像分类和目标检测领域,ImageNet广泛用于评估和训练深度学习模型。

2)ImageNet-21k。ImageNet-21k是ImageNet数据集的扩展,其标签数量达到了ImageNet的21倍。得益于类别多样性和样本丰富性,ImageNet-21k成为大规模图像预训练的主要数据集。为了优化大规模预训练的训练效率和数据处理,研究人员提出了ImageNet-21k-P数据集。通过对图像样本进行进一步的图像去重、类别整合和数据清洗等操作,ImageNet-21k-P数据集相比原ImageNet-21k数据集更易于深度学习模型的大规模数据训练。

3)JFT-300M/3B。JFT-300M是由谷歌开发的大型图像数据集。它包含约3亿幅从网络采集得到的图像。通过自动化算法得到了约10亿个标签,研究人员经过算法过滤后得到了约3亿个标签。由于其庞大的规模,它对于研究视觉模型处理大规模和复杂数据集的情况表现出很大价值。同时,由于JFT-300M的标签是由算法生成的,存在一定程度的噪声和偏差。JFT-3B是谷歌基于JFT-300M进行的数据集扩展,拥有30亿图像数据。研究人员通过半自动处理流程为这些图像获取了约3万个层次化分类的标签。目前JFT-3B尚未对外公开,仅在谷歌的研究工作中使用(Zhai等,2022a)。

4)COCO。COCO数据集是CV研究中广泛使用的大型图像数据集,可应用于目标检测、图像分割和图像描述等多种任务(Lin等,2014)。COCO数据集包括了超过32万幅图像,涵盖了各种日常物体和人体相关标注。COCO包括了80个目标检测对象类别以及91个图像分割类别。在人体数据方面,COCO包括了超过25万个人体关键点信息标注。

5)ADE20k。ADE20k数据集是一个得到广泛应用的图像数据集,主要关注语义分割领域(Zhou等,2017)。ADE20k由美国麻省理工学院的Computer Science and Artificial Intelligence Laboratory开发,包含2.2万幅图像,这些图像涵盖了室内、城市景观和自然环境。该数据集具有150个以上的类别和35个场景类别,具有细粒度标注的特性。

6)Object365。Object365由旷视科技和北京智源人工智能实验室合作开发(Shao等,2019)。Object365关注CV中开放世界数据的目标检测问题,包含365个类别、200万幅图像和3 000万个目标标注框。得益于其丰富的物体类别和高质量的标准,Object365成为视觉模型进行大规模预训练的重要数据集。

4.2 图像文本数据集

除了图像数据集之外，在3.3节中提到了借助语言模态信息辅助的视觉基础模型。此类模型需要大规模的图像—文本对齐数据。本文也整理了目前在此类模型的训练中使用较多的图像文本数据集。

1）Conceptual Captions数据集。Conceptual Captions是由谷歌创建的大型图像和文本匹配数据集，数据源是谷歌搜集的互联网图文数据（Sharma等，2018）。该数据集中每条样本都包括一幅图像和与这幅图像对应的文本描述。该数据集包括了从日常生活到自然景观等多种场景。Conceptual Captions数据集有300万数据和1 200万两个版本，分别称为CC3M和CC12M（Changpinyo等，2021）。

2）Flickr30k数据集。Flickr30k是一个广泛应用于CV和NLP研究的图像文本数据集（Young等，2014）。Flickr30k包含来自照片共享平台Flickr的大约3.1万幅图像，涵盖了各种日常场景和活动。每幅图像都配有5个不同的描述文本标注。这些信息由人工标注，对场景中的物体和活动进行详细描述。

3）SBU（stony brook university captioned photo）数据集。该数据集由美国纽约州立大学石溪分校的研究者开发，旨在促进图像描述自动生成任务的发展（Ordonez等，2011）。SBU数据集包含约100万幅图像，每幅图像都配有一段自然语言描述文字。这些描述是由原始上传者人工撰写。数据集中的内容包括了人物、动物、自然景观和城市场景等广泛场景。

