Journal of Image and Signal Processing 图像与信号处理, 2022, 11(3), 144-161
Published Online July 2022 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2022.113016

Hans汉斯

基于深度学习的视觉关系检测方法及应用

汤婧婧[1]，黄　晶[2]，石爱业[1]，张丽丽[1]，徐立中[1]

[1]河海大学计算机与信息学院，江苏　南京
[2]河海大学商学院，江苏　南京

收稿日期：2022年7月1日；录用日期：2022年7月12日；发布日期：2022年7月25日

摘　要

随着深度学习的不断发展和广泛应用，计算机视觉的许多领域也得到了长足的进步，例如在图像分类、对象检测、图像分割等任务中的表现。视觉关系检测(VRD)是计算机视觉的重要任务，旨在识别图像中物体之间的关系或相互作用，这对于理解图像及视觉世界都很重要，VRD也是计算机视觉技术应用研究的关键环节。与一般的物体检测任务相比，VRD不仅需要预测每个物体的类别和轨迹，还需要预测物体之间的关系，研究人员已经针对改任务提出了很多办法，特别在近年来基于深度神经网络的发展的深度学习也有所突破。本文介绍了VRD任务的内容，深度学习基本方法，VRD的传统方法和基于深度学习模型的一些分类和框架及其VRD在计算机视觉领域的应用。

关键词

计算机视觉，深度学习，神经网络，视觉关系检测

Method and Application of Visual Relationship Detection Based on Deep Learning

Jingjing Tang[1], Jing Huang[2], Aiye Shi[1], Lili Zhang[1], Lizhong Xu[1]

[1]College of Computer and Information, Hohai University, Nanjing Jiangsu
[2]Business School, Hohai University, Nanjing Jiangsu

Received: Jul. 1st, 2022; accepted: Jul. 12th, 2022; published: Jul. 25th, 2022

Abstract

With the continuous development and wide application of deep learning, many fields of computer

文章引用: 汤婧婧, 黄晶, 石爱业, 张丽丽, 徐立中. 基于深度学习的视觉关系检测方法及应用[J]. 图像与信号处理, 2022, 11(3): 144-161. DOI: 10.12677/jisp.2022.113016

vision have also made great progress, such as performance in image classification, object detection, image segmentation and other tasks. Visual relationship detection (VRD) is an important task for computer vision, aiming to recognize relations or interactions between objects in an image, which is important for understanding images even the visual world. Compared with the general object detection task, VRD requires not only to predict the categories and trajectories of each object, but also to predict the relationship between objects. Researchers have proposed to tackle this problem especially with the development of deep neural networks in recent years. In this survey, we provide a comprehensive review of VRD in computer vision and some categorization and frameworks of deep learning models for VRD with its applications.

Keywords

Computer Vision, Deep Learning, Neural Networks, Visual Relationship Detection

1. 引言

人脑是由 10^9~10^{11} 个细胞极其大约 10^{15} 突触互相连接高度复杂的脑网络，通过视觉、触觉、嗅觉和听觉感知外部事物的信息，其中视觉信息是主要来源，基于人类视觉仿生学的计算机视觉认知技术是现实世界应用的体现。到目前为止，视觉认知技术主要包括目标检测[1]、语义分割[2]、图像分类[3]、视觉关系检测(VRD) [4]和视觉问答(VQA) [5]等。

VRD 是对图像中成对物体的预测，它是计算机视觉中的重要任务之一，对于理解图形，连接图像和文本具有重要意义。视觉认知技术使计算机能够理解真实世界，与计算机科学与技术、人工智能等在许多领域的实际应用有关，如模式识别、图像处理等，也与视觉关系处理密切相关。视觉关系检测如图 1 所示，从输入图像中得到物体及其关系，获取更丰富的图像信息，达到更深层次的图像理解。

Figure 1. Visual relationships detection in images
图 1. 图像中的视觉关系检测

VRD 是一种中级视觉任务，可以从低级视觉任务(物体检测和识别)中获取信息，并有助于高级视觉任务(例如 VQA，看图说话，视觉推理等)。在 VRD 研究的早期阶段，因为图像中的物体可能没有正确定位，给定物体之间的关系可能没有完全标记。此外，物体关系可以通过多种方式指定，数据集中物体相互关系包括几何关系、语义关系、从属关系和其他关系，它们的外观会发生很大变化，而且关系的分布比物体的分布长得多。很难为所有可能的关系获得足够的示例训练，导致在 VRD 的早期作品中只发现了少数关系。随着技术和人工智能的不断发展，特别是与深度学习的融合，VRD 的技术改进正在突飞猛进。由于其丰富的特征表示，深度学习近年来已成为研究热点，并被广泛用于多种任务，也包括视觉相关的应用，例如：识别、医学图像分割、姿态估计、看图说话、视频描述生成、图像风格迁移、以及跨模态检索等。基于深度学习的视觉关系检测方法也得到了发展，具有不同的模型、相应的目标函数和算法。除了 VRD 中的物体检测外，还通过条件随机场(CRF)网络，关系检测网络(RePN)，视觉转化嵌入(VTransE)网络[6]和基于知识的特征细化等学习模块来研究关系检测。

由于科学的快速发展和信息技术的不断变化，基于深度学习的 VRD 理论研究和实践创造不断涌现。本文对深度学习在视觉关系检测中的研究进展进行文献综述，介绍视觉关系检测的研究现状，深度学习基本模型，视觉关系检测方法及其应用。

2. 背景及早期研究介绍

随着计算机视觉领域的不断发展，对图像语义的理解变得越来越重要，但是直接从图像中学习完整的高级语义难度较大。视觉关系检测可以提高计算机对图像更深层次的理解，更好地支持高层次语义信息的计算机视觉任务，已经引起越来越多的关注。在早期视觉关系研究中，是以改善对象检测性能为目的，只引入了对象对之间的共同关系，例如位置和大小的比较。应用 Max-Margin Learning [7]和结构化学习[8]等方法，来检测这些少数视觉关系，包括空间对象－对象相互作用，介词和比较形容词关系以及人类对象相互作用(HOIS)，大部分工作都利用物体之间的交互(例如物体共生，空间关系)来改善视觉任务[9]，这些研究忽略了非空间关系。在面向图像的视觉关系检测中，这些方法是从每张图像中选择重要特征是必要步骤。而随着类别数量的增加，特征提取变得越来越麻烦。要确定哪些特征最能描述不同的视觉关系类别，取决于 CV 工程师的判断和长期试错。此外，每个特征定义还需要处理大量参数，所有参数必须由 CV 工程师进行调整。Cewu Lu 等人(2016)将视觉关系预测形式化为一项任务，并提供了具有中等数量关系的数据集 VRD [4]，包含 5000 张图片，其中有 100 个物体类别和 70 个关系类别，总共 37993 个关系标注，6672 个关系三元组。

图像中物体关系识别的目的是根据计算机识别的物体类型来学习和识别物体之间的语义关系，物体关系通常以〈主语－谓词－宾语〉的三元组形式表示。主语和宾语都是物体的类型，谓词表示两者之间的语义关系[6]。通过对物体关系三元组的不断研究，可以在视觉问答、看图说话等高级图像理解任务中取得突破。然而，物体关系检测任务需要将所有信息结合起来，对相对主观的语义关系表示进行建模，这给特征学习带来了新的挑战和更高的要求。

为了实现大规模的视觉关系检测，Cewu Lu 等人(2016) [4]将关系的预测分解为两个单独的部分：检测物体和预测谓词。Bryan A. Plummer 等人(2017) [10]通过融合几个视觉特征，如外观，大小，边界框和语言线索(如描述属性信息的形容词)，在图像中建立基础短语。尽管专注于短语定位而不是视觉短语检测，但在 VRD 数据集上进行评估时，结果与文献[4]相当。最近，视觉关系检测任务有几种新的尝试：Liang Xiaodan 等人(2017) [11]提出在强化学习框架中检测关系和属性；Li Yikang 等人(2017) [12]训练了端到端系统，通过更好的物体检测来提升关系检测；Dai Bo 等人(2017) [13]通过关系建模框架来检测关系。Hu Zhiting 等人(2016) [14]将丰富的语言和视觉表示组合在一个端到端的深度神经网络中，该神经网络在训

练过程中使用师生框架吸收外部语言知识，以增强预测和泛化。与独立于关系预测检测物体的不同[4]，对物体和关系进行共同建模，而且使用语言知识来预测谓词，该语言知识模拟谓词与〈主语，宾语〉对之间的相关性。

