

分类号: TP391.41

单位代码: 10335

密 级: 无

学 号: 11521059

浙江大学

博士学位论文



中文论文题目: 面向复杂场景理解的视觉内容识别、
检测与推理方法研究

英文论文题目: Visual Recognition, Detection, and Reasoning
for Complex Visual Scene Understanding

申请人姓名: 陈隆

指导教师: 肖俊

合作导师:

专业名称: 计算机科学与技术

研究方向: 计算机视觉

所在学院: 计算机科学与技术学院

论文提交日期 2020 年 xx 月 xx 日

面向复杂场景理解的视觉内容识别、
检测与推理方法研究



论文作者签名: _____

指导教师签名: _____

论文评阅人 1: _____

评阅人 2: _____

评阅人 3: _____

评阅人 4: _____

评阅人 5: _____

答辩委员会主席: xx 教授 xx 大学

委员 1: xx 教授 xx 大学

委员 2: xx 教授 xx 大学

委员 3: xx 教授 xx 大学

委员 4: xx 教授 xx 大学

委员 5: _____

答辩日期: 2020 年 xx 月 xx 日

Visual Recognition, Detection, and Reasoning
for Complex Visual Scene Understanding



Author's Signature: _____

Supervisor's Signature: _____

Thesis reviewer 1: _____

Thesis reviewer 2: _____

Thesis reviewer 3: _____

Thesis reviewer 4: _____

Thesis reviewer 5: _____

Committee of oral defence:

Committee Chairman: _____

Committeeman 1: _____

Committeeman 2: _____

Committeeman 3: _____

Committeeman 4: _____

Committeeman 5: _____

Date of oral defence: xx June 2020

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 浙江大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解 浙江大学 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

(保密的学位论文在解密后适用本授权书)

学位论文作者签名： 导师签名：
签字日期： 年 月 日 签字日期： 年 月 日

摘要

随着近年来众多大规模人工标注的图像和视频数据集的出现，基于深度学习的计算机视觉技术取得了长足的进步。然而，对于复杂视觉场景的识别和理解，目前的计算机视觉模型的表现与人类的表现还相差甚远，远远没有达到落地应用和大规模普及的水平。但是，日常生活中的媒体数据通常都是复杂视觉场景。为了充分利用日常生活中的海量媒体数据，复杂视觉场景的感知和理解已经逐渐成为近年来计算机视觉领域的一个研究热点。

本文将针对这四个不同层次的场景识别和理解（物体识别、场景识别、场景理解和场景推理），逐步地对复杂视觉场景的识别、检测和推理进行研究。本文的关键技术线路主要研究零样本物体分类、图像场景图生成、图像描述生成、视频片段检索和视觉问答等具体研究任务。在此研究路线下，本文主要的研究内容和贡献如下：

1. 针对目前零样本物体分类模型中普遍存在的属性丢失的问题，本文提出一种全新的零样本学习网络：基于属性保持的对抗网络。该网络通过引入两个独立的映射网络分支，将图像分类和图像重建两个原本相互冲突的任务分离出来。然后利用对抗网络学习让重建网络的特征向量的部分属性能够迁移到分类网络的特征向量中，从而使得分类网络的特征向量保持尽可能多的属性，减缓语义丢失的问题。

2. 针对流行的图像场景图生成的优化目标（物体和视觉关系分类的交叉熵之和）忽略不同物体重要性的问题，本文提出一种全新的训练框架，将图像场景图生成问题看成一个多智能体协同决策问题。基于新的框架，我们可以直接使用最终的场景图生成质量作为优化目标。同时，本文提出一种全新的反事实基准模型，来近似目标智能体预测类别的局部贡献。不仅可以显著地提升物体的类别准确率，同时提升整个场景图的生成质量。

3. 对于图像描述生成，基于现有的空间注意力机制，本文提出一种全新的多层次空间和通道注意力网络。通过充分挖掘卷积神经网络的特征图的三个维度之间的联系，使得模型在生成文本的过程中不断的关注不同的空间区域和通道。该网络

不仅是一种编码能力更强的注意力网络，同时帮助人们理解在图像描述生成过程中卷积神经网络的特征图的变化过程。

4. 对于视频片段检索，针对稀疏型自底向上模型的设计缺陷，本文提出一种全新的密集型自底向上的框架。通过将边界预测问题分解成相关性预测和边界回归两个子问题，大大降低了模型对视频动作边界定位的难度。同时，本文提出一个基于图卷积的特征金字塔层，来增强骨干网络编码能力。该框架大大提升了视频片段检索的准确率。

5. 针对目前视觉问答模型忽略的两个重要特性（视觉可解释性和问题敏感性），本文提出了一种通用的反事实样本生成机制。通过人为地遮盖图像中的重要区域或问题中的重要单词，同时更改标准答案，生成大量的反事实样本。然后合并原始的训练样本和新生成的反事实样本可以得到全新的训练样本。通过使用全新的训练样本对视觉问答模型进行训练，迫使模型关注反事实样本中被遮盖的重要内容，提升视觉问答模型的准确率和鲁棒性。

关键词：复杂场景理解，零样本物体分类，图像场景图生成，视觉描述生成，视频片段检索，视觉问答

Abstract

With the appearance of large scale human-annotated image and video datasets, deep-learning-based computer vision techniques have achieved impressive progress. However, for complex visual scene understanding, today's computer vision models still perform far behind human, and cannot be applied in daily life on a large scale. Unfortunately, complex visual scenes are ubiquitous in our daily visual data. To fully take advantage of the massive amounts of visual data, complex visual scene understanding is becoming one of the hottest research areas in computer vision.

In this thesis, we focus on visual recognition, detection, and reasoning for complex visual scene understanding. Specifically, we try to solve four different levels of visual scene understanding problems, including instance-level recognition, scene-level recognition, scene-level understanding, and scene-level reasoning. The related tasks consist of zero-shot recognition, scene graph generation, image captioning, query-based video localization, and visual question answering. In summary, this thesis makes five technical contributions:

1. For the ubiquitous semantic loss problem in current zero-shot visual recognition models, we propose a novel zero-shot learning framework: Semantics-Preserving Adversarial Embedding Networks (SP-AEN). SP-AEN introduces two independent embedding networks for classification and reconstruction respectively. Thanks to this design, we can disentangle these two conflict tasks. Then, SP-AEN resorts to adversarial learning to transfer attributes from reconstruction embeddings to classification embeddings, which helps to preserve semantics in classification embeddings and mitigate the semantic loss problem.

2. The prevailing training object for scene graph generation is the sum of cross-entropy loss of all objects and visual relationships, which overlooks the different contributions of different objects. In this thesis, we propose a novel training scheme: Counterfactual critic Multi-Agent Training (CMAT), which formulate scene graph generation

as a multi-agent cooperative decision-making problem. Based on this formulation, we can directly utilize the whole scene graph generation quality as the training objective. Meanwhile, we propose a counterfactual baseline, which can approximate the contribution of each local prediction. CMAT can not only improve the accuracy of object classification, but also improve the whole scene graph generation quality.

3. Based on the existing spatial attention mechanism in the encoder-decoder framework, we propose a novel network for image captioning: Spatial and Channel-wise Convolutional Networks (SCA-CNN). SCA-CNN take full advantage of the three characteristics of CNN features, and attend to different visual regions and channels during the sentence generation. The contribution of SCA-CNN is not only the more powerful attention model, but also a better understanding of where (ie, spatial) and what (ie, channel-wise) the attention looks like in a CNN that evolves during sentence generation.

4. According to the weakness of the sparse bottom-up framework for query-based video localization, we propose a novel dense bottom-up model. Specifically, we disentangle the boundary prediction problem into two sub-problems: relatedness prediction and boundary regression. Meanwhile, we propose a Graph Feature Pyramid Network (Graph-FPN) layer to boost the feature from the backbone network. The proposed model significantly improves the localization accuracy.

5. Current Visual Question Answering (VQA) models overlook two indispensable characteristics of an ideal VQA model: visual-explainable and question-sensitive. In this thesis, we propose a model-agnostic Counterfactual Samples Synthesizing (CSS) mechanism. The CSS generates numerous counterfactual training samples by masking critical objects in images or words in questions, and assigning different ground-truth answers. After training with the complementary samples (ie, the original and generated samples), the VQA models are forced to focus on all critical objects and words, which significantly improve the accuracy and robustness of VQA models.

Keywords: Complex Visual Scene Understanding, Zero-Shot Recognition, Scene Graph Generation, Image Captioning, Query-based Video Localization, Visual Question Answering

目 次

摘要	I
Abstract	III
目次	
插图	IX
表格	XI
1 绪论	1
1.1 研究背景	1
1.2 研究内容	4
1.2.1 基于属性保持对抗网络的零样本物体分类	5
1.2.2 基于反事实多智能体学习的图像场景图生成	5
1.2.3 基于多层空间和通道注意力网络的图像描述生成	6
1.2.4 基于密集型自底向上网络的视频片段检索	6
1.2.5 基于反事实样本生成的视觉问答	7
1.3 本文组织结构	7
1.4 本章小结	8
2 相关研究综述	9
2.1 零样本物体分类	9
2.1.1 零样本学习	9
2.1.2 通用型零样本学习	10
2.1.3 零样本学习中的域偏移问题	10
2.1.4 零样本学习和对抗生成网络	10
2.2 图像场景图生成	11
2.2.1 场景图生成	11
2.2.2 多智能体策略梯度	12
2.3 图像描述生成	12

2.3.1 编码-解码框架	12
2.3.2 注意力机制	13
2.4 视频片段检索	14
2.4.1 视频片段检索	14
2.4.2 自顶向下与自底向上	14
2.5 图像视觉问答	15
2.5.1 视觉问答模型	15
2.5.2 视觉问答模型的文本偏置	15
2.5.3 视觉问答模型的特性	16
3 基于属性保持对抗网络的零样本物体分类方法	17
3.1 问题描述	17
3.2 属性保持对抗网络	20
3.2.1 零样本分类预备知识	20
3.2.2 分类任务优化目标	20
3.2.3 重建任务优化目标	21
3.2.4 对抗学习优化目标	21
3.2.5 总体优化目标	22
3.3 实验设置与性能分析	22
3.3.1 零样本物体分类数据集	22
3.3.2 实验设定与零样本物体分类评价指标	23
3.3.3 网络模型与训练细节	24
3.3.4 零样本物体分类的性能对比	24
3.3.5 零样本物体分类方法分析	25
3.4 本章小结	29
4 基于反事实多智能体学习的图像场景图生成方法	31
4.1 问题描述	31
4.2 反事实多智能体学习	35
4.2.1 多智能体协同决策	35
4.2.2 反事实多智能体学习	37
4.3 实验设置与性能对比	41
4.3.1 图像场景图生成数据集与实验设定	41

4.3.2 实验细节	41
4.3.3 场景图生成性能分析	42
4.3.4 场景图生成性能对比	44
4.4 本章小结	46
5 基于多层空间和通道注意力网络的图像描述生成方法	49
5.1 问题描述	49
5.2 空间和通道注意力机制	51
5.2.1 概述	51
5.2.2 空间注意力机制	52
5.2.3 通道注意力机制	53
5.3 实验设置与性能对比	54
5.3.1 图像描述生成任务的数据集和评价指标	54
5.3.2 实验细节设定	55
5.3.3 通道注意力机制的性能分析	56
5.3.4 多层注意力机制的性能分析	58
5.3.5 空间和通道注意力卷积神经网络的性能比较	59
5.3.6 空间注意力和通道注意力权重的可视化	61
5.4 本章小结	62
6 基于密集型自底向上网络的视频片段检索方法	63
6.1 问题描述	63
6.2 基于图特征金字塔的密集型预测	67
6.2.1 骨干网络	67
6.2.2 图特征金字塔层	69
6.2.3 密集型头网络	70
6.2.4 训练阶段和测试阶段	71
6.3 实验设置与性能对比	71
6.3.1 视频片段检索数据集和评价指标	71
6.3.2 实验设定	72
6.3.3 视频片段检索性能对比	73
6.3.4 视频片段检索性能分析	75
6.4 本章小结	78

7 基于反事实样本生成的视觉问答方法	79
7.1 问题描述	79
7.2 反事实样本生成	82
7.2.1 引言	82
7.2.2 反事实样本生成	83
7.3 实验设置与性能对比	86
7.3.1 CSS 对视觉问答的性能分析	87
7.3.2 视觉问题方法性能对比	88
7.3.3 CSS 对视觉可解释性的帮助	90
7.3.4 CSS 对问题敏感性的帮助	91
7.4 本章小结	93
8 总结和展望	95
8.1 本文工作总结	95
8.2 未来研究展望	96
参考文献	99
攻读博士学位期间主要研究成果	119
致谢	121

插 图

1-1	大数据时代下图像、视频等媒体数据呈现“爆炸式”增长	1
1-2	众多大规模人工标注的图像和视频数据集推动计算机视觉技术的发展	2
1-3	复杂场景识别、检测和推理的关键技术路线	3
1-4	复杂场景识别、检测和推理的关键技术研究方法	5
3-1	三种典型的零样本学习框架	18
3-2	模型 SAE 和模型 SP-AEN 的图像重建结果对比	19
3-3	模型 SP-AEN 的整体网络结构流程图	20
3-4	三种不同的图像重建网络框架	26
3-5	不同图像重建网络框架在四个零样本物体分类数据集中的重建结果	27
3-6	在四个零样本分类数据集中的已见-未见准确率曲线下区域面积	28
3-7	图像重建结果随优化目标权重 α 的影响	29
4-1	图像场景图生成任务示例	32
4-2	场景图生成中优化目标的整体一致性和局部敏感性	33
4-3	模型 CMAT 的总体流程图	35
4-4	单步智能体通信示意图	36
4-5	CMAT 中反事实基准模型	39
4-6	优化目标局部敏感性的重要性示例图	39
4-7	模型 CMAT 和模型 MOTIFS 在数据集 VG 上的场景图生成结果对比...	46
5-1	VGG19 网络中 conv5_4 层和 conv5_3 层的通道注意力机制示意图	50
5-2	空间和通道注意力卷积神经网络流程图	52
5-3	空间注意力和通道注意力权重的可视化结果	60
6-1	两种不同的视频片段检索任务	64
6-2	典型的稀疏型自底向上视频片段检索模型	65