表1 预训练数据集

Table 1 Pretrain datasets

数据集	类型	发布时间	规模	简介
ImageNet-1k	图像	2010	1 331 167幅标注图像	图像领域使用最广泛的数据集。
ImageNet-21k	图像	2021	21 000类别图像	ImageNet-1k的更大规模、更多类别扩展。
JFT-300M	图像	2017	3万幅图像	数据来源于网络，存在一定噪声。
JFT-3B	图像	2022	30亿幅图像	JFT-300M的扩展，未开源。
COCO	图像	2014	超过328 000幅图像	具有丰富的物体类别和人体标准信息，广泛用于视觉模型的训练数据和评价基准。
ADE20k	图像	2017	27 000幅图像	在广泛的视觉场景上进行了细粒度的语义标注。
Object365	图像	2019	2百万幅图像，3 000万个标注框	关注开放世界的目标检测问题，图像及标注具备广泛的视觉多样性。
CC3M	图像—文本	2018	3百万图文对	发布时间较早的大规模图像—文本描述数据集，爬取自互联网的图像和相关描述，广泛用于图像描述和视觉问答任务。
CC12M	图像—文本	2021	1 200万图文对	数据结构与CC3M类似，数据量扩展到1 200万，适合大规模图像文本训练。
Flickr30k	图像—文本	2014	31 000幅图像，每幅图像5条文本标注，总计155 000条描述	数据集主要针对图像描述和视觉问答任务设计，包括日常活动、人物行为等丰富场景，广泛应用于多模态任务的训练和评价中。
SBU	图像—文本	2011	100万幅图文对	从互联网收集图像，图像描述由原始上传者提供。数据集中的场景具有多样的场景丰富度。

5 思考与展望

随着深度学习技术的不断革新和数据资源的持续丰富，视觉基础模型的发展已经取得了显著的成就。当前的研究成果不仅在理论上推动了人们对视觉信息处理范式的认知，而且在实际应用中展现出巨大的潜力。然而，面对不断扩展的应用领域和日益增长的技术需求，视觉基础模型的发展仍面临着多方面的挑战。

1）当前NLP基础模型的发展相对CV领域来说模型范式相对趋于统一。NLP领域研究者们领域基于提示的方法几乎统一了大部分下游任务。而目前视觉基础模型距离实现这个目标还存在一定距离。

在NLP基础模型领域，“scaling law”理论可以作为研发基础模型时模型性能随数据规模扩展而提高情况的参考。但是视觉基础模型还没有探索出这样一个提高基础模型性能的有效路径方法。NLP基础模型发展历程中的一些经验确实启发了视觉基础模型的发展，但是如何根据图像与文本这两种模态间的本质差异探索出适合于视觉基础模型的道路，这是当前CV研究者们需要考虑的重要课题。

2）从表2可以看到，目前对于视觉基础模型的性能评价主要还是基于传统的视觉识别指标，如ImageNet上的分类准确率、COCO数据集上的目标检测正确率等。但是使用这些传统的视觉封闭集对视觉基础模型进行评测是过时的。

3）在视觉模型领域引入更多自然语言信息的趋势下，模型的评价标准也应该随之发展，以评估视觉模型在广泛任务上的性能。在不同的视觉任务子领域，传统的深度学习方法已经接近现在评价体系的极限。综上，需要开发面向视觉基础模型的新评价方法。这些评价指标应该能够考虑到模型的跨模态理解能力，以及在开放集和广泛下游任务场景中的表现。

4）在视觉下游任务中，预训练域与目标域的数据分布通常具有较大差异。这种情况下，考虑到视觉数据在空间和结构上的复杂性，目前还无法通过预训练基础模型直接解决下游任务问题。在事先对域间差异未知的情况下，如何找到一个统一的范式对模型进行微调需要进一步研究。

5）数据和计算资源的依赖性。当前视觉基础模型通常需要大规模视觉数据进行训练，这个过程会显著消耗算力资源。这种对数据量和算力资源的依赖性限制了高性能视觉基础模型的推广应用。

6）泛化能力。尽管视觉基础模型在开放域视觉数据上的识别能力已经超越了传统深度学习模型，但是在面对多样性输入和长尾分布的视觉场景时，视觉基础模型的泛化性和鲁棒性仍有提高空间。

7）解释性与鲁棒性。尽管视觉基础模型在性能表现上超过了传统深度学习模型，但是其本质仍未脱离深度学习的“黑盒”问题。随着模型结构的日益复杂，如何保证模型在一些敏感领域，如医疗诊断、自动驾驶等场景应用时的鲁棒性，是当前面临的重要挑战之一。

尽管存在以上挑战，当前视觉基础模型的发展仍处于方兴未艾的阶段，本文对未来潜在的研究方向进行以下几个方面的展望。

1）视觉基础模型的评价指标。在CV领域面向特定场景特定任务的视觉任务评测已经非常丰富，

表2　视觉基础模型主流视觉任务性能表

Table2　Downstream task results of vision foundation models

/%

模型	任务、数据集、测试参数			
	图像分类	目标检测		图像分割
	ImageNet	COCO		ADE20k
	Top-1	AP_box	AP_mask	mIoU
Swin-T(Liu 等，2021)	87.3	58.7	51.1	53.5
Swin-T V2(Liu 等，2022)	90.17	63.1	54.4	59.9
BEiT(Bao 等，2022)	88.4	–	–	53.3
BEiT V2(Peng等，2022b)	89	–	–	57
Internimage(Wang 等，2023)	89.6	56.2	48.8	62.9
INTERN(Shao 等，2022)	84.2	48.2	44	49.3
EVA (Fang等，2023c)	89.7	64.2	55	62.3
EVA-2 (Fang等，2023b)	90	64.1	55.4	62
HiViT(Zhang 等，2022a)	84.2	51.2	44.2	51.2
CLIP(Radford 等，2021)	85.4	–	–	–