1986 年，Rina Dechter 首次将术语“深度学习”(DL)引入机器学习(ML) [15]，2000 年，Igor N. Aizenberg 等人将其用于人工神经网络(ANN) [16]。2012 年，Geoffrey Hinton 的团队在 ImageNET 大赛的图像分类任务中获得冠军，他们的表现说明 DL 优于传统方法。近年来，使用深度神经网络的端到端学习方法取得了很大的成功，现在它被广泛用于计算机视觉、语音识别、自然语言处理、医学图像处理等领域[17]。深度学习方法使用多个层来学习不同复杂程度的数据特征，允许计算机通过从更简单的概念构建来学习复杂的概念。深度学习架构进行迭代所需的知识和专业技能取代了手动提取特征所需的知识和专业技能，提高了 VRD 传统算法的性能。

3. 深度学习基本模型

深度学习通常与人工神经网络有关,以下介绍常用的深度神经网络，但要注意其有泛化能力弱、表达能力弱、没有注意力机制、过度依赖训练数据等缺陷。

3.1. 卷积神经网络(CNN)

CNN，也称为 ConvNets，是多层神经网络，主要用于图像处理和目标检测。Yann LeCun 等人(1989) [18]创建了第一个 CNN 模型，称之为 LeNet-5。它包含三种类型的层，具有不同的卷积层，池化层和全连接层(图 2)，用于识别手写邮政编码中数字之类的字符。

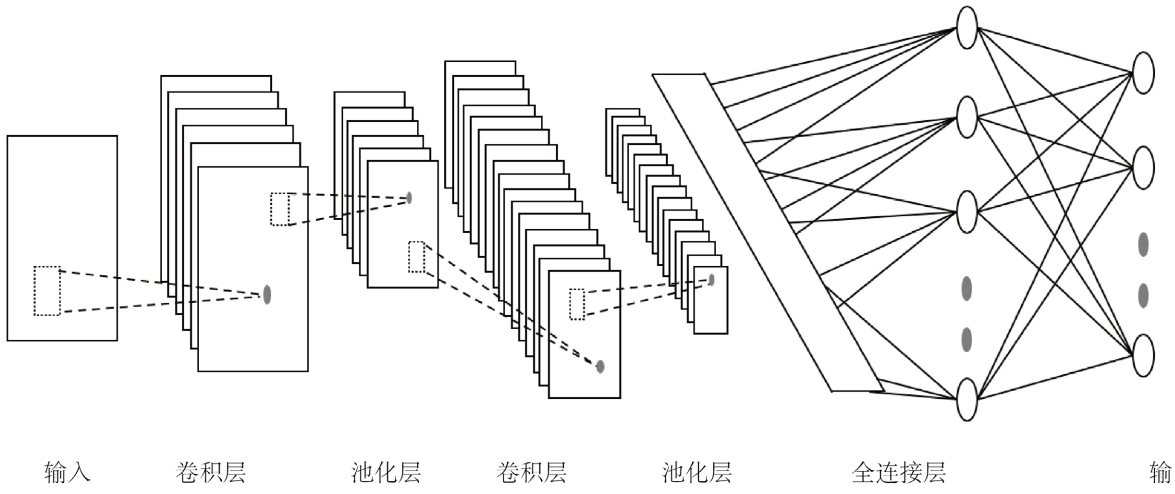

Figure 2. LeNet-5 architecture
图 2. LeNet-5 结构图

CNN 是图像处理和理解中最有效和适应性最强的模型，它能够适应图像的结构，通过结构重组和减少权值将特征抽取功能融合进多层感知器，自动的从图像中抽取出丰富的相关特性，能同时进行图像特征提取和分类。除了用于图像任务外，CNN 也可以处理具有局部空间相关性的数据，例如语音和自然语言等等。为了满足处理大规模数据的要求，CNN 模型加深了网络层数，随之产生了庞大的存储和计算量，还需要非常强大的处理器(如 CPU、GPU)支持，近年来 CNN 模型在算法层面上追求降低功耗，实现轻量化网络是进一步研究的方向。最常见的 CNN 架构除了 LeNet-5，还包括 AlexNet、VGGNet、GoogLeNet、

ZFNet、ResNet、DenseNet、NASNet 等等。

3.2. 循环神经网络(RNN)

二十世纪八十年代和九十年代神经生物学的探索中，研究人员发现大脑反应回路的兴奋和抑制受到大脑 α 节律(α-rhythm)调节和影响，并在 α-运动神经(α-motoneurones)中形成循环反馈系统[19]。在二十世纪七八十年代，一些数学模型被建立来模拟受 RNN 启发的循环反馈系统。John Hopfield (1982) [20]基于二进制节点的使用，建立了一个具有内容寻址记忆能力的神经网络来解决组合优化问题，即 Hopfield 神经网络。Michael Jordan (1997) [21]提出了一种认知模型，可以表示输出的动态特性，即 Jordan 网络。Jeffrey L. Elman (1990) [22]提出了第一个完全连接的 RNN，即 Elman 网络。Jordan 网络和 Elman 网络通过单层前馈神经网络建立递归连接，也称为简单循环网络(SRN)。Jürgen Schmidhube 等人提出了神经历史压缩器(NHC) [23]和长－短期记忆网络(LSTM) [24]。Mike Schuster and Kuldip K. Paliwal (1997) [25]提出了具有深层结构的双向 RNN (BRNN)，并进行了语音识别实验，其中双向 RNN 和长短时记忆网络(LSTM)是近年来常见的循环神经网络。

SRN 中循环层的输出在下次作为该层输入的一部分被延迟，然后输出被发送到网络的后续层，同时 SRN 是一个由输入层、隐藏层和输出层组成的三层连接神经网络，其基本结构如图 3 所示。

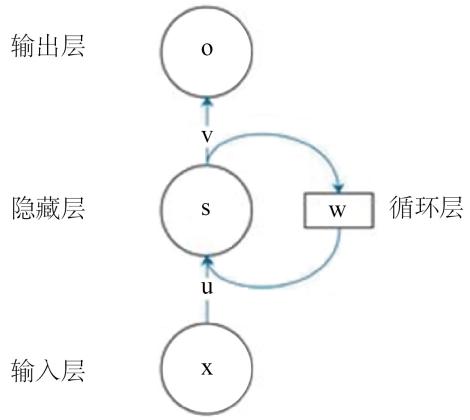

Figure 3. Simple recurrent network architecture
图 3. 简单循环网络结构

RNN 使模型上一个时间步长中生成的结果能够用作下一个时间步长的输入的一部分，并影响下一个时间步长的输出，即所谓的序列信息。根据时间轴，RNN 可以平滑扩展如图 4 所示。

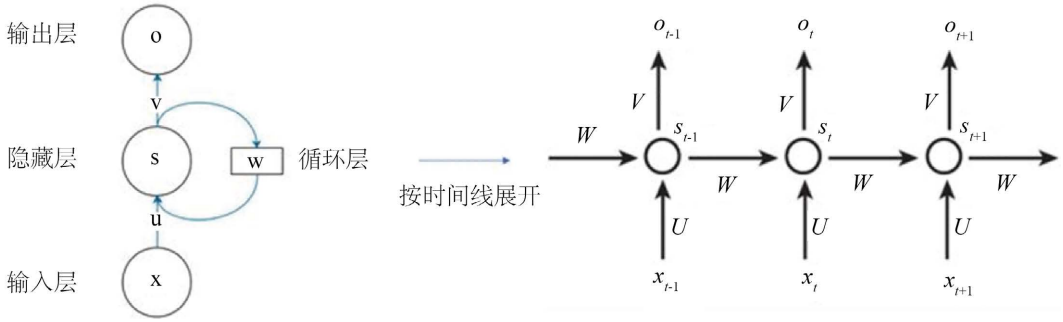

Figure 4. The architecture of simple recurrent network which unfolds by timeline
图 4. 简单循环网络时间线展开图

图 4 中循环神经网络的计算方法可以公式化为

$$O_t = g\left(V \cdot S_t\right) \tag{1}$$

$$S_t = f\left(U \cdot X_t + W \cdot S_{t-1}\right) \tag{2}$$

其中，向量 X_t， S_t 和 O_t 分别表示输入层、隐藏层和输出层的当前值，而 S_{t-1} 表示隐藏层的先前值。U 是从输入层到隐藏层的权重矩阵，V 是从隐藏层到输出层的权重矩阵。RNN 的隐藏层的值不仅取决于当前输入 X_t，还取决于隐藏层的先前值，W 是用作当前输入的隐藏层先前值的权重矩阵。