6-3	一个基于语句的视频片段检索示例	66
6-4	典型的稀疏型自底向上视频片段检索模型	67
6-5	QANet 的模型结构	68
6-6	图特征金字塔层	69
6-7	密集型头网络	70
6-8	时域池化示意图	71
6-9	GDP 模型分别在 ActivityNet Captions 和 Activity-VRL 的检索结果	73
6-10	同一骨干网络不同特征优化层的性能对比	75
6-11	场景空间中的节点可视化	76
7-1	VQA 模型的视觉可解释性和问题敏感性	80
7-2	V-CSS 和 Q-CSS 示意图	81
7-3	一个 I^+ 、 I^- 、 Q^+ 和 Q^- 的示例图	86
7-4	V-CSS 和 Q-CSS 中不同超参数对模型性能的影响	87
7-5	可视化结果	92

表 格

3-1 不同零样本分类模型在通用的四个零样本分类数据集上的性能对比	25
3-2 不同重建网络下重建图像与输入图像之间的平均像素差平方	26
3-3 SP-AEN 在不同优化目标条件下的性能对比	28
4-1 不同全局奖励函数的选择对性能的影响	43
4-2 不同基准模型对性能的影响	43
4-3 不同多智能体通信步数对性能的影响	43
4-4 不同场景图生成方法在 VG 数据集上的性能对比.....	45
5-1 VGG-19 网络和 ResNet-152 网络中单层注意力机制的性能对比	57
5-2 空间注意力模型在 VGG-19 网络和 ResNet-152 网络下不同层的性能 对比	58
5-3 空间和通道注意力模型在 VGG-19 网络和 ResNet-152 网络下不同层 的性能对比	59
5-4 不同描述语句生成算法在数据集 Flickr8k、Flickr30k 和 MSCOCO 上 的性能对比	61
5-5 不同图像描述语句生成算法在数据集 MSCOCO 的在线服务器上的 性能对比	61
6-1 不同基于语句查询的视频片段检索方法的性能对比	74
6-2 不同基于视频查询的视频片段检索方法的性能对比	74
6-3 基于语句查询的视频片段检索任务中模型 A、B、C、D 的性能对比 ...	76
6-4 基于视频查询的视频片段检索任务中模型 A、B、C、D 的性能对比 ...	76
6-5 密集型头网络和稀疏型头网络的性能对比	77
7-1 CSS 机制对不同 VQA 模型的性能影响.....	88
7-2 不同视觉问答模型在 VQA-CP v2 和 VQA v2 上的性能对比.....	89

7-3 不同视觉问答模型在 VQA-CP v1 上的性能对比	90
7-4 VQA-CP v2 测试集的准确率	91
7-5 VQA-CP v2 测试集的 \mathcal{AI} 分数	91
7-6 VQA-CP-Rephrasing 测试集的 $CS(k)$ 和 VQA-CP v2 测试集的 CI 分数	91

1 绪论

1.1 研究背景

视觉是人类感知外界客观世界的主要信息来源，而计算机视觉技术旨在让机器能够像人类一样准确地感知和理解物理世界。计算机视觉技术的发展和进步，是众多无人机交互技术的基石，也是人类社会迈向真正人工智能时代至关重要的一步。目前，随着互联网技术、社交媒体技术的快速发展以及数字媒体设备、便携式设备的全面普及，图像、动态图、视频等多种形式的视觉媒体数据都呈现“爆炸式”增长。如图 1-1 所示¹，截止到 2016 年，视频分享网站 YouTube 每分钟上传约 500 小时视频数据，图像分享社区 Instagram 每分钟上传约 65972 张图像数据。面对日益增长的海量视觉媒体数据，利用计算机视觉技术对其进行感知和理解，从而实现对海量视觉媒体数据的快速利用，对便利人们的日常生活、推动人类社会的进步都有着十分重大的意义和应用价值。

与此同时，如图 1-2 所示，随着近年来众多大规模人工标注的图像和视频数据集的出现^[1-5]，基于深度学习^[6,7]的计算机视觉技术取得重大突破，在多个视觉任务中达到甚至超过人类的表现。例如，在大规模图像分类数据集 ImageNet 上，最新的计算机视觉模型^[8]在 Top-1 类别的分类准确率高达 88.4%、在 Top-5 类别的分类





图 1-2 众多大规模人工标注的图像和视频数据集推动计算机视觉技术的发展

准确率高达 98.7%，而人类的在 Top-5 类别的分类准确率只有 94.9%^[2]。然而，对于复杂视觉场景的识别和理解，目前的计算机视觉模型的表现与人类的表现还相差甚远，远远没有达到落地应用和大规模普及的水平。这主要原因来自于复杂视觉场景中通常包含大量的物体以及物体间的交互，同时物体间还存在各种遮挡、尺度不同等问题，这些都大大增加了视觉场景识别和理解的难度。但是，日常生活中的媒体数据通常都是复杂视觉场景。为了充分利用日常生活中的海量媒体数据，复杂视觉场景的感知和理解已逐渐成为近年来计算机视觉领域的一个研究热点。

对复杂视觉场景的识别和理解，具体来说，主要包含四个不同的层次：1) 对场景内单个物体进行识别（**物体识别**）；2) 对场景内所有物体以及物体间的视觉关系进行识别（**场景识别**）；3) 对整个视觉场景的内容进行理解（**场景理解**）；4) 对整个视觉场景在理解的基础上进行知识推理（**场景推理**）。本文将针对这四个不同层次的场景识别和理解，逐步地对复杂视觉场景的识别、检测和推理进行研究。如图 1-3 所示，本文的关键技术线路主要包括物体分类（零样本物体分类）、场景图生成、视觉描述生成、视觉检索和视觉问答等具体研究任务：

1. 物体识别：视觉场景感知和理解的首要步骤就是对场景内包含的物体进行个体层次的识别。作为计算机视觉领域中一个最基本的问题，个体层次的物体识别结果将直接影响后续对整个视觉场景进行场景层次的识别、理解和推理。根据物体

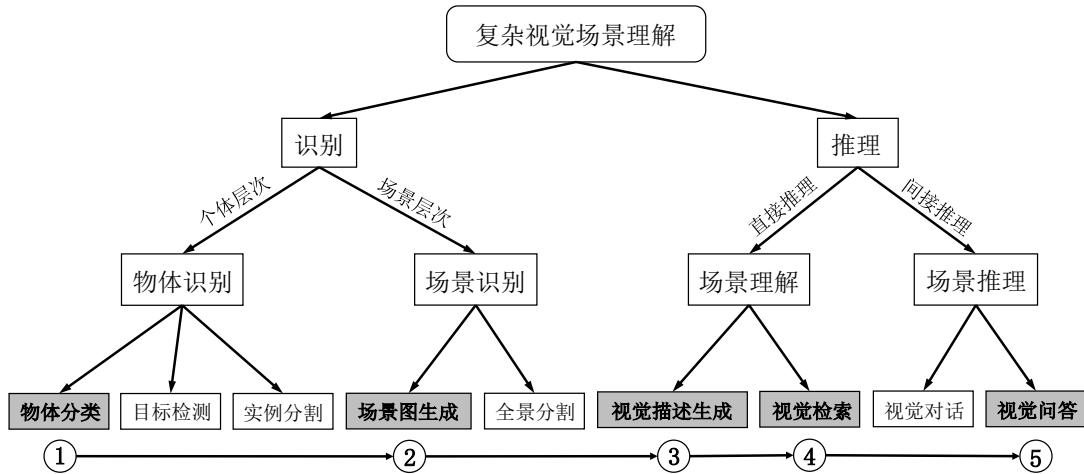


图 1-3 复杂场景识别、检测和推理的关键技术路线

识别的粒度，物体识别通常包含物体分类、目标检测、实例分割等多种具体任务。其中物体分类任务^[2,7,9–13]只是对物体进行多类别分类，而目标检测任务^[14–16]和实例分割任务^[17]需要在物体分类的基础上，同时对物体的大致边框位置或精确像素位置进行定位。随着卷积神经网络的发展，在理想实验条件下，当每个类别的训练样本足够充足时，物体识别的技术已经可以达到较高的准确率。然而，在日常生活中，所有类别的样本数量常常呈现“长尾分布”(long-tail distribution)，即大量的类别缺乏足够的训练样本。例如，在大规模图像分类数据集 ImageNet 中，除最常见的 1000 类图像以外，在剩余的 21814 个类别中，有 296 个类别只有一张训练样本图像^[2]。为了提升物体识别模型的通用性和鲁棒性，更加接近实际应用场景的少样本学习 (Few-Shot Learning, FSL)^[18] 或零样本学习 (Zero-Shot Learning, ZSL)^[19]逐渐成为近期的研究热点。然而，目前的零样本物体分类模型普遍存在语义丢失的问题 (semantic loss)，这将大大限制模型的迁移能力。本文主要聚集零样本物体分类任务，研究如何减少语义丢失，提升模型的通用性和鲁棒性。

2. 场景识别：视觉场景的组成元素除了大量的规则物体 (object) 外，还包括不规则物体 (stuff)，以及物体之间的视觉关系 (visual relationship) 等。因此，对整个视觉场景进行识别和理解的第二步就需要对所有的不规则物体 (全景分割^[20]) 和视觉关系 (场景图生成^[21]) 进行识别。尤其对于复杂场景来说，丰富的视觉关系通常“隐式地”包含场景内物体之间的内在联系。充分利用视觉关系不仅可以帮助提升单个物体的识别，同时通过与所有的物体结合，可以将非结构化的视觉场景转换成结构化的场景图 (scene graph)。这些场景图可以作为一种抽象的知识表达，辅助更高语义的场景理解和推理。然而，目前的场景图生成模型都集中于如何更加有

效地编码物体间的内在联系，使用物体和视觉关系分类的交叉熵之和作为优化目标。这个优化目标忽略了不同物体对整体场景图生成的不同贡献，大大降低了优化效率。本文主要聚集场景图生成任务，研究如何设计更加高效的场景图生成优化目标函数，提升场景图生成质量。

3. 场景理解：在对整个视觉场景中所有的组成元素都完成识别之后，就可以开始对场景内容进行理解和推理。就场景理解而言，如何判断一个计算机模型对视觉场景的理解程度，通常缺乏统一和标准的衡量方式和评价指标。随着自然语言处理领域的发展，众多视觉和文本融合的多模态任务开始当作视觉场景理解的代理任务：如视觉描述生成^[22]、视觉检索^[23]等。具体来说，视觉描述生成任务需要计算机模型生成一句自然描述语句，刚好能够正确描述整个视觉场景的内容。通过衡量最终描述语句的生成质量，来从侧面反映模型对视觉场景的理解程度。视觉检索任务需要计算机模型检索出与给定查询条件完全一致的视觉场景。通过衡量最终的检索排序结果，来从侧面反映模型对视觉场景的理解程度。本文将聚焦图像描述生成（image captioning）和视频片段检索（Query-based Video Localization, QBVL）两个具体的场景理解任务，通过分析目前模型框架的优缺点，研究如何设计更加合理的网络结构，提升模型对视觉场景的理解。

4. 场景推理：在对整个视觉场景内容进行充分理解之后，最后一步就是像人类一样对场景进行知识推理。视觉问答（Visual Question Answering, VQA）^[24]或视觉对话（visual dialog）^[25]等场景推理任务，通常被看成是一种视觉图灵测试^[26,27]。通过对视觉场景相关的内容进行提问，判断模型的场景推理能力。由于测试问题的自由性和开放性，理论上一个理想的计算机模型需要同时具备物体识别、场景识别、空间推理、常识推理等多方面的能力。然而，现阶段的视觉问答模型往往都过于依赖数据集内部的文本偏置（language bias），导致模型并没有充分理解场景内容。本文将聚焦到视觉问答任务，研究如何减少文本偏置对视觉问答模型的影响，帮助提升视觉问答模型的鲁棒性和推理能力。

1.2 研究内容

本文主要研究如何对复杂视觉场景进行不同层次的识别和理解，结合目前现有的研究技术，提出更加优化的学习算法和更加合理的网络结构设计，具体可以归纳为图 1-4 中所示方法。本文使用深度学习的方法对复杂场景理解中上述关键技术进行研究，具体包括以下内容：

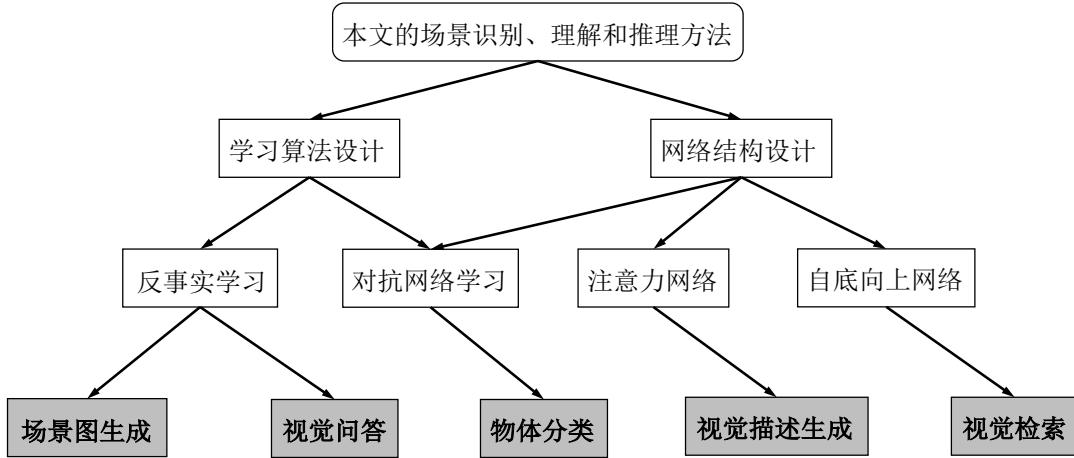


图 1-4 复杂场景识别、检测和推理的关键技术研究方法

1.2.1 基于属性保持对抗网络的零样本物体分类

复杂视觉场景的识别和理解，其首要步骤就是对视觉场景内的物体进行个体层次的识别。其中本文涉及到的关键技术为零样本物体分类，也称零样本学习。

目前主流的零样本物体分类模型都是基于嵌入映射的框架，这类方法主要是先在训练集上学习一个图像特征与类别属性特征之间的映射函数，然后直接将学习到的映射函数迁移到测试集中。由于训练集和测试集之间类别的差异，这些映射函数在测试集中不可避免地存在语义丢失的问题。针对这一常见问题，本文提出一种全新的零样本学习网络：基于属性保持的对抗网络。该网络通过引入两个独立的映射网络分支，将图像分类和图像重建两个原本相互冲突的任务分离出来。然后利用对抗网络学习让重建网络的特征向量的部分属性能够迁移到分类网络的特征向量中，从而使得分类网络的特征向量保持尽可能多的属性，减缓语义丢失的问题。本文提出的零样本学习网络不仅可以逼真地重建回原始图像，同时可以在多个数据集上大幅度提升零样本分类的准确率。