注：同一模型有多个参数规模时取其最好性能；“–”表示模型未进行或未进行同等设置的该类实验。

但是建立覆盖任务类型广泛、数据场景足够多样、可作为视觉基础模型的通用基准测评体系是CV领域所亟需的。

2)跨模态能力提升。当前对齐语言模态的视觉基础模型主要通过简单的双塔结构实现图像文本对齐。未来需要探索更复杂的网络结构,增强模型对视觉与语义的联合建模能力,从而实现对复杂跨模态场景的深度理解。此外,除语言外,也可以考虑与其他模态如音频、视频的结合。

3)更丰富的视觉任务覆盖。现有视觉基础模型主要聚焦在图像分类、目标检测等任务上,未来需要扩展到更多视觉子领域,如图像增强、图像修复、运动跟踪和三维建模等,构建覆盖更为全面的通用视觉基础模型。

4)数据驱动对基础模型性能提升具有关键作用。除了图像数据外,高质量的结构化知识库也可以为视觉模型注入背景知识和先验信息,帮助更好地理解视觉场景。如何获得更大规模且质量更高的视觉训练数据是需要继续努力的方向。

5)目前大规模视觉基础模型参数量巨大,计算耗费大。模型压缩与优化技术可以降低其计算和存储成本,使其更好地部署到实际场景中。如何在保证性能的前提下进行模型压缩也是一个可探索的方向。

通过进一步的研究与探索,视觉基础模型有望走向通用化和智能化,真正解决现实世界中的复杂视觉问题。模型的通用能力和泛化性能还需要不断提升,也需要研究者们持续努力。

6 结 语

本文综述了当前视觉基础模型的研究现状与发展趋势。首先追溯了计算机视觉学科从CNN到ViT的发展历程,并着重介绍了如Transformer这类关键模型结构对构建高效视觉基础模型的基础性贡献。此外,本文还探讨了各类预训练任务设计的多样性,并对当前流行的视觉基础模型进行了系统分类。在论文的后半部分,本文对视觉基础模型领域面临的挑战及未来可能的发展方向,包括评价指标的完善、跨模态能力的增强以及任务覆盖范围的扩大等方面,进行了深入的探讨。

总体来看,构建能够高效适应不同场景的视觉基础模型依然是一个充满挑战的任务。目前,采用数据驱动的大规模预训练方法被认为是实现这一目标的有效手段之一。未来,视觉基础模型的发展不仅需要在更多元的任务上进行全面评估,还应考虑从多种模态中获取更丰富的监督信号。随着通用视觉主干架构的不断完善和模型规模的逐步扩大,预计在不远的将来,视觉基础模型将具备更强大的跨任务和跨模态适配能力。视觉基础模型作为人工智能领域的关键分支,对于推动计算机视觉、模式识别等相关学科的发展具有深远的影响。这些模型的持续优化与进步,不仅极大地提升了处理和分析视觉信息的能力,而且为自动化、智能决策以及人机交互等多个应用领域提供了强大的技术支持。

参考文献(References)

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T D, Gong Z T, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Bińkowski M, Barreira R, Vinyals O, Zisserman A and Simonyan K. 2022. Flamingo: a visual language model for few-shot learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 23716-23736

Bao H B, Dong L, Piao S H and Wei F R. 2022. BEiT: BERT pre-training of image Transformers [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2106.08254.pdf

Bao H B, Dong L, Wei F R, Wang W H, Yang N, Liu X D, Wang Y, Piao S H, Gao J F, Zhou M and Hon H W. 2020. UniLMv2: pseudo-masked language models for unified language model pre-training//Proceedings of the 37th International Conference on Machine Learning. Virtually: JMLR.org: 642-652

Bommasani R, Hudson D A, Adeli E, Altman R, Arora S, von Arx S, Bernstein M S, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis J Q, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho D E, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh P W, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li X L, Li X C, Ma T Y, Malik A, Manning C D, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles J C, Nilforoshan H,

Nyarko J, Ogut G, Orr L, Papadimitriou I, Park J S, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H Y, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas A W, Tramèr F, Wang R E, Wang W, Wu B H, Wu J J, Wu Y H, Xie S M, Yasunaga M, You J X, Zaharia M, Zhang M, Zhang T Y, Zhang X K, Zhang Y H, Zheng L, Zhou K and Liang P. 2022. On the opportunities and risks of foundation models [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2108.07258.pdf

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1877-1901

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with Transformers//Proceedings of the 16th European Conference Computer Vision—ECCV 2020. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Caron M, Misra I, Mairal J, Goyal P, Bojanowski P and Joulin A. 2020. Unsupervised learning of visual features by contrasting cluster assignments//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 9912-9924

Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P and Joulin A. 2021. Emerging properties in self-supervised vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9630-9640 [DOI: 10.1109/ICCV48922.2021.00951]

Changpinyo S, Sharma P, Ding N and Soricut R. 2021. Conceptual 12M: pushing web-scale image-text pre-training to recognize long-tail visual concepts//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3557-3567 [DOI: 10.1109/CVPR46437.2021.00356]

Chen J N, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021a. TransUNet: Transformers make strong encoders for medical image segmentation [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2102.04306.pdf

Chen M, Radford A, Child R, Wu J, Jun H, Luan D and Sutskever I. 2020a. Generative pretraining from pixels//Proceedings of the 37th International Conference on Machine Learning. Virtually: JMLR.org: 1691-1703

Chen T, Kornblith S, Norouzi M and Hinton G. 2020b. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Virtually: JMLR.org: 1597-1607

Chen T, Saxena S, Li L L, Fleet D J and Hinton G. 2022. Pix2seq: a language modeling framework for object detection [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2109.10852.pdf

Chen X, Wang X, Changpinyo S, Piergiovanni A J, Padlewski P, Salz D, Goodman S, Grycner A, Mustafa B, Beyer L, Kolesnikov A, Puigcerver J, Ding N, Rong K R, Akbari H, Mishra G, Xue L T, Thapliyal A, Bradbury J, Kuo W C, Seyedhosseini M, Jia C, Ayan B K, Riquelme C, Steiner A, Angelova A, Zhai X H, Houlsby N and Soricut R. 2023. PaLI: a jointly-scaled multilingual language-image model [EB/OL]. [2023-12-06]. http://arxiv.org/pdf/2209.06794.pdf

Chen X K, Ding M Y, Wang X D, Xin Y, Mo S T, Wang Y H, Han S M, Luo P, Zeng G and Wang J D. 2024. Context autoencoder for self-supervised representation learning. International Journal of Computer Vision, 132(1): 208-223 [DOI: 10.1007/s11263-023-01852-4]

Chen X L and He K M. 2021. Exploring simple Siamese representation learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15745-15753 [DOI: 10.1109/CVPR46437.2021.01549]

Chen X L, Xie S N and He K M. 2021b. An empirical study of training self-supervised vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9620-9629 [DOI: 10.1109/ICCV48922.2021.00950]

Cho K, van Merrienboer B, Bahdanau D and Bengio Y. 2014. On the properties of neural machine translation: encoder-decoder approaches [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1409.1259.pdf

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung H W, Sutton C, Gehrmann S, Schuh P, Shi K S, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P C, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai A M, Pillai T S, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z W, Wang X Z, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S and Fiedel N. 2023. PaLM: scaling language modeling with pathways. Journal of Machine Learning Research, 24(240): 1-113

Dai Z H, Liu H X, Le Q V and Tan M X. 2021. CoAtNet: marrying convolution and attention for all data sizes//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 3965-3977

Dehghani M, Djolonga J, Mustafa B, Padlewski P, Heek J, Gilmer J,

Steiner A, Caron M, Geirhos R, Alabdulmohsin I, Jenatton R, Beyer L, Tschannen M, Arnab A, Wang X, Riquelme C, Minderer M, Puigcerver J, Evci U, Kumar M, van Steenkiste S, Elsayed G F, Mahendran A, Yu F, Oliver A, Huot F, Bastings J, Collier M P, Gritsenko A A, Birodkar V, Vasconcelos C, Tay Y, Mensink T, Kolesnikov A, Pavetić F, Tran D, Kipf T, Lučić M, Zhai X H, Keysers D, Harmsen J and Houlsby N. 2023. Scaling vision Transformers to 22 billion parameters//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 7480-7512

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR. 2009. 5206848]

Desai K and Johnson J. 2021. VirTex: learning visual representations from textual annotations//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11157-11168 [DOI: 10.1109/CVPR46437.2021. 01101]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional Transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/n19-1423]

Doersch C, Gupta A and Efros A A. 2015. Unsupervised visual representation learning by context prediction//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1422-1430 [DOI: 10.1109/ICCV.2015.167]

Donahue J, Jia Y Q, Vinyals O, Hoffman J, Zhang N, Tzeng E and Darrell T. 2014. DeCAF: a deep convolutional activation feature for generic visual recognition//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR.org: I-647-I-655

Dong L, Yang N, Wang W H, Wei F R, Liu X D, Wang Y, Gao J F, Zhou M and Hon H W. 2019. Unified language model pre-training for natural language understanding and generation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13063-13075

Dong X Y, Bao J M, Zhang T, Chen D D, Zhang W M, Yuan L, Chen D, Wen F, Yu N H and Guo B N. 2023. PeCo: perceptual codebook for BERT pre-training of vision Transformers//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 552-560 [DOI: 10.1609/aaai.v37i1.25130]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2010.11929.pdf