RNN 可以学习以自然有效的方式混合顺序和并行信息处理的程序，因此它可以用于以下任务中：在一对一映射中将单个输入映射到单个输出，例如，图像分类；单个输入以一对多关系映射到一系列输出，例如，看图说话；单个输出由一系列输入产生，例如，情绪分析(多个单词的二进制输出)；一组输入产生一组输出，例如，视频分类(将视频拆分为帧并单独标记每个帧)。

3.3. 生成对抗网络(GAN)

Ian Goodfellow 等人(2014) [26]提出了生成对抗网络(GAN)一般设计架构如图 5 所示。

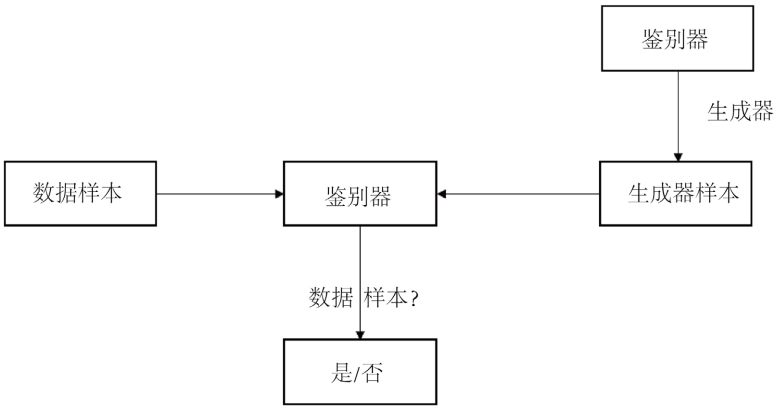

Figure 5. Conventional design of GAN model
图 5. 生成对抗网络模型的常规设计

为了学习生成器对数据 x 的分布 p_g，他们将输入噪声变量的先验表示为 $p_z\left(z\right)$，然后将到数据空间的映射表示为 $G\left(z;\theta_g\right)$，其中可微函数 G 由具有参数的多层感知器 θ_g 表示。第二个多层感知器 $D\left(x;\theta_d\right)$ 被定义为那些输出单个标量的感知器。$D\left(x\right)$ 表示来自数据 x 而不是 p_g 的概率，D 和 G 使用值函数 $V\left(D,G\right)$ 计算极小极大，该过程可由公式(3)所示

$$\min_G \max_D V\left(D,G\right) = \mathbb{E}_{x\sim p_{data}(x)}\left[\log D\left(x\right)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \tag{3}$$

实际操作中使用迭代的数值方法来实现以上过程，但是在训练的内部循环中优化 D 在计算上无法实现，并且出现在有限的数据集上会导致过度拟合的问题，所以采用在优化 D 的 k 个步骤和优化 G 的一个步骤之间进行交替，使得 G 变化足够慢，将 D 保持在接近其最优解的位置[24]。该策略与随机最大似然(SML)/持续性对比散度(PCD) [27]相似，GAN 仅使用反向传播即可获得梯度，避免了马尔可夫链产生的影响，在学习过程中省去推理的麻烦，有利于模型中多种功能的集成。

3.4. 图神经网络(GNN)

图神经网络(GNN)是在图域上运行深度学习的方法，由于其令人信服的性能，GNN 最近已成为一种

广泛应用的图形分析方法。Alessandro Sperduti 等人(1997)首次将神经网络用于有向无环图[28]，后来分别引入了递归神经网络和前馈神经网络来处理循环问题。虽然取得了成功，但这些方法背后的普遍思想是在图上构建状态转换系统并迭代直到收敛，这限制了可扩展性和表示能力。这些早期的研究属于递归图神经网络(RecGNNs)的范畴，他们通过迭代传播相邻信息直到到达稳定的不动点来学习目标节点的表示。这一过程的计算成本很高，最近人们越来越努力克服这些挑战[29]，受深度神经网络，特别是卷积神经网络在计算机视觉领域取得越来越多的成就，大量重新定义图形数据卷积概念的方法被并行开发，这些方法是基于卷积图神经网络(ConvGNNs)。ConvGNS 分为两大主流，基于光谱的方法和基于空间的方法。Joan Bruna 等人(2013) [30]提出了第一项关于基于光谱的 ConvGNS 的杰出研究，该研究基于谱图论开发了一种图卷积。自那时以来，基于光谱的 ConvGNNs 的改进、扩展和近似不断增加。基于空间的 ConvGNNs 的研究比基于光谱的 ConvGNNs 早得多。AlessioMicheli 等人(2009) [31]首先通过架构复合非递归层解决了图形的相互依赖性，同时继承了 Recgns 传递消息的思想。但是直到最近，相关学者才陆续提出许多基于空间的 ConvGNNs。

图形是一种对一组对象(节点)与其关系(边)进行建模的数据结构，通常，图形表示为 $G=(V,E)$，即顶点或节点 V 和边 E 的集合。让 $v_i \in V$ 来表示一个节点，$e_{ij}=(v_i,v_j)\in E$ 表示一条边 v_i 指向另一边 v_j。图 G 的邻接矩阵 A 是顶点 v_i 与顶点 v_j 之间的边数 a_{ij} 的集合，表示为 $A=A(G)=(a_{ij})_{n\times n}$。图可以有节点属性 X，其中 $X\in R^{n\times d}$ 节点特征矩阵表示 $x_v\in R^d$ 节点的特征向量 v。同时，一个图可以有边属性 X^e，其中 $X^e\in R^{m\times c}$ 是一个边特征矩阵，表示 $X^e_{v,u}\in R^c$ 一个边的特征向量 (v,u)。

有向图是所有边从一个节点定向到另一个节点的图形。无向图被认为是有向图的一种特例，其中如果连接了两个节点，则存在一对具有反方向的边。当且仅当邻接矩阵是对称的时，图才是无向的。时空图是一种属性图，其中节点属性随时间动态变化。时空图定义为 $G^{(t)}=(V,E,X^{(t)})X^{(t)}\in R^{n\times d}$ [32]。

GNN 的输入是图，并有多层图卷积和激活函数等各种操作，最终得到图中每个节点的表示，以方便节点分类、链接预测、图和子图生成等任务。图神经网络分为四类：递归图神经网络(RecGNNs)，卷积图神经网络(Con-vGNNs)，图自动编码器(GAEs)和时空图神经网络(STGNNs)。

3.5. 注意力机制(Attention mechanism)

日常生活中，人们看到一个场景时会关注其中的显著区域，并快速处理这些区域。为了模仿人类视觉系统的这一特点，注意力机制被引入了计算机视觉。注意力机制是将注意力转移到图像中最重要的区域，同时忽略不相关部分的方法。这种注意机制可以看作是基于输入图像特征的动态权重调整过程。注意力机制的发展可以大致分为四个阶段：

第 1 阶段采用 RNN 来构建注意力，一种代表性的方法是 RAM；第 2 阶段明确预测了重要区域，一种代表性的方法是 STN；第 3 阶段隐式完成了注意力过程，一个代表性的方法是 SENet；第 4 阶段使用自注意力方法。注意力机制过程可表述为

$$Attention=f(g(x),x) \tag{4}$$

其中 $g(x)$ 表示产生注意力，这与注意待识别区域的过程相对应。$f(g(x),x)$ 是指基于与处理关键区域和获取信息一致的注意力 $g(x)$ 来处理输入 x。一般来说，上述公式可以集中体现大多数注意力机制，包括自注意力[33]和压缩激发(SE)通道注意力[34]。自注意力 $g(x)$ 和 $f(g(x),x)$ 可以写成

$$Q,K,V=\text{Linear}(x) \tag{5}$$

$$g(x)=\text{Softmax}(QK) \tag{6}$$

$$f\left(g\left(x\right),x\right)=g\left(x\right)V \tag{7}$$

对于 SE， $g\left(x\right)$， $f\left(g\left(x\right),x\right)$ 可以写为

$$g\left(x\right)=\text{Sigmoid}\left(\text{MLP}\left(\text{GAP}\left(x\right)\right)\right) \tag{8}$$

$$f\left(g\left(x\right),x\right)=g\left(x\right)x \tag{9}$$

计算机视觉领域中注意力机制分为以下几种：通道注意力，空间注意力，时间注意力、分支注意力，通道和空间混合注意力和空间和时间混合注意力。

4. 基于深度学习的视觉关系检测方法

视觉关系检测常用的方法主要分为两种。一是以关系三元组作为检测的基本单位，把每个关系三元组作为一个范畴来训练分类器。这种方法的缺点是需要训练的类别太多，而且训练数据往往不足且分布不均匀，因此很难训练出一个好的分类器。二是将对象分类和关系分类分开，训练两种分类器。一个分类器实现对象的分类，即判断主语和宾语，另一个分类器实现关系的分类，即判断预测。这种方法克服了前一种方法的缺点，但对分类器进行训练以判断关系往往很困难。在同一关系中，主语和宾语可以不同，不同场景中图像外观可能差异很大。