1.2.2 基于反事实多智能体学习的图像场景图生成

在复杂视觉场景的识别中，除了需要对场景内所有的单个物体进行识别，还需要检测物体间的视觉关系。图像场景图生成任务主要研究如何充分利用视觉场景内的各种元素，提升整个场景层次识别的准确率。

现有的图像场景图生成方法基本都是将场景内所有的物体和视觉关系分类的交叉熵之和作为模型的优化目标。这个优化目标的本质是认为每个物体分类的损失是相互独立的，即忽略了不同物体对整体场景图生成质量的不同贡献，容易陷入局部最优解。本文提出一种全新的训练框架，将图像场景图生成问题看成一个多智能

体协同决策问题，并且直接使用最终的场景图生成质量（评价指标 Recall@K）作为优化目标。具体来说，我们将图像中的每个物体看成一个智能体，每个智能体的动作空间是所有可能的物体类别。同时，本文提出一种全新的反事实基准模型，通过固定其他智能体预测的类别，而反事实地“更改”目标智能体的预测类别，来近似目标智能体预测类别的局部贡献。本文提出的反事实智能体学习框架可以显著地提升物体的类别准确率，以及整个场景图的生成质量。

1.2.3 基于多层空间和通道注意力网络的图像描述生成

图像描述生成是一种典型的视觉场景理解任务。该任务要求模型在对整个视觉场景进行充分的理解之后，生成自然语言来准确地描述场景内容。

现有的图像描述生成模型都是基于编码-解码（encoder-decoder）框架：即利用编码网络（如：卷积神经网络）将图像编码成视觉特征向量，然后利用解码网络（如：递归神经网络）将编码的视觉特征向量解码成自然语句。早期的图像描述生成模型都是只将图像编码成一个固定的视觉特征向量，这大大限制了视觉特征向量的表达能力。之后，部分模型开始引入空间注意力机制，通过在每个时刻对不同空间区域的视觉特征赋予不同的权重，得到不同的视觉特征向量。然而，卷积神经网络的特征图（feature map）除了空间维度外，还包含通道和层级两个维度。在不同的通道和层级下，视觉特征所编码的视觉信息也往往不同。本文提出一种全新的多层空间和通道注意力网络，不仅充分挖掘了特征图的三个维度之间的联系，提升了编码网络的表达能力。同时，还可以帮助人们理解在图像描述生成过程中卷积神经网络的特征图的变化过程。

1.2.4 基于密集型自底向上网络的视频片段检索

视频片段检索任务也是一种具有广泛应用价值的视觉场景理解任务。给定一个查询内容（query），视频片段检索任务需要模型在视频中定位出与查询内容相匹配的视频片段。

对于视频片段检索任务，目前的方法都属于自顶向下框架或稀疏型自底向上框架。本文首先分析了现有的视频片段检索的主流框架的优缺点，然后针对稀疏型自底向上模型的设计缺陷，我们提出了一种全新的密集型自底向上框架。我们通过将边界预测问题分解成相关性预测和边界回归两个子问题，大大降低了模型对视频动作边界定位的难度。同时，我们提出一个基于图卷积的特征金字塔层，来增强骨干网络编码能力。该框架大大提升了自底向上模型的检索准确率。对于多种不同

的查询形式（如：自然语句和视频片段），本文提出的密集型自底向上模型都达到了目前最好的性能。

1.2.5 基于反事实样本生成的视觉问答

视觉场景理解的最终目标就是能够在对整个场景内容完成理解之后进行知识推理。视觉问答是最简单的一种视觉推理任务，给定一个与视觉场景内容相关的问题，模型需要在简单的逻辑推理之后给出问题答案。

尽管近年来已经出现了大量的视觉问答模型，但是这些方法都忽略了一个理想视觉问答模型应当具备的两个重要能力：视觉可解释性和文本敏感性。为了提升视觉问答模型的这两个能力，本文提出了一种全新的反事实样本生成机制。通过人为地遮盖图像中的重要区域或问题中的重要单词，同时更改标准答案，生成大量的反事实样本。然后合并原始的训练样本和新生成的反事实样本可以得到全新的训练样本。通过使用全新的训练样本对视觉问答模型进行训练，迫使模型关注反事实样本中被遮盖的重要内容，即让模型在决策时关注到“正确的”视觉区域和单词，提升模型的准确率和鲁棒性。本文提出的反事实样本生成机制可以无缝地运用到任意的视觉问答模型中，提升模型的性能。

1.3 本文组织结构

本文通过对复杂视觉场景理解中的识别、检测和推理中的一系列典型问题，提出了多个更加优化的学习算法和更加合理的网络结构。全文共分为八章，后续章节安排如下：

- 第二章介绍了与本文相关的关键技术研究，就零样本物体分类、图像场景图生成、图像描述生成、视频片段检索和视觉问答等几方面的相关工作和本文的关系进行综述。
- 第三章介绍了基于属性保持对抗网络的零样本物体分类方法。本章首次提出图像分类与图像重建本质上是相互冲突的两个子任务。算法通过引入两个独立的映射网络分支，将图像分类和图像重建两个任务分离出来。然后利用对抗网络学习让重建网络的特征向量的部分属性能够迁移到分类网络的特征向量中，从而使得分类网络的特征向量保持尽可能多的属性，减缓语义丢失的问题。此项工作发表在国际顶级计算机视觉会议 CVPR 上。
- 第四章介绍了基于反事实多智能体学习的图像场景图生成方法。本章首次将

图像场景图生成问题看成一个多智能体协同决策问题，并且直接使用最终的场景图生成质量作为优化目标，避免现有优化目标的缺陷。同时，本章提出一种全新的反事实基准模型，有效地对不同物体的预测类别赋予不同的梯度，显著提升物体类别的预测性能和整体场景图的生成质量。此项工作发表在国际顶级计算机视觉会议 ICCV 上。

- 第五章介绍了基于多层空间和通道注意力网络的图像描述生成方法。本章首次提出通道注意力机制，认为通道注意力机制本质上也属于一种特殊的语义注意力机制。同时，本章提出一种全新的多层空间和通道注意力网络，不仅充分挖掘了卷积神经网络的特征图的三个维度（空间、通道和层级）之间的联系，提升了编码网络的表达能力。同时，还可以帮助人们理解在图像描述生成过程中卷积神经网络的特征图的变化过程。此项工作发表在国际顶级计算机视觉会议 CVPR 上。

- 第六章介绍了基于密集型自底向上网络的视频片段检索方法。本章首次提出了一种密集型自底向上网络框架。通过将边界预测问题分解成相关性预测和边界回归两个子问题，大大降低了模型对视频动作边界定位的难度。同时，本章提出一个基于图卷积的特征金字塔层，来增强骨干网络编码能力。该框架显著地提升了视频片段检索的准确率。此项工作发表在国际顶级人工智能会议 AAAI 上。

- 第七章介绍了基于反事实样本生成的视觉问答方法。本章首次提出了一种通用的反事实训练样本生成方法，提升视觉问答模型的视觉可解释性和文本敏感性。本文提出的反事实样本生成机制可以无缝地运用到任意的视觉问答模型中，提升模型的回答准确率。此项工作已经发表在国际顶级计算机视觉大会 CVPR 上。

- 第八章对全文介绍的工作进行了总结，并提出了对进一步对复杂场景理解的识别、检测和推理的研究内容以及今后的研究展望。

1.4 本章小结

本章对复杂视觉场景的识别、检测和推理问题进行了叙述，分别介绍了研究背景、本文的主要研究内容以及全文的组织结构。

2 相关研究综述

本章将就零样本物体分类、图像场景图生成、图像描述语句生成、视频片段检索和图像视觉问答几方面的相关工作和本文的关系进行综述。

本文提出的算法和其相关工作的具体细节和对比将在之后各章节中展示。

2.1 零样本物体分类

2.1.1 零样本学习

零样本物体分类，或零样本学习，通过利用所有类别（即：已见类别和未见类别）在属性空间的关联性^[19,28–32]，将类别属性看成是一个共同语义空间的中间特征，实现知识从已见类别到未见类别之间的迁移。早期的零样本学习模型通常为二阶段模型^[30,33–36]。在第一阶段，模型预测出图像包含的属性特征；在第二阶段，模型根据预测的属性特征推测出物体的类别。

为了扩大零样本迁移能力，目前的大多数方法都是基于嵌入映射的^[29,37–44]：通过学习图像的视觉特征向量和类别的语义嵌入向量之间的映射函数，实现零样本分类。最早的基于嵌入映射的模型是 SOC^[37]。SOC 通过将图像特征从视觉空间直接映射到语义空间，然后在语义空间与类别属性的嵌入向量进行相似度对比。之后，ALE^[39] 和 DeViSE^[38] 开始提出使用排序损失函数（ranking loss）作为优化目标。除了将视觉图像特征线性地映射到语义空间，LatEm^[41] 开始使用非线性的映射函数。CMT^[42] 使用两层神经网络的将视觉图像特征映射到语义空间中。为了进一步提升图像视觉特征的表达能力，Ba 等人^[45] 提出要让图像编码网络和映射网络一起端到端训练。同样地，Zhang 等人^[46] 认为图像的视觉特征空间比类别的语义空间具有更强的区分能力，通过将类别属性的嵌入向量映射到视觉空间中，然后通过端到端训练可以进一步提升零样本分类效果。与此同时，部分模型开始提出将图像的视觉特征和类别的语义特征都映射到其他的共同空间。例如：SSE^[47] 直接将已见类别的线性组合作为共同空间。JLSE^[48] 分别讲图像的视觉特征和类别的语义特征分别

映射到两个单独的子空间，然后计算两个子空间的相似度。

类别在语义空间中的嵌入向量主要来自于属性标注，但是由于属性信息需要大量的人工标注。一些工作开始研究使用其他的辅助信息来减少模型需要的属性标注。如 SJE^[40] 使用了四种嵌入向量：属性、word2vec^[49]、GloVe^[50] 和 wordnet^[51]。另外，嵌入向量不仅可以来自于单词，也可以来自于图像的描述语句^[45,52,53]。

2.1.2 通用型零样本学习

在之前的零样本物体分类（即：传统型零样本学习）中，测试阶段的所有图像都只来自于未见类别，即模型预测类别的选择空间只限于未见类别。这种实验设定大大降低了零样本学习的难度以及在实际场景中的运用价值。随着通用型零样本学习^[54] 的提出，模型在测试阶段需要对所有类别（已见类别和未见类别）的图像进行类别预测。Bendale 等人^[55] 首次将通用型零样本分类问题看成异常检测问题，利用一个网络先判断图像是否来自未见类别。对于通用型零样本学习，由于训练时使用了大量的已见类别图像而没有使用任何未见类别图像，模型往往倾向于将图像分类成已见类别。为了更好的评估通用型零样本分类，许多工作提出了不同的评价指标^[56,57] 来权衡已见类别和未见类别。

在本文，我们也重点研究通用型零样本物体分类。据我们了解，本文提出的属性保持的对抗网络学习模型（SP-AEN）是零样本分类算法中，第一个可以用语义空间的嵌入向量来重建原始图像。

2.1.3 零样本学习中的域偏移问题

属性损失问题，也常常被称为域偏移问题（domain shift）^[58,59]，是零样本学习、少样本学习^[60] 或领域自适应^[61,62] 等任务中一个非常普遍存在的问题。只要训练集数据和测试集数据的分布不同，就存在域偏移。目前的研究工作都发现重建原始信号可以缓解域偏移问题^[63]。在零样本分类领域，Kodirov 等人^[43] 通过同时进行图像分类和语义属性重建来缓解属性丢失问题。在本文，我们通过实验发现利用同一个网络同时进行重建和分类这两个相互冲突的任务对于保持语义特征并不是很有效。与此同时，另一种缓解语义损失的办法是增加一个单独的属性分类器^[64]，但是这个方法需要额外的属性标注信息。

2.1.4 零样本学习和对抗生成网络

对抗生成网络^[65] 主要是训练一个生成器，使得生成器生成的样本和真实数据非常“相似”，可以“骗过”判别器。理论上，这种对抗的训练过程可以让生成器

生成的样本分布和真实的样本分布完全相同。对于零样本分类问题，一些模型开始借助于对抗生成网络来生成更多的训练样本，从而将零样本分类问题转化为普通的分类问题^[66–68]。尽管这类方法目前已经取得很好的实验性能，但是它们违背了零样本分类问题的一个基本假设：训练阶段中测试集的类别信息是无法知道的。相反，本文提出的零样本分类模型 SP-AEN 只是将对抗网络中的对抗思想运用到特征层面上^[69–72]，使得分类特征向量的分布向重建特征向量的分布靠近，进而让分类特征向量尽可能地保持更多的属性特征，提升模型的迁移能力。

2.2 图像场景图生成

2.2.1 场景图生成

随着视觉关系检测任务^[73] 和大规模图像场景图数据集^[3] 的出现，图像场景图生成任务逐渐成为一个新的研究热点。目前，绝大多数的场景图生成模型都是将场景图生成任务分为两个步骤：1) 利用预训练好的目标检测器^[73–78] 或在图像场景图生成数据重新微调的目标检测器^[79–84] 对图像进行目标检测。2) 对于所有检测的物体的类别以及两两物体间的视觉关系类别进行预测。在早期阶段，许多模型都将物体类别预测和视觉关系类别预测拆分成两个独立的任务^[73–75,83,85]。这样的拆分忽略了场景内所有所有物体间的内在联系。为了充分利用场景内所有物体间的诱导偏置（inductive bias），一些场景图生成模型开始利用信息传递机制（message passing）^[76,80,81,86–93]。例如：Xu 等人^[80] 将物体和视觉关系分别看成两个独立的递归神经网络的节点，在递归神经网络的每个迭代过程中两个节点互相传递信息，每个物体节点和视觉关系节点通过接收传递的信息来更新自身节点的内部特征表达，从而充分利用图像的上下文全局信息。Li 等人^[86] 通过引入图像密集描述生成任务，构建三层语义节点（物体、视觉关系以及密集描述框），然后同样利用信息传递机制更新三种节点的特征表达。Yang 等人^[89] 利用图卷积，对每个物体周围的特征进行加权。之后，Zellers 等人^[82] 发现目前数据集中标注的视觉关系包含大量的文本偏置，即利用物体类别和视觉关系类别之间的统计频率就可以得到较高的视觉关系预测准确率。目前，最新的模型都将训练集图像中视觉关系的统计频率作为场景图中视觉关系的先验知识。