Duong L, Cohn T, Bird S and Cook P. 2015. Low resource dependency parsing: cross-lingual parameter sharing in a neural network parser//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics: 845-850 [DOI: 10.3115/v1/P15-2139]

El-Nouby A, Izacard G, Touvron H, Laptev I, Jégou H and Grave E. 2021. Are large-scale datasets necessary for self-supervised pre-training? [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2112.10740.pdf

Fang Y X, Dong L, Bao H B, Wang X G and Wei F R. 2023a. Corrupted image modeling for self-supervised visual pre-training [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2202.03382.pdf

Fang Y X, Sun Q, Wang X G, Huang T J, Wang X L and Cao Y. 2023b. EVA-02: a visual representation for neon genesis [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2303.11331.pdf

Fang Y X, Wang W, Xie B H, Sun Q, Wu L, Wang X G, Huang T J, Wang X L and Cao Y. 2023c. EVA: exploring the limits of masked visual representation learning at scale//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 19358-19369 [DOI: 10.1109/CVPR52729.2023.01855]

Fedus W, Zoph B and Shazeer N. 2022. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research, 23(1): 5232-5270

Gao Y, Bai H P, Jie Z Q, Ma J Y, Jia K and Liu W. 2020. MTL-NAS: task-agnostic neural architecture search towards general-purpose multi-task learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11540-11549 [DOI: 10.1109/CVPR42600.2020.01156]

Goyal P, Caron M, Lefaudeux B, Xu M, Wang P C, Pai V, Singh M, Liptchinsky V, Misra I, Joulin A and Bojanowski P. 2021. Self-supervised pretraining of visual features in the wild [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2103.01988.pdf

Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, Doersch C, Pires B Á, Guo Z D, Azar M G, Piot B, Kavukcuoglu K, Munos R and Valko M. 2020. Bootstrap your own latent a new approach to self-supervised learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 21271-21284

Gupta T, Kamath A, Kembhavi A and Hoiem D. 2022. Towards general purpose vision systems: an end-to-end task-agnostic vision-language architecture//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16378-16388 [DOI: 10.1109/CVPR52688.2022.01591]

Hadsell R, Chopra S and LeCun Y. 2006. Dimensionality reduction by learning an invariant mapping//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). New York, USA: IEEE: 1735-1742 [DOI: 10.1109/CVPR.2006.100]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M and Gelly S, 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2790-2799

Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, Le Q V, Sung Y H, Li Z and Duerig T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 4904-4916

Jiang D and Ye M. 2023. Transformer network for cross-modal text-to-image person re-identification. Journal of Image and Graphics, 28(5): 1384-1395 (姜定, 叶茫. 2023. 面向跨模态文本到图像行人重识别的Transformer网络. 中国图象图形学报, 28(5): 1384-1395) [DOI: 10.11834/jig.220620]

Joulin A, Van Der Maaten L, Jabri A and Vasilache N. 2016. Learning visual features from large weakly supervised data//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 67-84 [DOI: 10.1007/978-3-319-46478-7_5]

Kolesnikov A, Beyer L, Zhai X H, Puigcerver J, Yung J, Gelly S and Houlsby N. 2020a. Big transfer (BiT): general visual representation learning [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1912.11370.pdf

Kolesnikov A, Beyer L, Zhai X H, Puigcerver J, Yung J, Gelly S and Houlsby N. 2020b. Big transfer (BiT): general visual representation learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 491-507 [DOI: 10.1007/978-3-030-58558-7_29]

Kornblith S, Shlens J and Le Q V. 2019. Do better ImageNet models transfer better?//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2656-2666 [DOI: 10.1109/CVPR.2019.00277]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Kudugunta S, Huang Y P, Bapna A, Krikun M, Lepikhin D, Luong M T and Firat O. 2021. Beyond distillation: task-level mixture-of-experts for efficient inference//Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics: 3577-3599 [DOI: 10.18653/V1/2021.FINDINGS-EMNLP.304]

Lepikhin D, Lee H, Xu Y Z, Chen D H, Firat O, Huang Y P, Krikun M, Shazeer N and Chen Z F. 2020. GShard: scaling giant models with conditional computation and automatic sharding [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2006.16668.pdf

Lester B, Al-Rfou R and Constant N. 2021. The power of scale for parameter-efficient prompt tuning//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics: 3045-3059 [DOI: 10.18653/V1/2021.EMNLP-MAIN.243]

Li A, Jabri A, Joulin A and van der Maaten L. 2017. Learning visual N-grams from web data//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4193-4202 [DOI: 10.1109/ICCV.2017.449]

Li L H, Zhang P C, Zhang H T, Yang J W, Li C Y, Zhong Y W, Wang L J, Yuan L, Zhang L, Hwang J N, Chang K W and Gao J F. 2022. Grounded language-image pre-training//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10955-10965 [DOI: 10.1109/CVPR52688.2022.01069]