4.1. 视觉短语

在目标检测任务中，在不同的场景中，要检测的某种物体的特定形状可能会有很大不同。这种被检测物体差异很大的现象，会使物体检测器难以训练。但是，如果不将对象用作检测的基本单元，而是将视觉短语用作检测的基本单元，则可以有效地改善此问题。在视觉短语中，组成元素的外观变化通常相对较小，因此更容易训练探测器。以同一视觉短语〈人 - 驾驶 - 汽车〉为例，在不同的图片中具有相对一致的外观。

在视觉关系检测的任务中，Mohammad Amin Sadeghi 等人(2011) [35]提出了视觉短语识别的方法[35]。与以物体为单元的检测方法相比，以视觉短语为单元的方法可以达到更好的检测效果。该方法从数据集 Pascal voc2008 [36]中选择了 8 种物体：人、马、狗、汽车、自行车、瓶子、椅子和沙发，然后将这些物体用于组成 17 种视觉短语，作为视觉关系检测的基本单元。视觉短语检测器在其数据集中的多类对象检测和关系推理方面表现良好。当视觉关系被视为一个短语时，它的检测可以表述为一个三个相互关联的识别问题。为了同时检测〈主语 - 谓词 - 宾语〉，Li Yikang 等人(2017) [12]提出了一种视觉短语引导卷积神经网络(ViP-CNN)，其中设计了一种短语引导消息传递结构(PMPS)来探索关系分量的连接。

4.2. 知识蒸馏

视觉关系检测的建模过程可以是检测由主语、谓词和宾语形成的关系三元组，这样可以更准确地模拟视觉关系，但关系三元组的多样性导致模型参数空间的扩展很多，现有的数据集难以满足模型的训练要求，而且数据集中一直存在长尾分布问题，增加了该模型的训练难度。为了更加充分利用图像训练数据，采用语言知识来标准化预测结果。

获取语言知识的方法有两种，一种是通过计算训练集中标记信息的条件概率 P (PRED|sub, obj)来获取语言知识，另一种是从外部知识库(如百度百科全书，维基百科全书等)获取语言知识。前者由于数据集大小限制，只能收集少部分知识，无法显著改善模型效果。后者从互联网知识库的公共文本中捕获外部语言知识，可以有效缓解长尾分布问题。外部知识库涵盖大量常用词的统计数据，这些词可用于描述主题、对象和对象之间的关系。由于外部知识库更新频率高，它包含训练数据所没有的组合形式，而且覆盖范

围更广，知识更一般，这样会产生噪音很大的问题。

Ruichi Yu 等人(2017) [37]提出语言知识蒸馏模型，利用视觉和语言表征的作用，采用内外部语言知识对端到端深度神经网络的学习过程进行标准化，增强其预测能力而且此模型可以用于小数据集上进行训练，效果良好。但是，这种方法也有其缺点，例如，输入图片中的信息未完全浏览，标准视觉模块有效地提取了视觉外观特征，但忽略了图片中物体之间的空间位置关系信息和周围物体提供的环境信息。

4.3. 语言先验

由于无法提供足够数量的训练样本用来训练数据集中所有的关系，相关研究工作将要预测的关系控制在一定范围内。在视觉关系检测的任务中，可以将语言模块添加到视觉模块中，从而引入一种先验的语言知识[4]。视觉模块首先检测物体，然后确定物体之间的谓词，最后将它们组合在一起形成关系三元组，语言模块将对象和关系映射到特征向量中来帮助进行预测。在该方法中，对于输入图片，首先使用基于区域的卷积神经网络(RCNN) [38]来检测图片中可能成为对象的区域，然后将可能的对象区域成对，将它们输入到后续的可视化模块和语言模块中，最后输出每个对象对之间最有可能的关系三元组。可视化模块和语言模块是整个方法的核心部分。可视化模块使用卷积神经网络来训练不同关系三元组的分类器，语言模块引入语义上的相似性辅助分类预测。文献[4]中的关系投影函数定义为

$$f\left(\mathcal{R}_{(i,k,j)},\mathbf{W}\right)=\mathbf{w}_k^{\mathrm{T}}\left[\text{word2vec}\left(t_i\right),\text{word2vec}\left(t_j\right)\right]+b_k \tag{10}$$

其中，word2vec()表示转换函数，并且 t_j 是第 j 个物体类别的单词。 $\mathbf{W}=\left\{\left\{\mathrm{w}_1,b_1\right\},\cdots,\left\{\mathrm{w}_k,b_k\right\}\right\}$ 是一组谓词。

Dai Bo 等人(2017) [13]提出了深度关系网络(DR-Net)来学习物体类别和关系谓词之间的关系。关系谓词和物体类别之间的统计依赖关系将有助于限制关系更加合理。基于 VRD 中的 DR-Net 的统计关系已被利用来解决由视觉或空间线索引起的歧义。关系 r 的后验概率公式为

$$\mathbf{q}_r=\sigma\left(\mathbf{W}_r\mathbf{x}_r+\mathbf{W}_{rs}\mathbf{q}_s+\mathbf{W}_{ro}\mathbf{q}_o\right) \tag{11}$$

其中 σ 表示激活函数， $\mathbf{W}_{rs}=\varphi_{rs}\left(r,s\right)$ 表示捕获关系谓词 r 和主语类别 s 之间统计关系的潜力， $\mathbf{W}_{ro}=\varphi_{ro}\left(r,o\right)$ 表示捕获关系谓词 r 和对象类别 o 之间统计关系的潜力。 $\mathbf{x}_r$ 表示合并封闭框的外观和空间配置的压缩对要素。s 和 o 的生成采用类似的方法，更新后的概率向量可以由下式表示

$$\mathbf{q}_s'=\sigma\left(\mathbf{W}_a\mathbf{x}_s+\mathbf{W}_{sr}\mathbf{q}_r+\mathbf{W}_{so}\mathbf{q}_o\right) \tag{12}$$

$$\mathbf{q}_r'=\sigma\left(\mathbf{W}_r x_r+\mathbf{W}_{rs}\mathbf{q}_s+\mathbf{W}_{ro}\mathbf{q}_o\right) \tag{13}$$

$$\mathbf{q}_0'=\sigma\left(\mathbf{W}_a x_o+\mathbf{W}_{os}\mathbf{q}_s+\mathbf{W}_{or}\mathbf{q}_r\right) \tag{14}$$

DR-Net 是通过将这种迭代更新过程展开到具有一系列计算层(即这些更新公式)的网络中来实现的。先验和后验统计将利用大规模关系检测中的词嵌入，以及语言先验已被用作许多方法中的基本组件，在对象和关系之间可以进一步利用结构化关联。

4.4. 基于 DNN 的方法

为了学习丰富多样的关系，特别是近年来随着深度学习的发展，学者们提出了大量基于 DNN 的视觉关系检测方法。Cewu Lu 等人(2016) [4]提出使用 RCNN 来检测具有语言先验的对象和谓词，以影响预测关系的可能性。从此，研究人员提出了各种模块来探索关系的具体特征，包括转化嵌入(TransE)；知识正则化；注意力模型；目标成本和损失[10]和基于强化学习(RL)的框架。

4.4.1. 转化嵌入(TransE)

关系是指对象和谓词的组合，对于 N 个对象和 K 个谓词，学习脱节关系的复杂性是 $O\left(N^2K\right)$。但是，由于相关对象，谓词的外观会发生了巨大变化。TransE 由低维向量 $\boldsymbol{s}, \boldsymbol{p}$ 和 $\boldsymbol{o}$ 分别表示〈主语－谓词－宾语〉，$\boldsymbol{s}+\boldsymbol{p} \approx \boldsymbol{o}$ 表示关系成立时关系的平移，否则由 $\boldsymbol{s}+\boldsymbol{p} \neq \boldsymbol{o}$ 表示。Zhang Hanwang 等人(2017) [6]提出了一种用于视觉关系检测的 VTransE 网络，他们通过在低维空间中映射对象和谓词的特征来建模视觉关系，并将谓词建模为主语和客体之间的平移向量。