然后，之前所有的方法都关注于如何编码周围的信息来增强物体特征，所有的这些模型都使用物体和视觉关系分类的交叉熵作为模型最终的优化目标。这个优化目标将图像中所有的物体的重要性看成完全相同，即忽略了不同物体间的重要性，

容易让模型陷入局部最优解。本文，首次将最终的场景图生成质量作为整体优化目标，并首次将场景图生成任务看成多智能体协同合作的决策问题。

2.2.2 多智能体策略梯度

策略梯度即是一种强化学习的优化策略，同样也是一种对不可导优化目标进行优化的方法。在视觉场景理解任务中，策略梯度已经广泛应用在多种任务中，如：图像描述生成^[94-99]，图像视觉问答^[100,101]，图文匹配^[102,103]，视觉对话^[104]，和目标检测^[105-107]。目前，所有的场景图生成模型中，只有 Liang 等人^[108] 将图像场景图生成任务看成一个单智能体的序列决策过程。它首先根据文本的先验信息构建一个有向的语义图，在决策过程的每一个步骤中，选取新的物体和视觉关系，逐渐构建视觉场景图。相反，本文将图像场景图生成任务看成一个多智能体协同决策过程。通过构建成多智能体协同决策过程，我们可以直接使用整体场景图生成质量作为优化目标。另外，与目前许多现有的多智能体协同工作不同^[109,110]，本文的模型中，单个图像中智能体的数量（64 个检测物体）与动作空间范围（151 个物体类别）都非常大。

2.3 图像描述生成

2.3.1 编码-解码框架

图像描述生成任务（Image Captioning）通常被认为是一种多模态的“翻译”任务，即模型将视觉图像“翻译”成自然语言描述。由于端到端编码-解码框架在机器翻译任务（Neural Machine Translation, NMT）^[111] 的成功，许多的图像描述生成模型也开始借鉴使用编码-解码框架。最早的基于编码-解码框架的图像描述生成模型是 NIC^[22]。NIC 用一个卷积神经网络将原始输入图像编码成一个固定的视觉特征向量，然后将该视觉特征向量作为一个递归神经网络的初始时刻的输入，利用递归神经网络逐步将视觉特征向量解码成描述语句。类似地，Karpathy 等人^[112] 将编码的视觉特征向量作为递归神经网络隐含状态的初始化，通过引入一个额外的“START”字符触发递归神经网络对视觉特征进行解码。

由于在大规模图像分类数据集 ImageNet 预训练的卷积神经网络（如：VGG^[9]、GoogLeNet^[10]、ResNet^[11] 等）通常可以提取较好的图像视觉特征，之后的许多基于编码-解码框架的改进工作主要集中于完善解码过程。例如，Donahue 等人^[113] 和 Mao 等人^[114] 提出在递归神经网络迭代的每个时刻都输入视觉特征向量，避免生成

句子过长时图像特征的影响逐渐减弱。Wang 等人^[115]提出使用双边递归神经网络作为解码器，避免单向递归神经网络 LSTM^[116]只考虑之前时刻的单词信息。

2.3.2 注意力机制

在解码器生成语句的过程中，可以通过引入注意力机制^[117]使得模型在预测每个单词的时候动态地调整视觉特征向量，增强编码器的表达能力。

空间注意力机制：Xu 等人^[118]首次将注意力机制应用于图像描述生成任务中。具体来说，Xu 等人在卷积神经网络的最后一层特征图中引入空间注意力机制，让模型在每个时刻动态地关注不同的空间区域，合成新的视觉特征。类似地，Zhu 等人将同样的空间注意力机制也运用到图像视觉问答任务^[119]。除了在最后一层特征图只使用一次空间注意力加权，Yang 等人^[120]和 Xu 等人^[121]提出通过叠加使用多次空间注意力加权来提升模型性能。相比于之前的模型只在卷积神经网络的特征图中使用空间注意力加权，Anderson 等人^[122]和 Li 等人^[123]提出先对图像进行目标检测，然后对物体级别特征使用空间注意力机制可以进一步提升模型性能。

语义注意力机制：除了对图像的不同视觉区域加权，一些模型开始引入图像包含的属性，作为额外的语义信息，来引导语句的生成^[124-127]。其中，You 等人^[125]提出语义注意力机制，通过对不同的属性赋予不同的权重，在生成单词的过程中不断关注相关的属性。Yao 等人^[127]不仅引入属性来提升图像的描述语句生成，同时研究如何设计网络结构使得模型在生成描述语句过程中能够更好的利用属性信息。Jia 等人^[128]通过将图像和当前生成语句的相关性看成一个全局的语义信息，来引导语句生成。但是，这些模型需要额外地提取语义信息，如属性等。在本文提出的多层空间和通道注意力网络中，每个卷积核可以看成是多个语义检测器^[129]。因此，通道注意力机制可以看成是一种特定的语义注意力机制。

自注意力机制：自注意力机制^[130]先将特征集合（如图像区域特征）通过不同的映射矩阵分别映射为查询特征矩阵（query）、键特征矩阵（key）和值特征矩阵（value）。通过查询特征矩阵和键特征矩阵计算相似度得到注意力权重矩阵，然后对值特征矩阵进行加权，得到加权后的特征向量。基于自注意力机制的图像描述生成主要是利用 Transformer 结构^[130]：用编码网络对图像进行编码，解码网络对编码特征进行解码。其中编码网络和解码网络中都包含大量的自注意力机制模块。Herdade 等人^[131]首次基于 Transformer 结构，并且将图像中物体间的几何位置作为重要的语义信息，帮助生成描述语句。与此同时，Li 等人^[132]通过引入外部的语义标签来增强图像描述的内容。Huang 等人^[133]引入额外的注意力机制对加权后的特征进行

再一次加权。不同于之前的模型都只基于原始的 Transformer 结构，Cornia 等人^[134]通过将编码网络中所有层级的特征都作为每层解码网络的输入，对 Transformer 结果进行改进，进一步提升描述语句生成质量。

2.4 视频片段检索

2.4.1 视频片段检索

基于语句查询的视频片段检索：基于语句查询的视频片段检索，是一个典型的多模态问题。目前，主流的方法都是基于自顶向下的框架，这些方法主要关注如何设计更强的多模态融合模型，如基于视频的查询注意力机制^[135]、基于查询的视频注意力机制^[136]、查询-视频的协同注意力机制^[137-139]。据我们了解，绝大多数的模型都是自顶向下或自底向上框架，只有两个例外：RWM^[140] 和 SM-RL^[141]。这两个方法都是将视频片段检索问题转换成时序决策问题，然后利用梯度策略进行优化，其中的动作空间为时序窗口的变化或帧的跳变。

基于视频查询的视频片段检索：基于视频查询的视频片段检索的主要困难来自于查询视频和参考视频之间巨大的场景差异，包括背景、物体、视角等不同。目前最好的视频查询的视频片段检索是 CGBM^[142]，它也是基于稀疏型自底向上模型。

2.4.2 自顶向下与自底向上

自顶向下和自底向上方法，是计算机视觉研究领域解决问题的两种重要的思考哲学。与本文提出的自底向上框架最为接近的概念是：

目标检测：随着目标检测器 Faster R-CNN^[14] 的出现，目前绝大多数的目标检测算法都是基于自顶向下的框架，即在每个位置上预先定义大量的锚框，然后对每个锚框进行类别分类和坐标位置回归。这类自顶向下的模型拥有和自顶向下的视频片段查询方法一致的缺点。随着第一个性能相当的自底向上目标检测器的出现：CornerNet^[143]，自底向上的目标检测模型开始逐渐受到关注，近期也推出了大量的自底向上的目标检测器^[144-147]。

注意力机制：自底向上注意力机制在许多的视觉-文本等多模态任务中发挥重要的作用，例如：图像描述生成^[118,148]，视觉问答^[121,149] 等。最近，Anderson 等人^[122]提出融合自底向上和自顶向下的注意力机制，进一步提升了多个任务的模型性能。因此，如何更好的融合或权衡自顶向下与自底向上两种方法，将是未来的一个重要的研究方向。

2.5 图像视觉问答

2.5.1 视觉问答模型

典型的图像视觉问答模型通常包含一个图像编码模块、一个文本编码模块、一个融合模块和一个分类器。图像编码模块和文本编码模块分别对图像和问句进行编码，然后利用融合模块将两个模态的编码向量进行融合，最后分类器直接基于融合后的多模态特征进行答案预测^[24,150,151]。由于每个不同的问句往往涉及到图像的不同区域，之前的模型利用图像编码模块对图像提取统一的全局特征容易忽略一些重要的细节。为了避免上述问题，一些视觉问答模型开始引入注意力机制来动态提取视觉特征。Chen 等人^[152] 将问句的特征向量映射成一个卷积核，然后对图像的视觉特征卷积层进行卷积。Yang 等人^[120] 通过迭代多个空间注意力机制，逐渐优化空间注意力的权重。Fukui 等人^[153]、Kim 等人^[154,155]、Yu 等人^[156,157]、Ben 等人^[158,159] 分别提出了不同的多模态双线性融合方式来融合文本特征（即问句）和不同图像区域特征。Anderson 等人^[122] 提出自底向上和自顶向下注意力机制，通过对物体框进行注意力机制，而不是直接对图像的不同特征层区域使用注意力机制。

除了对图像进行充分的理解外，视觉问答同时还需要对问句进行理解。因此，对问句使用注意力机制同样可以提取更加重要的文本信息。Lu 等人^[160] 提出协同注意力机制（co-attention mechanism）同时对图像和问句进行注意力加权。Yu 等人^[157] 通过在问句中使用自注意力机制对协同注意力机制进行简化。早期的协同注意力机制只是在单个模态下分别进行注意力加权，之后，Nguyen 等人^[161]、Kim 等人^[155]、Gao 等人^[162]、Yu 等人^[163] 分别提出“密集型”协同注意力机制，充分计算每一个单词特征和每一个图像视觉特征之间的交互，进一步提升视觉问答性能。

2.5.2 视觉问答模型的文本偏置

尽管所有的视觉问答模型都在融合了多模态的特征之后进行答案预测，但是大量的研究工作表明^[164-167]，目前的视觉问答模型预测答案时往往都依赖文本偏置。为了减少文本偏置对视觉问答的影响，主要有两种解决方案：

1. **收集更平衡的数据集：**减少文本偏置最简单的方法就是收集更加平衡的数据集。例如：Zhang 等人^[166] 通过对抽象数据集中所有的判断题收集互补的场景图像，使得判断题的答案刚好完全相反。Goyal 等人^[167] 对真实图像数据集的所有类型的问题都收集互补的图像，使得标准答案与原始图像不同。虽然这些数据集在一定程度上可以缓解文本偏置的问题，但是由于训练集和测试集的问题答案分布基

本一致，目前的模型仍然可以通过统计的偏置得到较高的准确率^[168]。例如，在数据集 VQA-CP 中，当训练集和测试集的问题答案分布不同时，模型都会出现明显的性能下降。本文，我们同样是参考这个思路，来生成更多的互补样本。相反，本文提出的反事实样本生成机制（Counterfactual Samples Synthesizing, CSS）不需要任何额外的人工标注。

2. 设计更合理的去偏置模型：另一种减少文本偏置的方法就是设计更合理的去偏置模型。到目前为止，效果最后的去偏置模型就是复合模型^[169-174]：通过引入一个辅助网络来约束视觉问答网络的训练，其中的辅助网络只用文本（即问题语句）作为输入。在本文，我们提出的反事实样本生成机制可以无缝地应用于复合模型中，进一步减少文本偏置的影响。

2.5.3 视觉问答模型的特性

视觉可解释性和文本敏感性是一个理想的视觉问答模型必不可少的两个特性：

1. 视觉可解释性：为了提升视觉可解释性，早期的视觉问答模型^[175-177]都直接使用人类关注的区域作为额外的监督信息。然而，由于强大的文本偏置，即使模块关注到正确的图像区域，后续的网络结构仍然会忽视这部分视觉信号^[178]。因此，最近的一些视觉问答工作^[178,179]开始使用工具 Grad-CAM^[180]来计算每个物体对正确答案的贡献，然后让不同物体的贡献度与人工标注的一致。但是，这类方法有两个明显的缺点：1) 它们需要额外的人工标注，即每个物体的贡献度大小或者排序；2) 目前这类方法都不是端到端训练的。

2. 文本敏感性：如果一个视觉问答模型能够充分理解问题的含义，那么这个模型应该对问题的变化十分的敏感，即具备文本敏感性。据我们所知，目前只有一个工作讨论了视觉问答模型的文本敏感性^[181]。具体来说，Shah 等人^[181]对于视觉问答和视觉问题生成两个对偶任务设计了一个循环一致的损失函数，然后通过引入随机噪声来生成不同的问题。然后，他们只考虑了对不同语句表达时的鲁棒性。相反，本文提出的反事实样本生成机制帮助模型能够感知重要单词的改变，提升文本敏感性。

目前的方法都只关注于视觉可解释性或者文本敏感性。相反，本文提出的反事实样本生成机制可以提升着两个特性。

3 基于属性保持对抗网络的零样本物体分类方法

零样本物体分类问题是指当训练集和测试集包含的图像类别不同时，模型仍然能够准确地对训练过程中从未见过类别的图像进行分类。目前，虽然主流的零样本物体分类方法都是基于嵌入映射的模型，但是这类方法不可避免地存在一定的语义丢失（semantic loss）。语义丢失是指在零样本分类的训练过程中，模型往往选择“丢失”一些对训练类别区分性不大的属性（即同一类别方差较小的属性）来提升训练集的分类准确率。然而，由于零样本分类中，训练类别和测试类别存在较大差异，即丢失的属性可能对测试集区分性较大。为了缓解语义丢失的问题，在本章，我们提出了一个全新的零样本物体分类网络：属性保持对抗网络（Semantics-Preserving Adversarial Embedding Networks, SP-AEN）。具体来说，SP-AEN 通过引入一个新的映射函数，将两个冲突的任务（图像分类和图像重建）进行分离，分别映射到两个不同的子空间中。通过对抗网络学习，SP-AEN 可以让重建子空间中部分属性迁移至分类子空间，让分类子空间中的特征向量保持尽可能多的属性。通过在大量的数据集上与现有的零样本分类方法进行对比，实验发现，SP-AEN 不仅仅在零样本分类性能上有大幅提升，同时可以重建出非常逼真的图像，表现出非常好的属性保持能力。在零样本物体分类任务通用的四个数据集 CUB、AWA、SUN、aPY 中，SP-AEN 比目前性能最好的零样本分类方法^[182] 在 H 值上能够分别提升了 12.2%、9.3%、4.0% 和 3.6%。