Lin J Y, Men R, Yang A, Zhou C, Ding M, Zhang Y C, Wang P, Wang A, Jiang L, Jia X Y, Zhang J, Zhang J W, Zou X, Li Z K, Deng X D, Liu J, Xue J B, Zhou H L, Ma J X, Yu J, Li Y, Lin W, Zhou J R, Tang J and Yang H X. 2021. M6: a Chinese multi-modal pretrainer [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2103.00823.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu P F, Qiu X P and Huang X J. 2016. Recurrent neural network for text classification with multi-task learning//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York,

USA: AAAI Press: 2873-2879

Liu S K, Fan L X, Johns E, Yu Z D, Xiao C W and Anandkumar A. 2024. Prismer: a vision-language model with an ensemble of experts [EB/OL]. [2024-01-07]. https://doi.org/10.48550/arXiv.2303.02506

Liu X D, He P C, Chen W Z and Gao J F. 2019. Multi-task deep neural networks for natural language understanding//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 4487-4496 [DOI: 10.18653/V1/P19-1441]

Liu Z, Hu H, Lin Y T, Yao Z L, Xie Z D, Wei Y X, Ning J, Cao Y, Zhang Z, Dong L, Wei F R and Guo B N. 2022. Swin Transformer V2: scaling up capacity and resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 11999-12009 [DOI: 10.1109/CVPR52688.2022.01170]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Long M S, Cao Z J, Wang J M and Yu P S. 2017. Learning multiple tasks with multilinear relationship networks//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 1593-1602

Lu J S, Clark C, Zellers R, Mottaghi R and Kembhavi A. 2023. UNIFIED-IO: a unified model for vision, language, and multi-modal tasks//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Lu Y X, Kumar A, Zhai S F, Cheng Y, Javidi T and Feris R. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1131-1140 [DOI: 10.1109/CVPR.2017.126]

Mahajan D, Girshick R, Ramanathan V, He K M, Paluri M, Li Y X, Bharambe A and van der Maaten L. 2018. Exploring the limits of weakly supervised pretraining//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 185-201 [DOI: 10.1007/978-3-030-01216-8_12]

Misra I, Shrivastava A, Gupta A and Hebert M. 2016. Cross-stitch networks for multi-task learning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 3994-4003 [DOI: 10.1109/CVPR.2016.433]

Mori Y, Takahashi H and Oka R. 1999. Image-to-word transformation based on dividing and vector quantizing images with words [EB/OL]. [2024-01-07]. https://www. semanticscholar. org/paper/Image-to-word-transformation-based-on-dividing-and-Mori-Takahashi/8b29ffb4207435540ddecf4b14a8a32106b33830

Noroozi M and Favaro P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles//Proceedings of the 14th European Conference on Computer Vision—ECCV 2016. Amsterdam, the Netherlands: Springer: 69-84 [DOI: 10.1007/978-3-319-46466-4_5].

Ordonez V, Kulkarni G and Berg T L. 2011. Im2Text: describing images using 1 million captioned photographs//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain: Curran Associates Inc.: 1143-1151

Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A and Tran D. 2018. Image Transformer//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 4055-4064

Patterson D, Gonzalez J, Le Q V, Liang C, Munguia L M, Rothchild D, So D, Texier M and Dean J. 2021. Carbon emissions and large neural network training [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2104.10350.pdf

Peng X Y, Wang K, Zhu Z, Wang M and You Y. 2022a. Crafting better contrastive views for Siamese representation learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16010-16019 [DOI: 10.1109/CVPR52688.2022.01556]

Peng Z L, Dong L, Bao H B, Ye Q X and Wei F R. 2022b. BEiT v2: masked image modeling with vector-quantized visual Tokenizers [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2208.06366.pdf

Peters M E, Neumann M, Iyyer M, Gardner M, Clark C, Lee K and Zettlemoyer L. 2018. Deep contextualized word representations// Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA: Association for Computational Linguistics: 2227-2237 [DOI: 10.18653/V1/N18-1202]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8748-8763

Radford A and Narasimhan K. 2018. Improving language understanding by generative pre-training [EB/OL]. [2024-01-07]. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I. 2019. Language models are unsupervised multitask learners [EB/OL]. [2024-01-07]. https://www. semanticscholar. org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y

Q, Li W and Liu P J. 2023. Exploring the limits of transfer learning with a unified text-to-text Transformer [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1910.10683.pdf

Razavian A S, Azizpour H, Sullivan J and Carlsson S. 2014. CNN features off-the-shelf: an astounding baseline for recognition//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, USA: IEEE: 512-519 [DOI: 10.1109/CVPRW.2014.131]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1804.02767.pdf