要将特征空间投影到关系空间，矩阵 $\mathbf{W}_s$ 和 $\mathbf{W}_o$ 由 VTransE 学习。当 $\mathbf{x}_s, \mathbf{x}_o \in \mathbb{R}^M$ 表示主体和宾语的 M 维特征时，$\boldsymbol{s}$ 和 $\boldsymbol{o}$ 可以分别重写为 $\boldsymbol{s}=\mathbf{W}_s\mathbf{x}_s$ 和 $\boldsymbol{o}=\mathbf{W}_o\mathbf{x}_o$。因此，视觉转化可以表述为

$$\mathbf{W}_s\mathbf{x}_s+\mathbf{t}_p \approx \mathbf{W}_o\mathbf{x}_o \tag{15}$$

其中 $\mathbf{t}_p \in \mathbb{R}^r\left(r \ll M\right)$ 表示要学习的关系平移向量。损失函数定义为

$$L_{rel}=\sum_{(s,p,o)\in\mathcal{R}}-\log \operatorname{softmax}\left(\mathbf{t}_p^{\mathrm{T}}\left(\mathbf{W}_o\mathbf{x}_o-\mathbf{W}_s\mathbf{x}_s\right)\right) \tag{16}$$

其中 softmax 通过 p 计算。在更快速区域的卷积神经网络之后设计了一个特征提取层，以整合对象和关系之间的知识转移，包括类概率、位置(即边界框的坐标和比例)以及感兴趣区域(ROI)视觉特征。TransE 在对关系数据进行建模方面非常有效，可以应用于社交网络分析和推荐系统。

4.4.2. 知识正则化

常识知识帮助人类推理视觉关系，因此它可以用来提炼物体和关系的特征。通过从训练注释(内部)和公开可用的文本(例如维基百科(外部))中获得知识，然后计算给定的〈subject, object〉对。这些知识被提炼成深度学习模型，并在师生知识提炼框架中进行训练，使用 T-Net 和 S-Net 分别代表教师和学生网络。构建 T-Net 后，其优化函数定义为 T-Net 和 S-Net 预测分布的 KL 散度为

$$\min_{t\in T}\mathrm{KL}\left(t(Y)\,\|\,s_\phi(Y\mid X)\right)-\lambda\mathbb{E}_t\left[L(X,Y)\right] \tag{17}$$

其中 $t(\cdot)$ 和 $s_\Phi(\cdot)$ 分别表示 T-Net 和 S-Net 的预测；Φ 是 S-Net 的参数集；并且 $L(\cdot)$ 是一个约束函数。λ 是一个平衡项。

在文献[39]中，通过限制直接使用大型外部语料库的负担来处理广泛的类并增强可伸缩性。网络可以更好地利用对象和谓词类之间的统计和语义依赖关系，这些依赖关系是从预先计算的模型和训练注释中提取的。基于知识的特征细化是为了通过利用常识关系来改善特征表示。使用外部知识库，其中包括语义实体。通过使用对象标签从知识库中检索，可以识别最常见的关系。这种方法的优点是它能够解决关系检测中的长尾问题，因为外部知识由用于表示主语和宾语对之间关系的单词的统计信息组成。

4.4.3. 注意力模型

注意力机制被引入计算机视觉，目的是模仿人类视觉系统的元素，该元素可以自然有效地发现复杂情况下的显著区域，它们在包括视觉连接识别的许多视觉任务中取得了巨大的成功。由 Bohan Zhuang 等人(2017) [40]创建的上下文感知注意力模型被加于特征图上，以选择特定于交互的显著特征区域。基于这些模型开发了视觉连接识别的交互识别框架。在这些数据集中，“谓词”与本文文献中的“交互作用”相当。上下文使用产生词向量的模型存储在语义空间中，这将有助于零次泛化。Ranjay Krishn 等人(2018) [41]构建了注意力和谓词转移模块，对连接实体的谓词建模，在不同实体上进行注意力转移。注意力模块通过软注意力 $\mathrm{Att}(\cdot)$ 进行近似最大化，可以表述为

$$\hat{x}^0=\mathrm{Att}(\mathbf{f},S)=\mathrm{ReLU}\left(\mathbf{f}\cdot\mathrm{Emb}(S)\right) \tag{18}$$

$$\hat{y}^0 = \mathrm{Att}(\mathbf{f}, O) = \mathrm{ReLU}\left(\mathbf{f} \cdot \mathrm{Emb}(O)\right) \tag{19}$$

其中，$\mathrm{Emb}(\cdot)$将实体嵌入到 C-维语义空间中，$\mathbf{f}$ 表示从图像中提取的特征映射，$\mathrm{ReLU}(\cdot)$是经过校正的线性单位运算符，$\hat{x}^0$，$\hat{y}^0$ 分别表示只使用实体对主语和宾语的初始注意。

为了理解参考表达式，Wang Peng 等人(2019) [42]引入了一种基于图形的、语言引导的注意力方法。此技术的节点和边缘注意组件分别针对对象区域和关系而设计。然而，由于上述方法忽略了三元组级连接，Li Mi 和 Zhenzhong Chen (2020) [43]开发了一个分层图注意网络，以更好地捕获三元组关系。

注意力机制基于关注基本部分而不是整个图像。软注意力已经以上述基于注意力的方式得到解决，但其他注意力模型还需进一步发展。

4.4.4. 目标成本和损失

由于单个类别中的谓词样本明显不同，因此已利用具有目标成本和损失函数的多个线索来减少关系的模糊性。对于短语定位和关系检测，已经使用了语言和视觉线索的集合，以及专门的成本函数[10]。多元计算，如语义嵌入线索、空间位置和视觉外观，已被整合为深度结构排名(DSR)框架的输入，并且设计了一个具有结构排名损失的排名目标函数，通过强制注释的关系来学习一对本地化对象的相关谓词，以获得更高的相关性分数。损失函数定义为

$$\mathcal{L}(x) = \sum_{r \in R} \sum_{r' \in R'} \left[\Delta(r, r') + \Phi(x, r') - \Phi(x, r)\right]_+ \tag{20}$$

其中 x 是输入图像，$r = (s, p, o)$ 是关系实例，$\mathcal{R} = (s, p, o) \mid (s, o) \in \mathcal{P} \wedge p / \notin \mathcal{P}_{s,o}$ 表示一个图像中存在的可视关系，$\mathcal{R}' = (s', p', o') \mid (s', p', o') \mid (s', o') \in \mathcal{P} \wedge p' / \notin \mathcal{P}_{s',o'}$ 并且是未注释的关系实例。$[\cdot]_+ = \max(0, \cdot)$ 工作以保留正极部分。$\Delta(\cdot, \cdot)$是一个边距函数，用于测量视觉关系检测的不完整性，定义为

$$\Delta(r, r') = \Delta(s, p, o, s', p', o') = 1 + P(p \mid c_s, c_o) - P(p' \mid c_{s'}, c_{o'}) \tag{21}$$

其中 c_s 和 c_o 表示主语和宾语的类别。Φ 是一个用于测量 x 和 r 之间兼容性的函数：

$$\Phi(x, r) = \Phi(x, s, p, o) = \mathrm{W}_p^{\mathrm{T}} f(x, s, o) \tag{22}$$

其中 w_p 表示要学习的第 p 个谓词的参数。该框架旨在使关系共生更容易，并解决注释不完整的问题。

Zhu Yaohui 等人(2017) [44]提出了一种基于空间分布表达物体位置关系(PR)并在物体之间传递结构信息的概念的方法。空间分布包括 PR、距离关系(DR)、形状关系(SIR)和大小关系(SHR)。使用图形对比损失解决了两种类型的错误：1) 近端连接歧义 2) 实体实例混淆。对于近端连接歧义，还创建了类敏感谓词损失。

损失函数的设计对学习模态的训练技术有巨大的影响，并且根据成本和损失方面，模态将用于各种视觉交互。

4.4.5. 基于强化学习(RL)的框架

深度强化学习的目标是学习最佳或接近最优的策略，以最大化从即时奖励中获得的“奖励功能”。为了捕捉关系和特征之间的全局语义相互依赖性，Liang Xiaodan 等人(2017) [11]提出了一个深度变异结构 RL(VRL)框架。有向语义动作网络、变异结构化遍历方案、状态空间和奖励函数都是 VRL 的重要组成部分。识别视觉联系和质量的挑战表现为 VRL 框架中的顺序决策过程。它使用全局上下文信号和先前提取的单词的语义嵌入来生成预测。

MnihVolodymyr 等人(2013) [45]提出了一个深度 Q 网络(DQN)，它用于分别估计属性 $\mathcal{A}$ 谓词类别 $\mathcal{P}$ 和对象类别 c 的网络权重 θ_a, θ_p 和 θ_c。奖励函数包括 $\mathcal{R}_a(\mathbf{f}, \mathbf{g}_a)$，$\mathcal{R}_p(\mathbf{f}, \mathbf{g}_p)$ 和 $\mathcal{R}_c(\mathbf{f}, \mathbf{g}_c)$，这些函数反映了