3.1 问题描述

零样本物体识别（Zero-Shot Recognition, ZSR），或零样本学习（Zero-Shot Learning, ZSL）是为了能够让模型对训练过程中从未见过的新类别图像进行分类识别。目前，关于零样本物体分类问题的难点，学界的共识是如何将训练集中已见类别的知识迁移到测试集的未见类别上。到目前为止，已经有非常多的零样本分类方法，并且这些方法基本都是依据一些非常容易理解的直觉。例如，虽然训练集中不包含“浣熊”这个类别的图像，但是在测试阶段我们仍然可以通过检查浣熊这个类别特

有的一些属性特征对浣熊图像进行识别，像“有条纹的尾巴”^[19,28,183,184]、“像狐狸的外观”^[184,185]、以及“浣熊”这个类别的语义信息^[49,50]等。这些属性特征通常在训练阶段被建模，然后在测试阶段时对所有的类别（已见类别和未见类别）共享。经过数十年的发展，目前的零样本物体分类方法已经从初始的基于属性分类器的模型^[19]发展到基于嵌入映射的模型^[38,39,186]。如图 3-1 (a) 所示，这种基于嵌入映射的模型首先将图像从视觉空间 (\mathcal{V}) 映射到语义空间 (\mathcal{S})，同时，类别的属性特征向量也在语义空间 \mathcal{S} 中。通过这样的空间映射之后，零样本物体分类问题就简化成了语义空间中一个最近邻类别查找问题——模型在语义空间中选择最近的类别作为图像的分类结果。

然而，这种嵌入映射模型的知识迁移能力受限于语义丢失的问题。如图 3-1 所示，模型在训练过程中往往丢失一些对训练集图像方差较小的属性（即，不同类别之间区别小的属性），这样有利于训练集的分类。然而，由于训练集和测试集之间的类别存在差异，这些丢失的属性可能对测试集的类别图像来说具有较明显

的区别性，进而就会增加测试集分类的困难。虽然每个类别在语义空间 \mathcal{S} 中都只是一个单独的“点”，能够具有丰富的语义信息，但是将所有同类的图像从视觉空间 \mathcal{V} 映射到这个点附近，就不可避免地造成部分属性的丢失^[58,187]。

为了尽可能多地减少训练过程中属性的丢失，一个可能的解决方法是通过图像重建，即先将图像从视觉空间映射到语义空间中，然后将语义空间的特征映射回视觉空间。如果映射回视觉空间的特征能够重建初始的图像，说明语义空间的特征已经尽可能多地保持了原有的属性，否则图像将无法重建^[63,188-190]。然而，图像重建和图像分类是两个相互冲突的目标：前者希望能够尽可能多地保持图像的细节，而后者希望只关注类别差异性大的特征，忽略不相关的特征。例如，只用“头”或者“躯干”就可以充分地对“人”这个类别进行识别分类，而一些其他的颜色属性，如“红色”或者“白色”就需要忽略。为了进一步展示重建和分类这两个冲突任务，

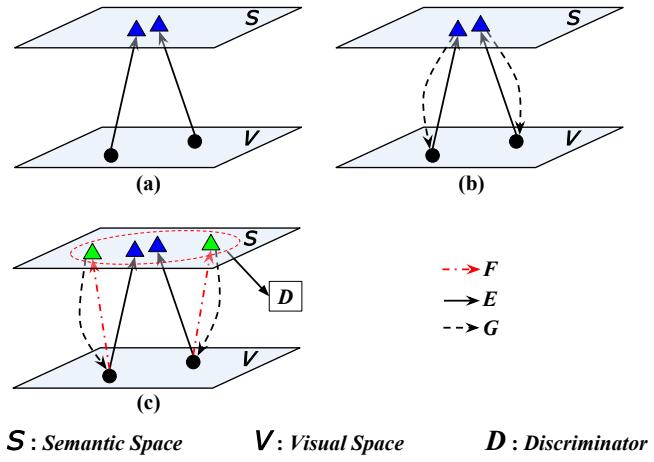


图 3-1 三种典型的零样本学习框架

如图 3-1 (b) 所示, 假设 $E: \mathcal{V} \rightarrow \mathcal{S}$ 和 $G: \mathcal{S} \rightarrow \mathcal{V}$ 是视觉空间 \mathcal{V} 与语义空间 \mathcal{S} 中的两个映射函数。对于图像分类任务, 我们希望视觉空间中同一类别的两个图像 $x, x' \in \mathcal{V}$ 在语义空间能够接近 $s, s' \in \mathcal{S}$, 即 $E(x) = s \approx s' = E(x')$; 对于图像重建任务, 我们希望 $G(s) \approx x$ 和 $G(s') \approx x'$, 这样就很难满足 $s \approx s'$ 。因此, 同时训练图像分类和图像重建两个目标, 对于保持属性的效果往往有限 (如模型 SAE^[43])。如图 3-2 所示, 当我们想实现好的图像分类结果时, 图像重建往往会失败。



图 3-2 模型 SAE 和模型 SP-AEN 的图像重建结果对比

为了缓解图像分类任务和图像重建任务之间的冲突, 本文提出一个全新的零样本物体分类网络: 属性保持对抗网络 (SP-AEN)。如图 3-1 (c) 所示, 我们引入一个新的映射函数 $F: \mathcal{V} \rightarrow \mathcal{S}$ 和一个对抗优化目标^[65]。映射函数 F 和对抗训练的目的是让判别器 D 无法区分这两个不同的映射分布 $E(x)$ 和 $F(x)$ 。具体来说, 这样做有两个好处: (1) 语义迁移: 虽然对于单独的分类映射函数 E 来说, 语义丢失问题是不可避免的。但是, 我们通过利用判别器 D 的训练, 让分类映射向量 $E(x)$ 和重建映射向量 $F(x)$ 在同一个分布下, 实现属性特征的迁移, 让 $E(x)$ 尽可能地保持更多的属性。(2) 分类任务与重建任务分解: 映射网络 F 和 G 实现重建任务, 而映射网络 E 实现分类任务。通过将分类任务和重建任务进行分解, 之前的严格条件 $G(E(x)) \approx x$ 和 $G(E(x')) \approx x'$ 变成了 $G(F(x)) \approx x$ 和 $G(F(x')) \approx x'$, 同时 $F(x)$ 与 $F(x')$ 在语义空间中也不需要非常接近。如图 3-2 所示, 我们的映射 $G(F(x))$ 可以重建出较好的输入图像, 说明属性特征能够更好地保持。

本文在零样本分类任务通用的四个数据集 CUB^[191]、AWA^[19]、SUN^[192] 和 aPY^[28] 中对模型 SP-AEN 的效果进行验证。相比于目前性能最好的零样本分类方法^[182], SP-AEN 在评价指标 H 值 (harmonic mean value) 上对于上述四个数据集分别提升了 12.2%、9.3%、4.0% 和 3.6%。同时, SP-AEN 是第一个能够直接重建回原始图像的零样本分类模型。

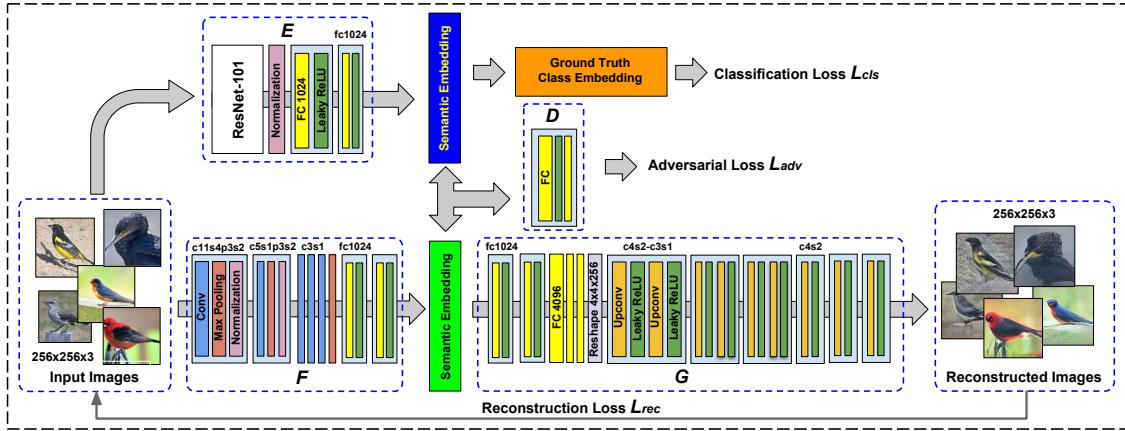


图 3-3 模型 SP-AEN 的整体网络结构流程图

3.2 属性保持对抗网络

在本节，我们首先介绍零样本分类任务，然后再具体介绍本章提出模型 SP-AEN 的各个优化目标的细节。

3.2.1 零样本分类预备知识

给定一个训练集 $\{x_i, l_i\}$ ，其中 $x_i \in \mathcal{V}$ 是图像在视觉空间的映射向量， $l_i \in \mathcal{L}_s$ 是已见类别的类别标签。零样本物体分类任务的目标是学习一个分类器，不仅可以预测已见类别的图像 (\mathcal{L}_s)，也可以预测未见类别的图像 (\mathcal{L}_u)。按照之前零样本物体分类方法的总结^[45,182]，几乎所有先进的零样本分类方法都是基于嵌入映射框架的。这类方法旨在找到一个映射函数 $E: \mathcal{V} \rightarrow \mathcal{S}$ ，其中所有的类别标签在语义空间 \mathcal{S} 都是编码成 $y_l \in \mathbb{R}^d$ 。因此，预测类别标签 l^* 时，可以直接通过简单的最近邻查找得到：

$$l^* = \max_{l \in \mathcal{L}} y_l^T E(x) \quad (3-1)$$

特别地，如果 $l \in \mathcal{L}_u$ ，这类零样本分类问题称为**传统型零样本分类** (conventional ZSL)；如果 $l \in \mathcal{L}_s \cup \mathcal{L}_u$ ，这类零样本分类问题称为**通用型零样本分类** (generalized ZSL)。另外，公式 (3-1) 中映射函数 E 即可以是线性函数，也可以是基于深度神经网络的非线性函数。

3.2.2 分类任务优化目标

因为公式 (3-1) 的类别预测问题本质上是一个排序问题，所以我们直接使用排序损失函数 (ranking loss) 作为分类任务的优化目标^[38,186]。具体来说，给定一个训

练习样本 (x, l) , 我们希望 \mathbf{y}_l 和 $E(x)$ 之间的点积相似度较大; 而对于负样本 (x, l') , $\mathbf{y}_{l'}$ 和 $E(x)$ 之间的点积相似度较小; 同时正样本的相似度相比负样本的相似度要大于一定的阈值 γ :

$$L_{cls} = \sum_{l \neq l'} \max\{0, \gamma - \mathbf{y}_l^T E(x) + \mathbf{y}_{l'}^T E(x)\} \quad (3-2)$$

其中, 阈值 γ 是一个大于 0 的超参数。在每次训练迭代过程中, l' 是从所有错误的类别标签中随机任选的一个。

分类任务的优化目标 L_{cls} 主要是让所有的同一类图像的语义空间映射向量 $E(x)$ 都接近与其类别标签在语义空间的映射向量 \mathbf{y} 。它造成的语义丢失问题将由后续介绍的两个优化目标 (重建任务优化目标和对抗学习优化目标) 来解决。

3.2.3 重建任务优化目标

重建任务的目标是学习一个从语义空间到视觉空间的映射函数 $G: \mathcal{S} \rightarrow \mathcal{V}$, 使得将语义映射向量 $s \in \mathcal{S}$ 可以重建出输入图像, 并且使差别 $\|G(s) - x\|$ 很小。由于在自编码器形式的网络中重建任务 $s = E(x)$ 与分类任务是相互冲突的, 我们引入一个新的视觉空间到语义空间的映射函数 F , 使得 $s = F(x)$ 。另外, 不同于 SAE^[43] 利用卷积神经网络^[9,11] 的高层输出向量作为视觉空间 \mathcal{V} 的特征, 我们直接利用原始的 $256 \times 256 \times 3$ 大小的 RGB 色彩空间来进行图像重建。这样做的主要原因是卷积神经网络的输出特征在网络的预训练阶段已经存在语义丢失。

为了最小化重建损失, 映射向量 $F(x)$ 会尽可能地保持多的属性, 以便于能够重建回原始输入图像。我们参考最新的图像生成工作^[193–195], 将重建任务的优化目标定义为:

$$L_{rec} = L_{feat} + \lambda_p L_{pixel} \quad (3-3)$$

其中 $L_{feat} = \|\phi(G(F(x))) - \phi(x)\|_2^2$ 是特征空间上的重建损失函数, 帮助图像能够保持细节感知上的相似度。我们使用卷积神经网络 AlexNet^[7] 中 conv5 层来表示函数 ϕ 。 $L_{pixel} = \|G(F(x)) - x\|_2^2$ 是像素空间上的重建损失函数, 有利于重建算法训练过程的稳定性。

3.2.4 对抗学习优化目标

到目前为止, 语义向量 $E(x)$ 和 $F(x)$ 之间还没有进行知识迁移。我们的目标是让 $E(x)$ 从 $F(x)$ 中迁移部分丢失的属性特征。因为手工直接定义 $E(x)$ 与 $F(x)$

之间的迁移方式比较困难，所以我们借助对抗学习的思想，来鼓励 $E(x)$ 的分布向 $F(x)$ 的分布靠近，通过“欺骗”判别网络 D ，来实现 $F(x)$ 的知识向 $E(x)$ 迁移：

$$L_{adv} = \mathbb{E}_x (\log D(F(x))) + \mathbb{E}_{x'} (\log [1 - D(E(x'))]) \quad (3-4)$$