Riquelme C, Puigcerver J, Mustafa B, Neumann M, Jenatton R, Pinto A S, Keysers D and Houlsby N. 2021. Scaling vision with sparse mixture of experts//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 8583-8595

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Sariyildiz M B, Perez J and Larlus D. 2020. Learning visual representations with caption annotations//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 153-170 [DOI: 10.1007/978-3-030-58598-3_10]

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B and Rueckert D. 2019. Attention gated networks: learning to leverage salient regions in medical images. Medical Image Analysis, 53: 197-207 [DOI: 10.1016/j.media.2019.01.012]

Shao J, Chen S Y, Li Y G, Wang K, Yin Z F, He Y, Teng J N, Sun Q H, Gao M Y, Liu J H, Huang G S, Song G L, Wu Y C, Huang Y M, Liu F G, Peng H, Qin S, Wang C Y, Wang Y J, He C H, Liang D, Liu Y, Yu F W, Yan J J, Lin D H, Wang X G and Qiao Y. 2022. INTERN: a new learning paradigm towards general vision [EB/OL]. [2022-09-20]. http://arxiv.org/pdf/2111.08687.pdf

Shao S, Li Z M, Zhang T Y, Peng C, Yu G, Zhang X Y, Li J and Sun J. 2019. Objects365: a large-scale, high-quality dataset for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8429-8438 [DOI: 10.1109/ICCV.2019.00852]

Sharma P, Ding N, Goodman S and Soricut R. 2018. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics: 2556-2565 [DOI: 10.18653/v1/P18-1238]

Shi Z H, Li C J, Zhou L, Zhang Z J, Wu C W, You Z Z and Ren W Q. 2023. Survey on Transformer for image classification. Journal of Image and Graphics, 28(9): 2661-2692 (石争浩, 李成建, 周亮, 张治军, 仵晨伟, 尤珍臻, 任文琦. 2023. Transformer驱动的图像分类研究进展. 中国图象图形学报, 28(9): 2661-2692) [DOI: 10.11834/jig.220799]

Strubell E, Ganesh A and McCallum A. 2019. Energy and policy considerations for deep learning in NLP//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 3645-3650 [DOI: 10.18653/V1/P19-1355]

Subramanian S, Trischler A, Bengio Y and Pal C J. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1804.00079.pdf

Sun Q, Fang Y X, Wu L, Wang X L and Cao Y. 2023. EVA-CLIP: improved training techniques for CLIP at scale [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2303.15389.pdf

Tian Y J, Xie L X, Wang Z Z, Wei L H, Zhang X P, Jiao J B, Wang Y W, Tian Q and Ye Q X. 2023. Integrally pre-trained Transformer pyramid networks//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18610-18620 [DOI: 10.1109/CVPR52729.2023.01785]

Tian Y L, Sun C, Poole B, Krishnan D, Schmid C and Isola P. 2020. What makes for good views for contrastive learning?//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6827-6839

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jegou H. 2021. Training data-efficient image Transformers and distillation through attention//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 10347-10357

van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/1807.03748.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang J F, Yang Z Y, Hu X W, Li L J, Lin K, Gan Z, Liu Z C, Liu C and Wang L J. 2022a. GIT: a generative image-to-text Transformer for vision and language [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2205.14100.pdf

Wang P, Yang A, Men R, Lin J Y, Bai S, Li Z K, Ma J X, Zhou C, Zhou J R and Yang H X. 2022b. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 23318-23340

Wang W H, Bao H B, Dong L, Bjorck J, Peng Z L, Liu Q, Aggarwal K, Mohammed O K, Singhal S, Som S and Wei F R. 2022c. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2208.10442.pdf

Wang W H, Dai J F, Chen Z, Huang Z H, Li Z Q, Zhu X Z, Hu X W, Lu T, Lu L W, Li H S, Wang X G and Qiao Y. 2023. InternImage: exploring large-scale vision foundation models with deformable convolutions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14408-14419 [DOI: 10.1109/CVPR52729.2023.01385]

Wei C, Fan H Q, Xie S N, Wu C Y, Yuille A and Feichtenhofer C. 2022a. Masked feature prediction for self-supervised visual pre-training//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14648-14658 [DOI: 10.1109/CVPR52688.2022.01426]

Wei L H, Xie L X, Zhou W G, Li H Q and Tian Q. 2022b. MVP: multimodality-guided visual pre-training//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 337-353 [DOI: 10.1007/978-3-031-20056-4_20]

Wu H P, Xiao B, Codella N, Liu M C, Dai X Y, Yuan L and Zhang L. 2021. CvT: introducing convolutions to vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 22-31 [DOI: 10.1109/ICCV48922.2021.00009]

Wu Z R, Xiong Y J, Yu S X and Lin D H. 2018. Unsupervised feature learning via non-parametric instance discrimination//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3733-3742 [DOI: 10.1109/CVPR.2018.00393]