在状态 $\mathbf{f}$ 下采取行动 $\left(\mathbf{g}_a,\mathbf{g}_p,\mathbf{g}_c\right)$ 的检测精度。在训练阶段的变异结构化行动中，使用贪婪策略选择行动 $\mathbf{g}_a,\mathbf{g}_p$ 和 $\mathbf{g}_c$，然后在测试阶段选择估计 Q 值最高的最佳行动来发现对象，关系和属性。基于 RL 的方法将在很大范围内搜索最佳结果。然而，由于前向搜索的多重计算，其复杂性是显著的。RL 的基本组成部分(规划、价值功能、奖励和政策等)需要进一步研究。

5. 视觉关系检测的应用

在处理了 VRD 的任务之后，研究人员试图理解在生成过程中为提高性能而做的所有工作，我们现在将探索其不同的应用。

5.1. 场景图

场景图被提出以更清晰和有组织的方式表达图像属性和对象连接，通过显式建模对象，它们的属性以及它们与其他对象的关系来捕获视觉场景的综合语义[46]。在图像中节点表示场景中的项目，连接它们的边表示动态图形框架中各种元素之间的交互。在图 6 中可以看到例子，其中对象，属性和关系以图形方式表示。

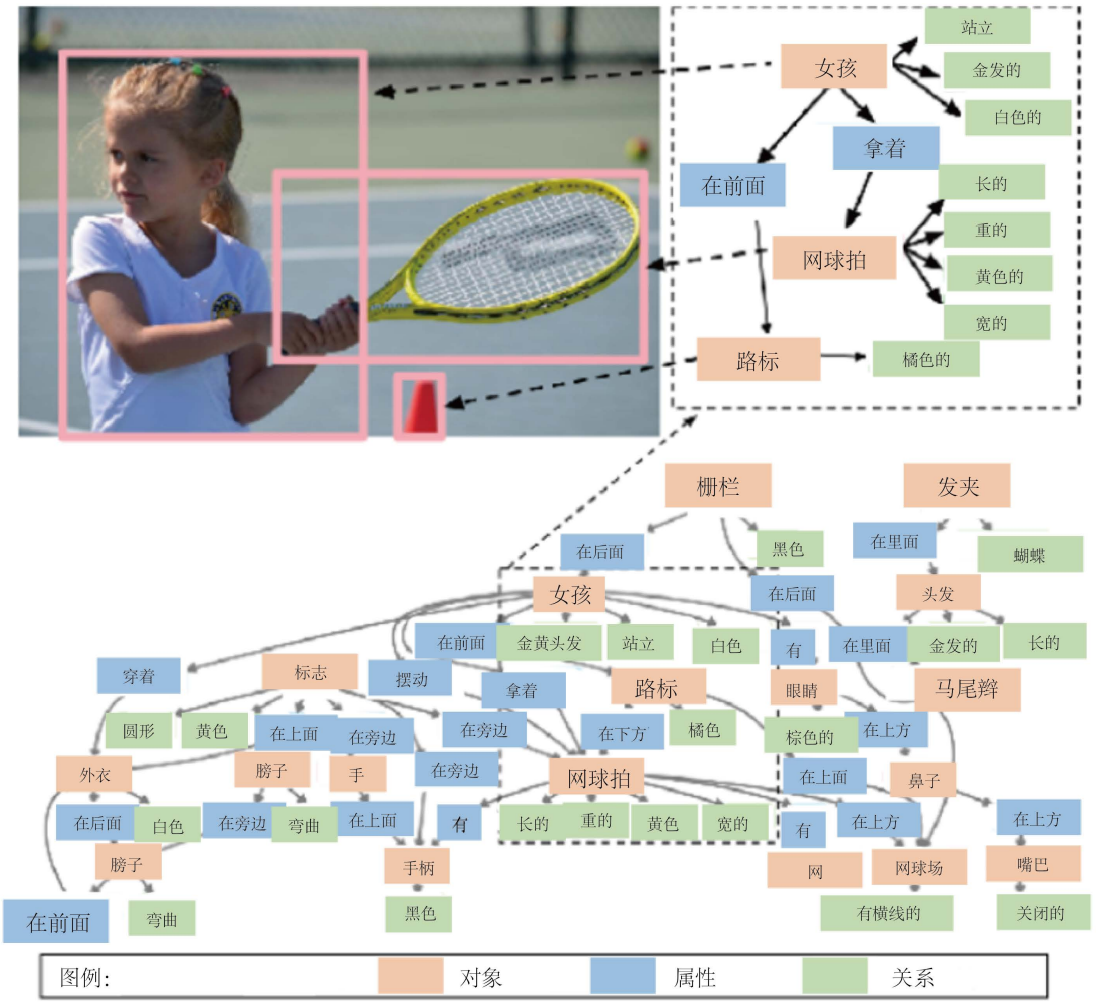

Figure 6. An example of a scene graph [46]
图 6. 场景图生成示例[46]

场景图是一种描述场景内容的图形数据结构，它对对象实例、对象属性以及对象之间的关系进行编码。给定一组对象类 $\mathcal{C}$、一组属性类型 $\mathcal{A}$ 和一组关系类型 $\mathcal{R}$，我们将场景图 G 定义为元组 $G=(O,E)$，其中 $O=\left(o_q,\cdots,o_n\right)$ 是一组对象，$E\subseteq O\times\mathcal{R}\times O$ 是一组边。每个对象的形式为 $o_i=\left(c_i,A_i\right)$，其中 $c_i\in\mathcal{C}$ 是对象的类，$A_i\subseteq\mathcal{A}$ 是对象的属性。

Justin Johnson 等人(2015) [46]引入了一种新的条件随机场(CRF)模型，用于从给定的场景图中检索图像，优于基于低级视觉特征的检索方法。他们定义了一组位置框 B 来表示一个图像，一个映射 $\gamma:O\to B$ 表示场景图的基础。然后，目标可以表述为

$$\gamma^{*}=\arg\max_{\gamma}\prod_{o\in O}P\left(o\mid\gamma_0\right)\prod_{(o,r,0')\in E}P\left(\gamma_0,\gamma_{0'}\mid o,\gamma,o'\right) \tag{23}$$

其中 $P\left(o\mid\gamma_0\right)$用于测量 γ_0 和 o 之间的一致性，而 $P\left(o\mid\gamma_0\right)P\left(\gamma_0,\gamma_{0'}\mid o,\gamma,o'\right)$用于模拟位置框对 $\left(\gamma_0,\gamma_{0'}\right)$ 表示元组 $\left(o,\gamma,o'\right)$ 的程度。可以用下式提取特征以编码相对位置和比例：

$$f\left(\gamma_0,\gamma_{0'}\right)=\left(\left(x-x'\right)/w,\left(y-y'\right)/h,w'/w,h'/h\right) \tag{24}$$

其中 $\gamma_0=\left(x,y,w,h\right)$ 和 $\gamma_{0'}=\left(x',y',w',h'\right)$ 表示位置框的坐标。

场景图已被用作语义图像检索查询，以模拟对象对之间的多种交互模式。CRF 模型旨在对场景图所有可能接地的分布进行建模。文献[46]中的场景图数据集由 5000 个以图像为基点的样本组成，用于测试图像检索性能。实验表明，场景图方法比仅使用对象或低级特征进行图像检索的方法表现更好。在计算机图形学中，图形已被用于表示结构关系，这些关系是在场景中学习的[47]。这是为了解决三维场景语料库中的数据挖掘问题，因为表示虚拟环境的场景的模型密度和复杂度正在迅速上升。

为了比较两个图中的虚拟子结构，定义了一个这些关系图之间的核，应用于场景建模问题，例如查找相似场景，相关性反馈和基于上下文的模型搜索。在文献[48]提出了常识空间知识的表示，并将其应用于文本到三维场景生成的任务。在文献[10]中建立一个有向语义动作图将所有可能的对象名词、属性和关系组织成一个紧凑且语义上有意义的表示。基于高效环境模型将在智能代理的自治系统中起关键作用的考虑，在文献[49]中提出了一个三维场景图作为环境模型的构建框架。除了与二维图形类似的表示形式外，三维场景图还提取环境中的物理属性，包括三维位置。在厨房模拟环境中的 VQA 和任务规划中，三维场景图的适用性得到了验证。

5.2. 人与物体交互(HOI)