其中网络 E 为了减少优化目标 L_{adv} ，而网络 D 希望增大优化目标 L_{adv} ，即： $E^* = \arg \min_E \max_D L_{adv}$ 。

目前，许多研究工作都发现目标函数 L_{adv} 的优化过程容易导致塌陷问题^[196] (mode collapse)。在零样本物体分类任务中，如果两个同类别的两个图像 x 和 x' 非常相似，同样容易导致 $\|F(x) - E(x')\| \approx 0$ ，引起塌陷问题。为了避免塌陷问题，我们参考 WGAN^[196] 的训练策略，极大地增加了模型训练的稳定性。

3.2.5 总体优化目标

将前文提到的分类任务优化目标、重建任务优化目标、以及对抗学习优化目标合在一起，得到模型 SP-AEN 最终整体的优化目标：

$$L(E, F, G, D) = L_{cls}(E) + \alpha L_{rec}(E, F, G) + \beta L_{adv}(E, F, G, D) \quad (3-5)$$

其中， α 和 β 是超参数用于权衡不同的优化目标函数的权重。我们最终的目标是得到：

$$E^* = \arg \min_{E, F, G} \max_D L(E, F, G, D) \quad (3-6)$$

如图 3-3 所示，网络 F 是重建任务编码网络，网络 G 是重建任务解码网络，重建任务映射向量 $F(x)$ 可以看成是瓶颈层，来约束分类任务映射向量 $E(x)$ 。另外，模型 SP-AEN 也可以无缝地运用到其他实验设定下的零样本物体分类问题，例如在半监督条件下，我们只需要给 $F(x)$ 增加一个额外的对抗学习优化目标来近似先验的映射空间。

3.3 实验设置与性能分析

3.3.1 零样本物体分类数据集

我们在四个通用的零样本物体分类数据集（CUB、SUN、AWA、aPY）上对模型 SP-AEN 的性能进行了验证。其中，我们参考 Xian 等人^[182] 使用新的数据集划分。因为之前的零样本分类工作都使用 ILSVRC ImageNet^[2] 上的 1000 个常见类图

像对卷积神经网络进行预训练，而这 1000 类与这四个数据集的原始划分都有重叠类别，即违背了零样本物体分类的基本设定。以下是这四个数据集的详细介绍：

CUB^[191]: 全称是 Caltech-UCSD-Birds 200-2011 数据集。它是一个细粒度鸟类数据集，总共包含 11788 张来自 200 个细粒度类别的鸟图像，并且每张图像标注了 312 个语义属性。在通用型零样本分类的实验设定中，训练集包含 150 个已见类别的 7057 张图像，测试集包含 150 个已见类别的 1764 张图像和 50 个未见类别的 2967 张图像。

SUN^[192]: 全称是 SUN attribute 数据集。它是一个细粒度场景分类数据集，总共包含 14340 张来自 717 个场景类别的场景图像，并且每张图像标注了 102 个语义属性。在通用型零样本分类的实验设定中，训练集包含 645 个已见类别的 10320 张图像，测试集包含 645 个已见类别的 2580 张图像和 72 个未见类别的 1440 张图像。

AWA^[19]: 全称是 Animals with Attributes 数据集。它是一个动物类别分类数据集，总共包含 30475 张来自 50 个类别的动物图像，并且每张图像标注了 85 个语义属性。在通用型零样本分类的实验设定中，训练集包含 40 个已见类别的 23527 张图像，测试集包含 40 个已见类别的 5882 张图像和 10 个未见类别的 7913 张图像。由于原始 AWA 数据集图像版权的问题，我们这里的 AWA 数据集实际上使用的是 AWA2^[57,182]。

aPY^[28]: 全称是 Attribute Pascal and Yahoo 数据集。它是一个通用的物体分类数据集，总共包含 12051 张来自 32 个类别，并且每张图像标注了 64 个语义属性。在通用型零样本分类的实验设定中，训练集包含 20 个已见类别 5932 张图像，测试集包含 20 个已见类别 1483 张图像和 12 个未见类别的 7924 张图像。

为了公平地和其他零样本物体分类模型进行比较，我们参考 Xian 等人^[182] 使用完全相同的类别嵌入映射向量，并对所有的嵌入映射向量进行 l_2 范数归一化。

3.3.2 实验设定与零样本物体分类评价指标

实验设定: 为了充分地评估模型 SP-AEN 的零样本物体分类性能，我们在三种不同的实验设定下进行对比：

1. U→U: 测试图像的类别和预测的类别标签都只限于未见类别；
2. S→T: 测试图像的类别只限于未见类别，但预测的类别标签可以是未见类别或已见类别；
3. U→T: 测试图像的类别和预测的类别标签都可以是未见类别或已见类别。

通常， $U \rightarrow U$ 被称为传统型零样本分类，而 $U \rightarrow T$ 被称为通用型零样本分类。

评价指标：我们参考之前的工作^[182]，使用每个类别的平均准确率作为三个实验设定下的评价指标（即： $Acc_{U \rightarrow U}$ 、 $Acc_{S \rightarrow T}$ 和 $Acc_{U \rightarrow T}$ ）。对于通用型零样本分类，我们另外使用常用的 H 值作为主要的评价指标，其中 H 值是已见类别的准确率 $Acc_{S \rightarrow T}$ 和未见类别的准确率 $Acc_{U \rightarrow T}$ 的调和平均数：

$$H = 2 \times Acc_{S \rightarrow T} \times Acc_{U \rightarrow T} / (Acc_{S \rightarrow T} + Acc_{U \rightarrow T}) \quad (3-7)$$

3.3.3 网络模型与训练细节

网络结构：整个模型 SP-AEN 的网络结构如图 3-3 所示。其中分类映射网络 E 是基于网络 ResNet-101^[11]。网络 E 的输入是大小为 $224 \times 224 \times 3$ 的原始图像，输出是 d 维的编码向量，然后输入到分类优化函数中（即公式 (3-2)）。重建编码网络 F 是基于网络 AlexNet^[7]，并加上两层额外的全连接层。和分类网络 E 类似，输入是 $224 \times 224 \times 3$ 的原始图像，输出是 d 维的编码向量。但是，输出的编码向量再输给重建解码网络 G 。重建解码网络 G 通过五个连续的反卷积和非线性激活函数 (leaky ReLU^[197]) 将向量转换成三维特征图。另外，在重建解码网络的头部我们使用两个全连接层先将编码向量的维度映射到 4096 维。判别器网络 D 是一个两层的全连接层加一个非线性 ReLU 层，网络 D 的输入为 d 维的编码向量。

训练细节：对于本章所有的零样本分类实验，我们都先将训练图像的短边放缩到 256，然后参照 AlexNet^[7] 数据增强的策略增大十倍训练数据。为了提升训练速度，分类映射网络 E 中 ResNet-101 采用 ImageNet 数据集上预训练好的 ResNet-101 的参数，并始终固定参数。重建编码网络 F 的参数初始化采用 ImageNet 数据集上预训练好的 AlexNet 的参数，重建解码网络 G 采用预训练好的生成器^[194] 作初始化。剩余网络中的所有参数都采用 MSRA 初始化^[197]。初始的学习率设置为 $1e^{-4}$ ，训练过程中当损失函数不再下降时，学习率降低 10 倍继续训练。

3.3.4 零样本物体分类的性能对比

本节将本章提出的模型 SP-AEN 与目前最先进的零样本物体分类方法进行对比。这些方法主要可以分类：(1) 基于嵌入映射的模型：**DeViSE**^[38]、**ALE**^[39]、**SJE**^[40]、**ESZSL**^[29]、**LTM**^[41]、**CMT**^[42] 和 **SAE**^[43]。这类方法和 SP-AEN 一样，都是将图像从视觉空间映射到语义空间，其中 SAE 是目前现有的唯一一个利用信号重建来解决语义损失的模型。(2) 基于属性的模型：**DAP**^[19]、**IAP**^[19]、**SSE**^[47]、**CSE**^[30] 和 **SYNC**^[198]。特别地，这类方法只适用于所有类别都有人工标注属性的情况下。

Dataset		DAP	IAP	SSE	CSE	SYNC	CMT	LTM	DeViSE	ALE	SJE	ESZSL	SAE	SP-AEN
SUN	$Acc_{U \rightarrow U}$	39.9	19.4	51.5	38.8	56.3	39.9	55.3	56.5	58.1	53.7	54.5	40.3	59.2
	$Acc_{U \rightarrow T}$	4.2	1.0	2.1	6.8	7.9	8.1	14.7	16.9	21.8	14.7	11.0	8.8	24.9
	$Acc_{S \rightarrow T}$	25.1	37.8	36.4	39.9	43.3	21.8	28.8	27.4	33.1	30.5	27.9	18.0	38.6
CUB	H	7.2	1.8	4.0	11.6	13.4	11.8	19.5	20.9	26.3	19.8	15.8	11.8	30.3
	$Acc_{U \rightarrow U}$	40.0	24.0	43.9	34.3	55.6	34.6	49.3	52.0	54.9	53.9	53.9	33.3	55.4
	$Acc_{U \rightarrow T}$	1.7	0.2	8.5	1.6	11.5	7.2	15.2	23.8	23.7	23.5	12.6	7.8	34.7
AWA	$Acc_{S \rightarrow T}$	67.9	72.8	46.9	72.2	70.9	49.8	57.3	53.0	62.8	59.2	63.8	54.0	70.6
	H	3.3	0.4	14.4	3.1	19.8	12.6	24.0	32.8	34.4	33.6	21.0	13.6	46.6
	$Acc_{U \rightarrow T}$	46.1	35.9	61.0	44.5	46.6	37.9	55.8	59.7	62.5	61.9	58.6	54.1	58.5
aPY	$Acc_{S \rightarrow T}$	0.0	0.9	8.1	0.5	10.0	0.5	11.5	17.1	14.0	8.0	5.9	1.1	23.3
	H	84.7	87.6	82.5	90.6	90.5	90.0	77.3	74.7	81.8	73.9	77.8	82.2	90.9
	$Acc_{U \rightarrow U}$	0.0	1.8	14.8	1.0	18.0	1.0	20.0	27.8	23.9	14.4	11.0	2.2	37.1
	$Acc_{U \rightarrow T}$	33.8	36.6	34.0	26.9	23.9	28.0	35.2	39.8	39.7	32.9	38.3	8.3	24.1
	$Acc_{S \rightarrow T}$	4.8	5.7	0.2	0.0	7.4	1.4	0.1	4.9	4.6	3.7	2.4	0.4	13.7
	H	78.3	65.6	78.9	91.2	66.3	85.2	73.0	76.9	73.7	55.7	70.1	80.9	63.4

表 3-1 不同零样本分类模型在通用的四个零样本分类数据集上的性能对比

性能对比结果: 表 3-1 总结了不同的零样本分类模型在数据集 SUN、CUB、AWA 和 aPY) 和三种不同实验设定 ($U \rightarrow U$ 、 $U \rightarrow T$ 和 $S \rightarrow T$) 下的性能对比。从表 3-1 的结果可以发现: (1) 在通用型零样本分类中, SP-AEN 能够显著提升实验性能。例如: 在 $Acc_{U \rightarrow T}$ 和 H 值两个指标下, SP-AEN 能比目前最好的模型提升大约 4% 到 12%。尤其当数据集中训练集和测试集所有属性方差的余弦相似度越大时 (数据集 SUN、CUB、AWA、aPY 中, 训练集和测试集所有属性方差的余弦相似度分别为 0.9851、0.9575、0.7459、0.5847。), 性能提升更加明显, 这也从侧面反映模型 SP-AEN 可以有效地缓解语义丢失问题。(2) 在传统型零样本分类中, 在绝大多数的数据集中, SP-AEN 同样可以得到最佳的性能。造成部分实验效果欠佳的主要原因可能是因为在 $U \rightarrow U$ 的实验设定下, 图像类别的搜索空间仅限于未见类别。然而, 语义丢失可能导致未见类别的图像和已见类别非常接近, 引起错误的预测。

3.3.5 零样本物体分类方法分析

分类任务与重建任务的冲突: 为了验证本章提出模型 SP-AEN 的设计动机, 即分类任务和重建任务是相互冲突的。我们设计了三种图像重建网络框架, 如图 3-4 所示, 可以实现图像重建: (1) **DirectMap**: 对于输入图像, 我们首先使用网络 E 将图像从视觉空间映射到语义空间, 得到语义嵌入向量, 然后使用网络 G 将语义嵌入向量映射回视觉空间。在 DirectMap 中, 我们固定网络 E 的参数 (即预训练好

Method	SUN	CUB	AWA	aPY
DirectMap	0.079	0.069	0.075	0.085
SAE	0.285	0.281	0.259	0.275
SplitBranch	0.070	0.058	0.059	0.076
SP-AEN	0.053	0.040	0.047	0.055

表 3-2 不同重建网络下重建图像与输入图像之间的平均像素差平方

的 AlexNet 网络参数), 只训练网络 G 的参数。DirectMap 模型可以衡量初始的语义嵌入向量能够包含多少语义信息。(2) **SAE**: 我们采用与 SAE 模型^[43] 相同的框架, 用重建网络 G 作为解码网络, 分类网络 E 作为编码网络, 并且利用其中的瓶颈层来进行分类任务。在训练过程中, 我们同时更新网络 E 和网络 G 的参数。(3) **SplitBranch**: 我们将网络 E 的输出分别输入到两个不同的支路网络中, 并且对其中一条支路进行分类任务。然后将两条主路的输出合并到一起, 并输入到网络 G 中进行重建。

图 3-5 和表 3-2 分别表示四个数据集中测试集未见类别图像的重建图像和重建差异。从实验结果中, 我们可以发现:(1) 在数据集 CUB 和 SUN 中, DirectMap 重建的图像和 SP-AEN 非常接近, 都具有较好的重建结果。然而, 在数据集 AWA 和 aPY 中, DirectMap 的重建效果有明显下降。同样地, 这是由于在数据集 CUB 和 SUN 中, 训练集和测试集所有属性方差的余弦相似度较大, 而数据集 AWA 和 aPY 中两者的余弦相似度较小。(2) 对于 SAE 模型, 当同时训练分类网络 E 和重建网络 G 时, 所有的图像样本都重建失败。对于 SplitBranch 模型, 当同时训练分类网络 E 和重建网络 G 时, 重建效果得到显著的提升, 接近模型 SP-AEN。然而, 在 SplitBranch 中, 我们发现分类分支合并时的权重几乎为零。这一方面说明分类的映射向量基本对重建任务没有任何的贡献, 另一方面这也引导我们借助对抗