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with Transformers//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 12077-12090

Xie Q Z, Luong M T, Hovy E and Le Q V. 2020. Self-training with noisy student improves ImageNet classification//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR42600.2020.01070]

Xue L T, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A and Raffel C. 2021. mT5: a massively multilingual pre-trained text-to-text Transformer//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics: 483-498 [DOI: 10.18653/v1/2021.naacl-main.41]

Yang Z Y, Gan Z, Wang J F, Hu X W, Ahmed F, Liu Z C, Lu Y M and Wang L J. 2022. UniTAB: unifying text and box outputs for grounded vision-language modeling//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 521-539 [DOI: 10.1007/978-3-031-20059-5_30]

Yao L W, Huang R H, Hou L, Lu G S, Niu M Z, Xu H, Liang X D, Li Z G, Jiang X and Xu C J. 2022. FILIP: fine-grained interactive language-image pre-training//Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net

Young P, Lai A, Hodosh M and Hockenmaier J. 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2: 67-78 [DOI: 10.1162/tacl_a_00166]

Yu J H, Wang Z R, Vasudevan V, Yeung L, Seyedhosseini M and Wu Y H. 2022. CoCa: contrastive captioners are image-text foundation models [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2205.01917.pdf

Zeng W, Ren X Z, Su T, Wang H, Liao Y, Wang Z W, Jiang X, Yang Z Z, Wang K S, Zhang X D, Li C, Gong Z Y, Yao Y F, Huang X J, Wang J, Yu J F, Guo Q, Yu Y, Zhang Y, Wang J, Tao H T, Yan D S, Yi Z X, Peng F, Jiang F Q, Zhang H, Deng L F, Zhang Y H, Lin Z, Zhang C, Zhang S J, Guo M Y, Gu S Z, Fan G J, Wang Y W, Jin X F, Liu Q and Tian Y H. 2021. PanGu-α: large-scale autoregressive pretrained Chinese language models with auto-parallel computation [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2104.12369.pdf

Zhai X H, Kolesnikov A, Houlsby N and Beyer L. 2022a. Scaling vision Transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1204-1213 [DOI: 10.1109/CVPR52688.2022.01179]

Zhai X H, Wang X, Mustafa B, Steiner A, Keysers D, Kolesnikov A and Beyer L. 2022b. LiT: zero-shot transfer with locked-image text tuning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18102-18112 [DOI: 10.1109/CVPR52688.2022.01759]

Zhang H Y, Wang T B, Li M Z, Zhao Z, Pu S L and Wu F. 2022. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of Image and Graphics, 27(9): 2652-2682 (张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞. 2022. 视觉语言多模态预训练综述. 中国图象图形学报, 27(9): 2652-2682) [DOI: 10.11834/jig.220173]

Zhang X S, Tian Y J, Huang W, Ye Q X, Dai Q, Xie L X and Tian Q. 2022a. HiViT: hierarchical vision Transformer meets masked image modeling [EB/OL]. [2024-01-07]. http://arxiv.org/pdf/2205.14949.pdf

Zhang Y H, Jiang H, Miura Y, Manning C D and Langlotz C P. 2022b. Contrastive learning of medical visual representations from paired images and text//Proceedings of the 7th Machine Learning for Healthcare Conference. Durham, USA: PMLR: 2-25

Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, Fu Y W, Feng J F, Xiang T, Torr P H S and Zhang L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6877-6886 [DOI: 10.1109/CVPR46437.2021.00681]

Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A and Torralba A. 2017. Scene parsing through ADE20K dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5122-5130 [DOI: 10.1109/CVPR.2017.544]

Zhou K Y, Yang J K, Loy C C and Liu Z W. 2022. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9): 2337-2348 [DOI: 10.1007/s11263-022-01653-1]

Zhu J G, Zhu X Z, Wang W H, Wang X H, Li H S, Wang X G and Dai J F. 2022. Uni-Perceiver-MoE: learning sparse generalist models with conditional MoEs//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 2664-2678

作者简介

张燚钧,男,博士,高级工程师,主要研究方向为计算机视觉、基础大模型和大规模预训练。E-mail: yijzhangme@gmail.com

齐骥,通信作者,男,博士,主要研究方向为计算机视觉、智能安防和云计算。E-mail: qiji@cmss.chinamobile.com

张润清,男,博士,主要研究方向为机器学习和目标检测。E-mail: zhangrunqing@cmss.chinamobile.com

周华健,男,硕士,主要研究方向为计算机视觉、目标检测和小样本学习。E-mail: zhouhuajian@cmss.chinamobile.com

余肇飞,男,助理教授,博士生导师,主要研究方向为类脑视觉计算和脉冲视觉。E-mail: yuzf12@pku.edu.cn

黄铁军,男,教授,博士生导师,主要研究方向为视频编解码和通用人工智能。E-mail: tjhuang@pku.edu.cn