理解人与物体之间的交互是视觉分类的基本问题之一，也是详细场景理解的重要一步。人与物体交互(HOI)检测致力于定位人和对象，以及识别它们之间的复杂交互。HOI 是一种特殊的视觉关系，其中谓词始终表示动作。Georgia Gkioxari 等人(2018) [50]提出一种以人为中心的方法，将人的外表(即姿势，衣服和动作)作为定位与人互动的对象的强大线索。引入了一个基于快速区域的卷积神经网络(Fast-RCNN)的框架来检测。该框架包括三个分支：1) 对象检测；2) 以人为中心；3) 互动。为了应对 HOI 类别长尾分布的挑战，基于训练数据集和外部来源的地面事实注释。为了检索描述检测到的〈人，对象〉对的动词，学习语义结构感知嵌入空间以利用语义相似性。该方法已在 V-COCO 和 HICO-DET 数据集上进行了测试。在图像中，人始终是中心角色，因此 HOI 在许多应用中都很重要，例如人机交互和机器人技术。

5.3. 视觉问答(VQA)

VQA 包括自然语言处理和计算机视觉，这被称为多模态任务。通过在 VQA 中输入模型的图像，基

于该图像的相关问题的正确答案将是输出。图 7 是视觉问答的示例。

Figure 7. An example of VQA (from: visualqa.org)
图 7. 视觉问答的示例(来源：visualqa.org)

VQA 是在图像内容中进行搜索和推理，图 8 显示了 VQA 的过程。

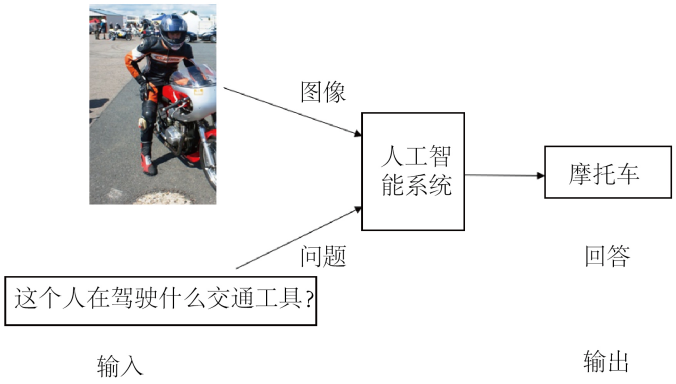

Figure 8. Process of VQA
图 8. 视觉问答过程

如图 8 所示，系统首先要检测到图像中的物体，对物体和场景进行分类、识别，对于问题还必须检测出物体之间的关系，进行常识性推理和可能性的知识推理。因此，VQA 也将在图像理解方面受益于 VRD。

VQA 始终被视为一个分类问题，表述为

$$\hat{a} = \arg\max_{a\in\Omega} p(a \mid Q, I; \Theta) \tag{25}$$

其中 $\hat{a}$ 代表最可能的答案，Q 表示一个问题，是表示 I 一个模型参数的相关图像，Θ 是候选答案的集合。

Zhou Su 等人(2018) [51]，它将结构化的人类知识与〈主体，关系，目标〉和深度视觉特征的结构三元组结合到记忆网络中。他们的可视化知识库建立在 VQA 数据集[35]提出了一种视觉知识记忆网络(VKMN)和 VG 关系数据集。Remi Cadene 等人(2019) [52]提出了一个用于 VQA 的多模态关系网络(MRA-Net)，并且已经利用区域关系来推理彼此交互的多个对象，将学习空间和语义概念。Liang Peng 等

人(2022) [53]在 MRA-Net 中设计了两个自适应视觉关系(二元和三元)注意模块，可以提供大量的更深层次的语义来增强推理能力。为了克服流行基准的缺陷(例如答案分布中的现实世界先验和统计偏差)，最近出现了一个新的数据集 GQA，用于视觉推理和组合问答[54]。由于医学照片的分辨率通常较差，并且经常复杂地解释医学图像，因此医学图像上的 VQA 很困难。

VQA 的主要目标是根据输入图像和问题，计算机输出符合自然语言规则的合理答案，因此，用户需要对图像的内容，问题的含义和意图以及相关的常识有一定的了解。短期记忆[55]和不同的推理模式[56]将有助于 VQA 的研究。

6. 总结

深度学习自提出以来，一直是图域中计算机学习问题的关键和实用方法，尤其近年来深度学习得到了广泛应用和快速增长，促进了大数据向大模型的转变，但在计算机视觉领域中仍然存在一些不适合深度学习的难题，如 3D 建模、视频处理和场景理解等，需要以后进一步深入研究。在深度学习和视觉关系检测的融合进程中，先后开发出许多优于传统的方法和模型，需要在视觉关系检测的应用场景中验证和完善，寻找出性能优越、简易快捷视觉关系检测新方法和新技术；不断增加视觉关系检测的数据规模来提升深度学习的性能；扩展和优化全视野、3D 视觉领域等视觉关系检测的深度学习模型。

致　谢

本文得到国家自然科学基金(No. 51979085)资助。

参考文献

[1] Wang, Q., Zou, L., Yao, Y., Wang, Y., Li, J. and Yang, W. (2021) An Interconnected Feature Pyramid Networks for Object Detection. *Journal of Visual Communication and Image Representation*, **3**, Article ID: 103260. https://doi.org/10.1016/j.jvcir.2021.103260

[2] Zhang, L., Hu, X., Zhou, Y., Zhou, G. and Duan, S. (2021) Memristive DeepLab: A Hardware Friendly Deep CNN for Semantic Segmentation. *Neurocomputing*, **451**, 181-191. https://doi.org/10.1016/j.neucom.2021.04.061

[3] Zhu, Y., Li, L. and Wu, X. (2021) Stacked Convolutional Sparse Auto-Encoders for Representation Learning. *ACM Transactions on Knowledge Discovery from Data*, **15**, Article No. 31. https://doi.org/10.1145/3434767

[4] Lu, C., Krishna, R., Bernstein, M. and Li, F.-F.(2016) Visual Relationship Detection with Language Priors. *European Conference on Computer Vision* (*ECCV*) 2016, Amsterdam, 11-14 October 2016, 852-869. https://doi.org/10.1007/978-3-319-46448-0_51

[5] Liu, P., Xiang, C., Jia, D., Zhao, X., Meng, W. and Wang, J. (2020) Stacked Attention Recurrent Relational Networks for Question Answering. *Journal of Physics Conference Series*, **1570**, Article ID: 012072. https://doi.org/10.1088/1742-6596/1570/1/012072

[6] Zhang, H., Kyaw, Z., Chang, S.F. and Chua, T.-S. (2017) Visual Translation Embedding Network for Visual Relation Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, 21-26 July 2017, 3107-3115. https://doi.org/10.1109/CVPR.2017.331

[7] Desai, C., Ramanan, D. and Fowlkes, C. (2009) Discriminative Models for Multi-Class Object Layout. 2009 *IEEE* 12*th International Conference on Computer Vision*, Kyoto, 29 September-2 October 2009, 229-236. https://doi.org/10.1109/ICCV.2009.5459256

[8] Yao, B. and Li, F.F. (2010) Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 17-24. https://doi.org/10.1109/CVPR.2010.5540235

[9] Mensink, T., Gavves, E. and Snoek, C.G.M. (2014) COSTA: Co-Occurrence Statistics for Zero-Shot Classification. *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 2441-2448. https://doi.org/10.1109/CVPR.2014.313

[10] Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J. and Lazebnik, S. (2017) Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. 2017 *IEEE International Conference on Computer Vision* (ICCV), Venice, 22-29 October 2017, 1946-1955. https://doi.org/10.1109/ICCV.2017.213

[11] Liang, X., Lee, L. and Xing, E.P. (2017) Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, 21-26 July 2017, 4408-4417. https://doi.org/10.1109/CVPR.2017.469

[12] Li, Y., Ouyang, W., Wang, X. and Tang, X. (2017) ViP-CNN: Visual Phrase Guided Convolutional Neural Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, 21-26 July 2017, 7244-7253. https://doi.org/10.1109/CVPR.2017.766

[13] Dai, B., Zhang, Y. and Lin, D. (2017) Detecting Visual Relationships with Deep Relational Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, 21-26 July 2017, 3298-3308. https://doi.org/10.1109/CVPR.2017.352

[14] Hu, Z., Yang, Z., Salakhutdinov, R. and Xing, E. (2016) Deep Neural Networks with Massive Learned Knowledge. 2016 *Conference on Empirical Methods in Natural Language*, Austin, 1-4 November 2016, 1670-1679. https://doi.org/10.18653/v1/D16-1173

[15] Dechter, R. (1986) Learning While Searching in Constraint-Satisfaction-Problems. *Proceedings of the* 5*th AAAI National Conference on Artificial Intelligence*, Philadelphia, 11-15 August 1986, 178-183.