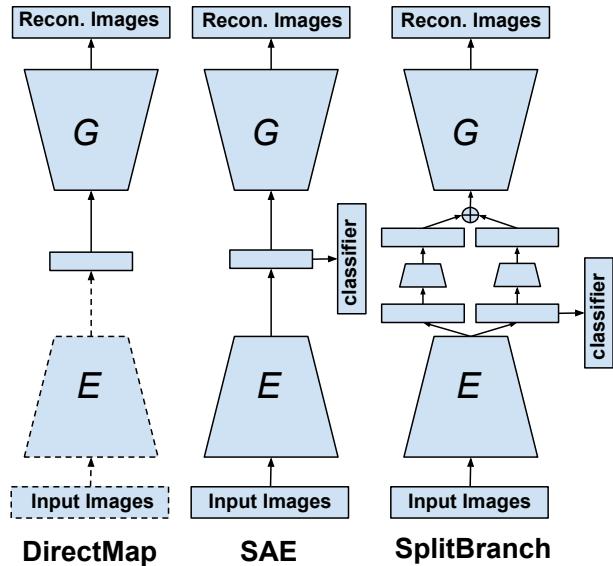


图 3-4 三种不同的图像重建网络框架

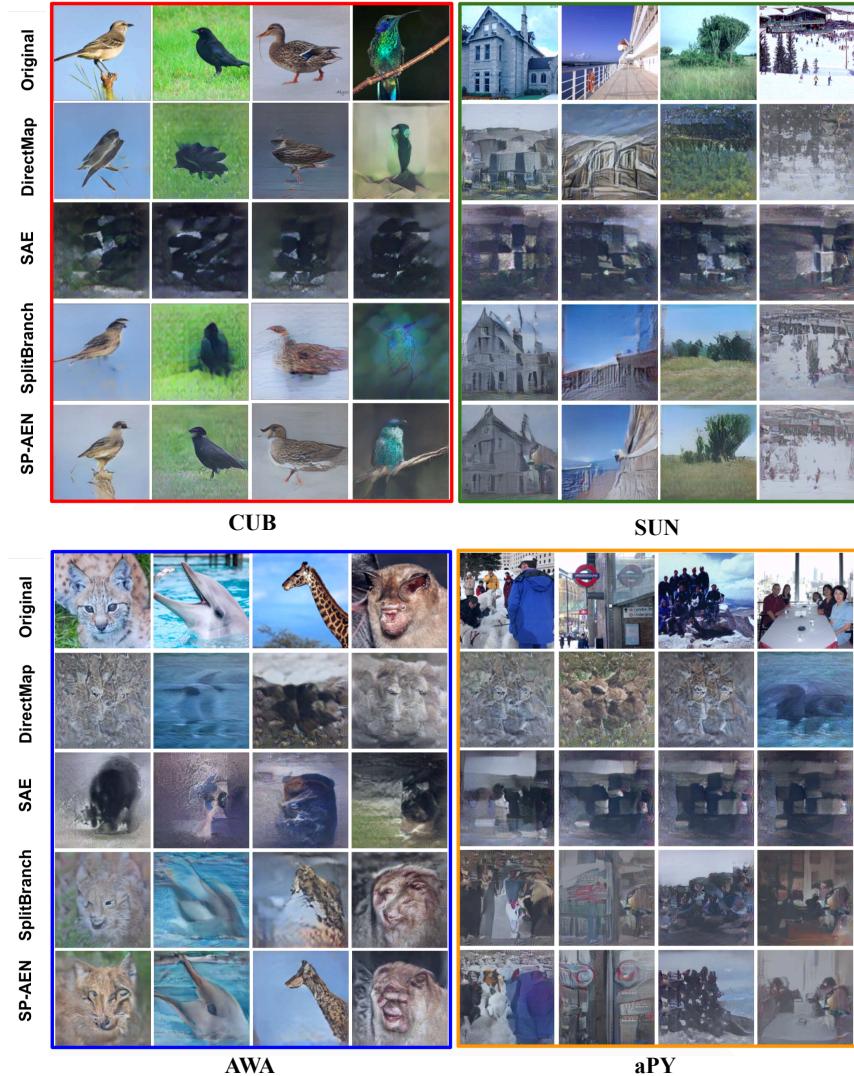


图 3-5 不同图像重建网络框架在四个零样本物体分类数据集中的重建结果

学习来实现语义迁移和高质量图像的重建。

网络 D 和网络 G 的有效性: 由于训练过程模型“见过”大量的已见类别图像，模型往往倾向于给已见类别赋予更高的分数。为了解决这个问题，通常使用“校准规则”(calibrated stacking rule)^[56]，即对已见类别的分数减掉一个常数偏置，然后再和未见类别的分数一起进行排序比较：

$$l^* = \max_{l \in \mathcal{L}_u \cup \mathcal{L}_s} \mathbf{y}_l^T E(x) - \gamma \mathbf{1}[l \in \mathcal{L}_s] \quad (3-8)$$

其中指示函数 $\mathbf{1}[\cdot]$ 用于判断类别 l 是否是已见类别， $\gamma \in \mathbb{R}$ 是校准系数。这个校准规则能够有效地实现对已见类别和未见类别预测之间的权衡。通过不同调整校准系数 γ ，可以得到一系列分类准确率 ($Acc_{U \rightarrow T}$ 和 $Acc_{S \rightarrow T}$) 和已见-未见准确率曲线 (Seen-Unseen accuracy Curve, SUC)。已见-未见准确率曲线下区域面积 (Area Under

Seen-Unseen accuracy Curve, AUSUC) 也是通用型零样本分类问题中一个常用评价指标，用来评估 $Acc_{U \rightarrow T}$ 和 $Acc_{S \rightarrow T}$ 之间的权衡。

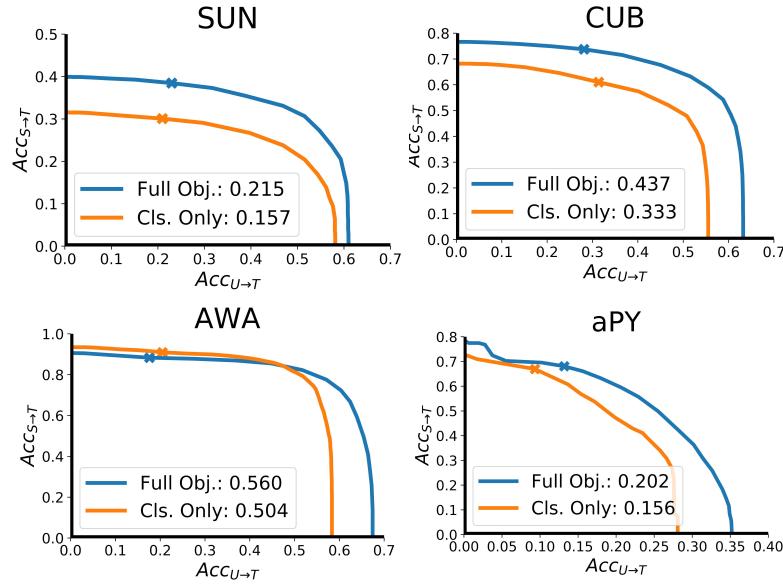


图 3-6 在四个零样本分类数据集中的已见-未见准确率曲线下区域面积

Dataset	SUN				CUB			
	U→U	U→T	S→T	H	U→U	U→T	S→T	H
Cls. Only	56.8	17.2	29.0	21.6	52.2	23.5	55.0	32.9
Full Obj.	59.2	24.9	38.6	30.3	55.4	34.7	70.6	46.6
Dataset	AWA				aPY			
	60.2	17.5	76.7	28.5	35.8	5.5	72.9	10.2
Full Obj.	58.5	23.3	90.9	37.1	24.1	13.7	63.4	22.6

表 3-3 SP-AEN 在不同优化目标条件下的性能对比

如图 3-6 所示，蓝色曲线表示模型 SP-AEN 使用所有的优化目标 (Full Obj.)，而黄色曲线表示 SP-AEN 只使用分类任务优化目标 (Cls. Only)。在所有的数据集中，SP-AEN (Full Obj.) 相比于 SP-AEN (Cls. Only) 都可以显著提升模型性能。表 3-3 展示了两种模型的定量性能对比。从实验结果可以看出，当使用对抗学习优化目标和重建学习优化目标时，在所有的数据集中 H 值可以显著提升超过 10%。如图 3-7，当逐渐减少重建网络的权重 α (从左向右变化，红框为原始输入图像)，图像重建的质量也逐渐降低。

图 3-7 图像重建结果随优化目标权重 α 的影响

3.4 本章小结

本章我们提出了一个全新的零样本分类网络：属性保持对抗网络（SP-AEN），用于解决目前零样本分类方法中普遍存在的语义丢失的问题。SP-AEN 主要通过两个设计来解决语义丢失：(1) 引入一个独立的重建编码网络和重建解码网络，使得重建任务的优化目标不直接影响分类任务网络的优化。(2) 通过引入对抗网络学习，实现重建网络编码向量和分类网络编码向量之间的知识迁移。在通用的四个标准零样本分类数据集中，我们通过大量的实验和可视化结果都证明了模型 SP-AEN 的有效性。

4 基于反事实多智能体学习的图像场景图生成方法

图像场景图 (scene graph) 是将图像中每个物体当成一个节点，两两物体（节点）之间视觉关系看成有向边，即视觉三元组“主语（物体） \rightarrow 谓语（视觉关系） \rightarrow 宾语（物体）”。图像场景图描绘了整个图像视觉场景中所有物体的类别、位置关系以及物体间的交互。为了生成准确的场景图，现有的场景图生成方法几乎都是通过“信息传递机制” (message passing)，让每个物体和视觉关系都能充分地考虑和融合周围的视觉元素。例如，物体“人”和物体“自行车”之间一个很常见的视觉关系就是“骑” (即“人 \rightarrow 骑 \rightarrow 自行车”)；同样，视觉关系“骑”也能够提升这两个物体 (“人” 和 “自行车”) 的类别预测。最后，这些方法都是直接利用物体和视觉关系分类的交叉熵之和作为模型最终的优化目标。然而，这个优化目标 (交叉熵之和) 将所有节点的重要性看成完全相同，即每个不同节点的分类损失对总的损失函数影响相同，将大大限制模型融合周围信息的能力。

在本章，我们提出一种全新的反事实多智能体学习方法 (Counterfactual critic Multi-Agent Training, CMAT)。CMT 是一种基于多智能体策略梯度优化方法。它通过将每个物体看成一个智能体，从而将整个图像场景图生成任务转换成一个多智能体协同决策任务。基于这个框架，CMAT 就可以直接利用整个场景图的生成质量作为优化目标 (即全局奖励函数)。另外，为了给每个智能体分配适当的奖励，我们设计来一个反事实基准模型 (counterfactual baseline)。这个反事实基准模型通过改变目标智能体的预测类别，固定其他所有智能体的预测，来推测当前智能体预测的局部贡献。通过在大规模图像场景图生成数据集 Visual Genome (VG) 上进行大量的对比实验，CMAT 在多个实验设定和评价指标下可以达到目前最好的性能。

4.1 问题描述

视觉场景理解是计算机视觉研究领域一个重要的研究领域。它不仅仅需要对场景中所有物体的类别以及位置进行预测，同时需要对两两物体之间的视觉关系进

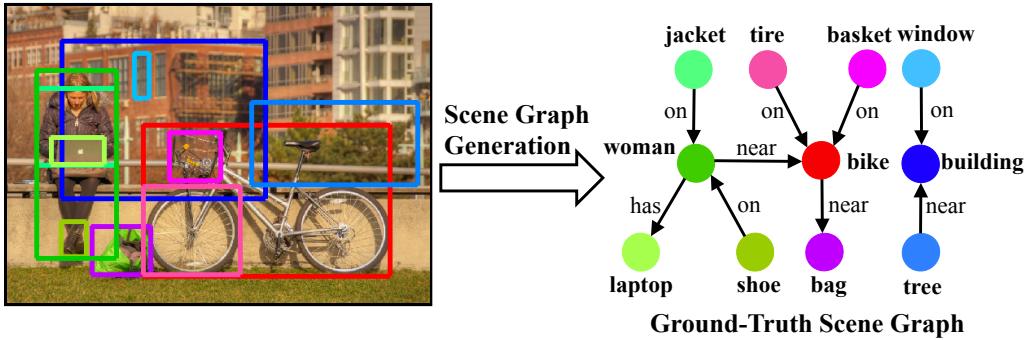


图 4-1 图像场景图生成任务示例

行预测。随着物体检测^[14,15]与分割技术^[17,199]的成熟，计算机已经可以准确地识别物体的类别、位置以及属性。然而，视觉场景理解不仅仅只是对单个物体的识别，还需要进一步对视觉关系进行识别。所有的物体和物体间视觉关系组合在一起，就构成了场景图^[21]。如图 4-1 所示，场景图中每个节点和边分别表示图像中的物体和对应物体间的视觉关系。与此同时，图像场景图通常作为一种结构化的视觉知识，辅助许多视觉场景理解任务，例如：图像描述生成^[200–202]、视觉问答^[203,204]和视觉推理^[205,206]等。

对于图像场景图生成任务（Scene Graph Generation, SGG），一个最直接的解决思路就是将场景图生成任务分解成物体分类和视觉关系分类两个独立的子任务，即先用一个物体检测器检测物体框，然后分别预测每个物体框的类别以及两两物体框之间视觉关系的类别^[73,75,77]。尽管这类方法的结构十分简单，但是它们都忽略了图像中所有元素之间的内在联系，即每个物体（视觉关系）周围的视觉元素往往会展提供一些归纳偏置^[207]（inductive bias）来辅助物体（视觉关系）的预测。如图 4-1 所示，物体“窗户”（window）和物体“建筑物”（building）常常会同时出现在同一张图像中，“在附近”（near）也是物体“树”（tree）和物体“建筑物”（building）之间最常见的视觉关系。因此，从视觉三元组“窗户 → 在上面 → ?”或“树 → 在附近 → ?”中，很容易推测出物体“?”是“建筑物”。这些归纳偏置带来的辅助信息已经被广泛地用来提升场景图生成性能^[76,79–82,86–88,90–93,208]。具体来说，这些方法都是通过借助条件随机场^[209]（Conditional Random Field, CRF）来建模所有节点和边的联合分布，然后通过信息传递机制来更新节点和边的特征^[210]。最后，整个模型利用所有节点（物体）和边（视觉关系）分类的交叉熵之和作为损失函数进行参数优化。