[16] Aizenberg, I.N., Aizenberg, N.N. and Vandewalle, J. (2000) Multi-Valued and Universal Binary Neurons: Theory, Learning and Applications. Springer, New York. https://doi.org/10.1007/978-1-4757-3115-6

[17] Huang, M.L. and Wu, Y.Z. (2022) Semantic Segmentation of Pancreatic Medical Images by Using Convolutional Neural Network. *Biomedical Signal Processing and Control*, **73**, Article ID: 103458. https://doi.org/10.1016/j.bspc.2021.103458

[18] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., *et al*. (1989) Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, **1**, 541-551. https://doi.org/10.1162/neco.1989.1.4.541

[19] Cullheim, S., Kellerth, J.O. and Conradi, S. (1977) Evidence for Direct Synaptic Interconnections between Cat Spinal α-Motoneurons via the Recurrent Axon Collaterals: A Morphological Study Using Intracellular Injection of Horseradish Peroxidase. *Brain Research*, **132**, 1-10. https://doi.org/10.1016/0006-8993(77)90702-8

[20] Hopfield, J.J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 2554-255. https://doi.org/10.1073/pnas.79.8.2554

[21] Jordan, M.I. (1997) Serial Order: A Parallel Distributed Processing Approach. In: Donahoe, J.W. and Dorsel, V.P., Eds., *Neural-Network Models of Cognition*: *Biobehavioral Foundations*, Vol. 121, North-Holland, Amsterdam, 471-495. https://doi.org/10.1016/S0166-4115(97)80111-2

[22] Elman, J.L. (1990) Finding Structure in Time. *Cognitive Science*, **14**, 179-211. https://doi.org/10.1207/s15516709cog1402_1

[23] Schmidhuber, J. (1992) Learning Complex, Extended Sequences Using the Principle of History Compression. *Neural Computation*, **4**, 234-242. https://doi.org/10.1162/neco.1992.4.2.234

[24] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[25] Schuster, M. and Paliwa, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. https://doi.org/10.1109/78.650093

[26] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., *et al*. (2014) Generative Adversarial Nets. arXiv preprint arXiv:1406.2661

[27] Tieleman, T. (2008) Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient. *Proceedings of the* 25*th International Conference on Machine Learning*, Helsinki, 5-9 July 2008, 1064-1071. https://doi.org/10.1145/1390156.1390290

[28] Sperduti, A. and Starita, A. (1997) Supervised Neural Networks for the Classification of Structures. *IEEE Transactions on Neural Networks*, **8**, 714-735. https://doi.org/10.1109/72.572108

[29] Ruiz, L., Gama, F. and Ribeiro, A. (2020) Gated Graph Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **68**, 6303-6318. https://doi.org/10.1109/TSP.2020.3033962

[30] Bruna, J., Zaremba, W., Szlam, A. and LeCun, Y. (2013) Spectral Networks and Locally Connected Networks on Graphs. arXiv preprint arXiv:1312.6203.

[31] Micheli, A. (2009) Neural Network for Graphs: A Contextual Constructive Approach. *IEEE Transactions on Neural Networks*, **20**, 498-511. https://doi.org/10.1109/TNN.2008.2010350

[32] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. https://doi.org/10.1109/TNNLS.2020.2978386

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2016) Attention Is All You Need. arXiv preprint arXiv:1706.03762

[34] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[35] Sadeghi, M.A. and Farhadi, A. (2011) Recognition Using Visual Phrases. *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Colorado Springs, 20-25 June 2011, 1745-1752. https://doi.org/10.1109/CVPR.2011.5995711

[36] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2015) The PASCAL Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, **111**, 98-136. https://doi.org/10.1007/s11263-014-0733-5

[37] Yu, R., Li, A., Morariu, V.I. and Davis, L.S. (2017) Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. 2017 *IEEE International Conference on Computer Vision* (CVPR), Venice, 22-29 October 2017, 1068-1076. https://doi.org/10.1109/ICCV.2017.121

[38] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[39] Plesse, F., Ginsca, A., Delezoide, B. and Prêteux, F. (2018) Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. 2018 *IEEE International Conference on Multimedia and Expo* (ICME), San Diego, 23-27 July 2018, 1-6. https://doi.org/10.1109/ICME.2018.8486503

[40] Zhuang, B., Liu, L., Shen, C. and Reid, I. (2017) Towards Context-Aware Interaction Recognition for Visual Relationship Detection. 2017 *IEEE International Conference on Computer Vision* (ICCV), Venice, 22-29 October 2017, 589-598. https://doi.org/10.1109/ICCV.2017.71

[41] Krishna, R., Chami, I., Bernstein, M. and Li, F.-F. (2018) Referring Relationship. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6867-6876. https://doi.org/10.1109/CVPR.2018.00718

[42] Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L. and van den Hengel, A. (2019) Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, 15-20 June 2019, 1960-1968. https://doi.org/10.1109/CVPR.2019.00206

[43] Mi, L. and Chen, Z. (2020) Hierarchical Graph Attention Network for Visual Relationship Detection. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Seattle, 13-19 June 2020, 13883-13892. https://doi.org/10.1109/CVPR42600.2020.01390

[44] Zhu, Y., Jiang, S. and Li, X. (2017) Visual Relationship Detection with Object Spatial Distribution. 2017 *IEEE International Conference on Multimedia and Expo* (ICME), Hong Kong, China, 10-14 July 2017, 379-384. https://doi.org/10.1109/ICME.2017.8019448

[45] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., *et al*. (2013) Playing Atari with Deep Reinforcement Learning. arXiv preprint arXiv.1312.5602.

[46] Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D.A., Bernstein, M.S., *et al*. (2015) Image Retrieval Using Scene Graphs. *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3668-3678. https://doi.org/10.1109/CVPR.2015.7298990

[47] Fisher, M., Savva, M. and Hanrahan, P. (2011) Characterizing Structural Relationships in Scenes Using Graph Kernels. *ACM Transactions on Graphics*, **30**, Article No. 34. https://doi.org/10.1145/2010324.1964929

[48] Chang, A.X., Savva, M. and Manning, C.D. (2014) Learning Spatial Knowledge for Text to 3D Scene Generation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (EMNLP), Doha, 25-29 October 2014, 2028-2038. https://doi.org/10.3115/v1/D14-1217

[49] Kim, U.-H., Park, J.-M., Song, T,-J. and Kim, J.-H. (2020) 3-D Scene Graph: A Sparse and Semantic Representation of Physical Environments for Intelligent Agents. *IEEE Transactions on Cybernetics*, **50**, 4921-4933. https://doi.org/10.1109/TCYB.2019.2931042

[50] Gkioxari, G., Girshick, R., Dollár, P. and He, K. (2018) Detecting and Recognizing Human-Object Interactions. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Salt Lake City, 18-23 June 2018, 8359-8367. https://doi.org/10.1109/CVPR.2018.00872

[51] Su, Z., Zhu. C., Dong, Y., Cai, D., Chen, Y. and Li, J. (2018) Learning Visual Knowledge Memory Networks for Visual Question Answering. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Salt Lake City, 18-23 June 2018, 7736-7745. https://doi.org/10.1109/CVPR.2018.00807

[52] Cadene, R., Ben-Younnes, H., Cord, M. and Thome, N. (2019) MUREL: Multimodal Relational Reasoning for Visual

Question Answering. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, 15-20 June 2019, 1989-1998. https://doi.org/10.1109/CVPR.2019.00209

[53] Peng, L., Yang, Y., Wang, Z., Huang Z. and Shen, H.T. (2022) MRA-Net: Improving VQA via Multi-Modal Relation Attention Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 318-329. https://doi.org/10.1109/TPAMI.2020.3004830

[54] Hudson, D.A. and Manning, C.D. (2019) GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, 15-20 June 2019, 6700-6709. https://doi.org/10.1109/CVPR.2019.00686

[55] Gupta, R., Hooda, P., Sanjeev and Kumar Chikkara, N. (2020) Natural Language Processing Based Visual Question Answering Efficient: an Efficient Det Approach. 4*th International Conference on Intelligent Computing and Control Systems* (*ICICCS*), Madurai, 13-15 May 2020, 900-904. https://doi.org/10.1109/ICICCS48265.2020.9121068

[56] Andreas, J., Rohrbach, M., Darrell, T. and Klein, D. (2016) Neural Module Networks. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Las Vegas, 27-30 June 2016, 39-48. https://doi.org/10.1109/CVPR.2016.12