现有的图像场景图生成方法没有充分地利用场景中视觉元素之间的内在联系，一个重要的原因就是将物体和视觉关系分类的交叉熵之和作为场景图生成的优化目标。这个优化目标不具备**整体一致性**。所谓的“整体一致性”是指所有预测的物

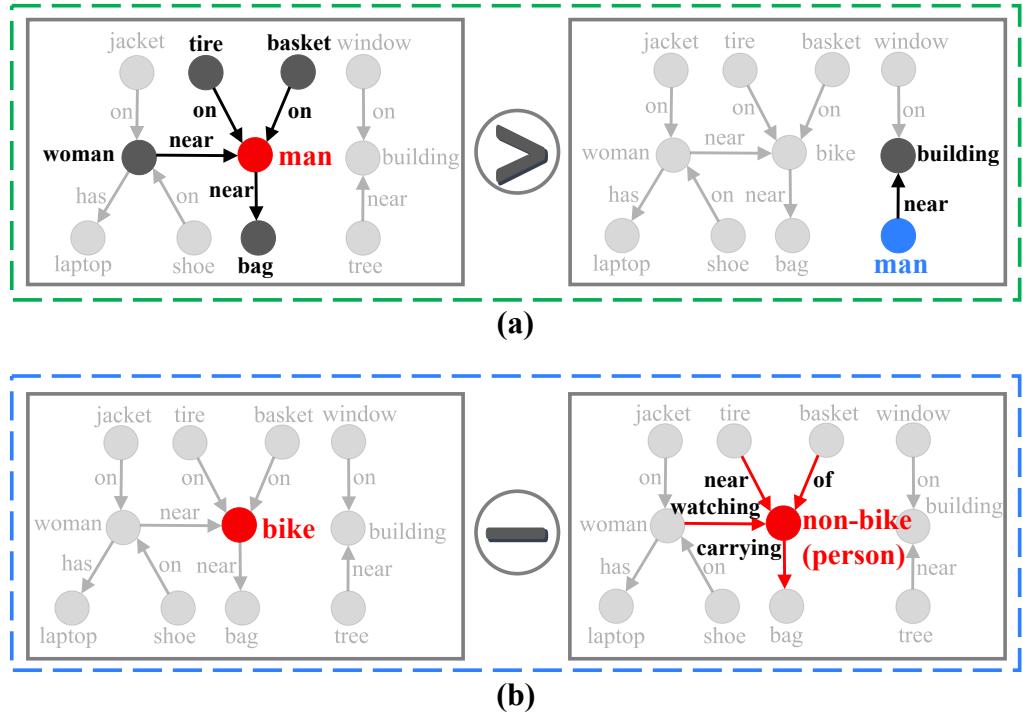


图 4-2 场景图生成中优化目标的整体一致性和局部敏感性

体类别和视觉关系类别之间应该保持整体一致。而交叉熵之和将所有的物体和视觉关系的预测看成是相互独立的。如图 4-2 (a) 所示，考虑两种极端情形，分别只有红色节点（“bike”）或蓝色节点（“tree”）被错误地分类成同一类别 “man”，而其他所有的物体和视觉关系都正确。对于这两种情况，根据交叉熵之和的损失函数，它们最终的损失大小是完全相同的。然而，因为红色节点连接的边远多于蓝色节点，即对红色节点的错误分类将影响更多的视觉关系。因此，错误分类红色节点相比于错误分类蓝色节点应该导致更大的损失。因此，我们提出直接使用 Recall@K^[73] 或者 SPICE^[211] 等图像场景图的全局评价指标作为优化目标。另一方面，场景图生成的优化目标还应该具备**局部敏感性**。所谓的“局部敏感性”是指优化目标应该能够感知每个节点类别的预测变化。由于场景图的全局评价指标是一个整体生成质量的评价数值，忽略了单个节点的预测贡献。因此我们需要设计一种机制，可以分解出每个节点各自的贡献，进而为每个局部预测计算更加有效的优化梯度。

在本章，我们提出了一种全新的场景图优化模型，可以同时满足优化目标的整体一致性和局部敏感性：反事实多智能体学习 (CMAT)。具体来说，我们将图像场景图生成任务转换成一种多智能体协同决策任务。其中将每个物体看成是一个智能体，每个智能体的动作空间是所有可选择的物体类别。每个智能体之间可以进行通信，来编码周围的视觉元素，提升智能体内部的特征表达。经过多轮智能体通信

之后，我们直接利用一个视觉关系预测模型来得到智能体之间的视觉关系，得到最终的场景图预测结果。通过与人工标注的场景图对比，得到一个全局的奖励。

为了优化目标的整体一致性，我们直接将整体场景图生成的评价指标（如：Recall@K 或 SPICE）作为全局奖励，并且使用策略梯度（policy gradient）的方法对参数进行学习^[212]。从多智能体强化学习^[213,214]（Multi-Agent Reinforcement Learning, MARL）的观点来看，尤其是“演员-评论家”方法^[214]（actor-critic），CMAT 中视觉关系预测模型可以看成是评论家（critic），而物体类别的分类模型可以看成是策略网络（policy network）。为了优化目标的局部敏感性，对于每个智能体，我们都从全局奖励中减去一个特定的反事实基准^[215]。这个反事实基准模型通过改变目标智能体的预测类别同时固定其他智能体的预测类别，来推测目标智能体预测的局部贡献。如图 4-2 (b) 所示，为了得到红色节点预测为“自行车”（bike）的贡献，我们可以固定其他节点的预测，而将红色节点的预测替换成其他的“非自行车”（non-bike）类，如“人”（person）等。进而通过计算出这种反事实的替换对整体的场景图生成效果带来多大的影响，来得到红色节点预测为“自行车”的贡献。

为了更好的编码物体周围的视觉元素信息和物体间的内在联系，我们还设计了一种更加有效的智能体通信（agent communication）模型。相比于现有的信息传递机制^[79-81,86-88]，我们不再将视觉关系也看成节点进行信息传递。通过这个设计，我们可以将智能体通信与视觉关系预测两个任务分离出来，让前者关注如何编码物体间内在联系，同时让后者作为评论家提供全局奖励来引导网络的优化。

我们在目前最大的图像场景图生成数据集 Visual Genome (VG) 上对 CMAT 的性能进行了验证。通过大量的对比实验，我们在通用的三种不同的实验设定下都可以达到目前最好的效果。

在本章，我们主要有三个贡献：

1. 我们提出了一种全新的图像场景图生成的优化方式：反事实多智能体学习(CMAT)。据我们了解，我们是第一次将图像场景图生成任务转化成一个多智能体协同决策问题，使得优化目标满足整体一致性要求。
2. 我们设计了一个全新的反事实基准模型，可以使多智能体策略梯度算法的优化目标同时具备局部敏感性。
3. 我们设计了一个有效的多智能体通信模型，有效地将智能体通信与视觉关系预测两个任务分离出来。

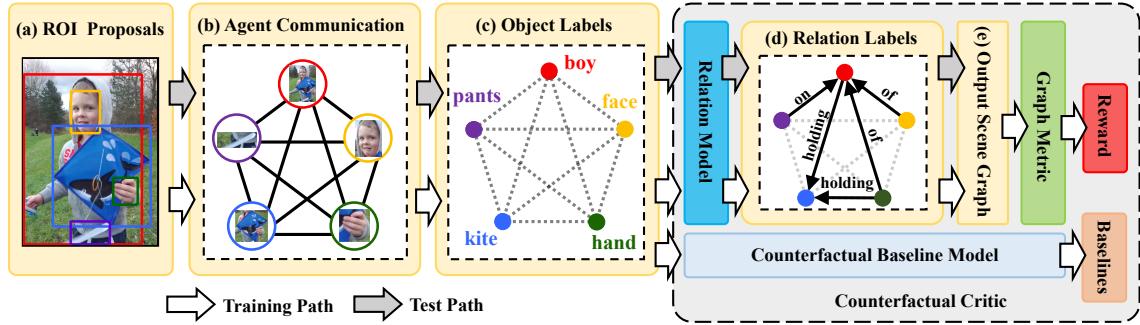


图 4-3 模型 CMAT 的总体流程图

4.2 反事实多智能体学习

给定一个物体类别集 \mathcal{C} (包括背景) 和一个视觉关系类别集 \mathcal{R} (包括没有视觉关系)，图像场景图可以表示成： $\mathcal{G} = \{\mathcal{V} = \{(v_i, l_i)\}, \mathcal{E} = \{r_{ij}\} | i, j = 1 \dots n\}$ ，其中 \mathcal{V} 和 \mathcal{E} 分别表示所有节点 (物体) 和边 (视觉关系) 的结合。 $v_i \in \mathcal{C}$ 表示第 i 个节点的物体类别， $l_i \in \mathbb{R}^4$ 表示第 i 个节点的物体位置， $r_{ij} \in \mathcal{R}$ 表示第 i 个节点和第 j 个节点之间的视觉关系。图像场景图生成任务就是检测出图像中所有的物体以及物体间的视觉关系。

在本节，我们先介绍模型 CMAT 中的每个组成部分。然后，我们再介绍模型 CMAT 的优化目标。

4.2.1 多智能体协同决策

物体候选框检测：我们首先使用预训练好的目标检测器 Faster R-CNN^[14] 对输入图像进行物体检测，得到一系列候选框。对于每个候选框，同时可以得到其位置坐标 l_i ，特征向量 x_i^0 ，以及初始的物体类别预测概率分布 s_i^0 。上角标 0 表示 T 轮智能体通信的初始输入。我们参考之前的场景图生成工作^[80,82]，固定所有候选框的位置 $\{l_i\}$ 作为物体坐标的最终预测结果。为了后续表达的简洁性，我们在后续内容中省略位置坐标 l_i 。

智能体通信：给定 n 个物体候选框，我们将每个候选框看成是一个智能体。智能体之间将通过 T 轮的通信来编码各自周围的视觉元素和物体间的内在联系。如图 4-4 所示，在单步智能体通信过程中，共有三个模块参与：信息提取模块 (extract module)、信息合成模块 (message module) 和状态更新模块 (update module)。为了减小整个 CMAT 模型的参数量，这三个模块在所有的智能体之间都共享参数。关于这三个模块的具体细节如下：

(a) 信息提取模块：我们使用递归神经网络 LSTM^[116] 作为信息提取模块。LSTM

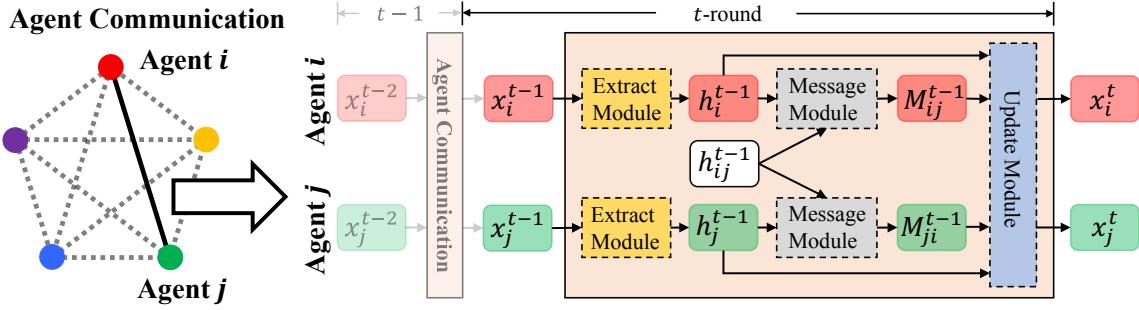


图 4-4 单步智能体通信示意图

不仅可以编码智能体之间的交互历史，同时也可以提取智能体自身的内部状态。具体来说，对于第 i 个智能体，在第 t ($0 < t \leq T$) 轮通信时：

$$\begin{aligned} \mathbf{h}_i^t &= \text{LSTM}(\mathbf{h}_i^{t-1}, [\mathbf{x}_i^t, \mathbf{e}_i^{t-1}]), \\ \mathbf{s}_i^t &= \mathbf{s}_i^{t-1} + \mathbf{W}_h \mathbf{h}_i^t, \\ v_i^t &\sim \mathbf{p}_i^t = \text{softmax}(\mathbf{s}_i^t), \\ \mathbf{e}_i^t &= \sum_{\tilde{v}} \mathbf{p}_i^t(\tilde{v}) \mathbf{E}[\tilde{v}], \end{aligned} \quad (4-1)$$

其中 $\mathbf{h}_i^t \in \mathbb{R}^h$ 是 LSTM 的隐含状态向量， $\mathbf{x}_i^t \in \mathbb{R}^d$ 是每个 LSTM 时刻的输入特征， $\mathbf{s}_i^t \in \mathbb{R}^{|\mathcal{C}|}$ 是预测的物体类别概率分布。初始的（即第 0 步通信时）输入特征和类别预测概率都来自于物体候选框检测。 $\mathbf{E}[\tilde{v}] \in \mathbb{R}^e$ 是类别标签 $\tilde{v} \in \mathcal{C}$ 的特征编码向量，以及 $\mathbf{e}_i^t \in \mathbb{R}^e$ 是一个基于类别概率 \mathbf{p}_i^t 加权的类别标签编码向量， $\mathbf{W}_h \in \mathbb{R}^{h \times |\mathcal{C}|}$ 是一个可学习的映射矩阵，以及 $[,]$ 是向量间的连接操作。所有的隐藏状态 $\{\mathbf{h}_i^t\}$ 都输入到之后的信息合成模块用于合成通信信息。

(b) 信息合成模块：对于第 i 个智能体和第 j 个智能体之间的通信，信息合成模块将分别合成信息 M_{ij}^t 和 M_{ji}^t 。具体来说，对于第 i 个智能体收到的信息 $M_{ij}^t = (\mathbf{m}_j^t, \mathbf{m}_{ij}^t)$ ，主要包含两部分：

$$\mathbf{m}_j^t = \mathbf{W}_u \mathbf{h}_j^t, \quad \mathbf{m}_{ij}^t = \mathbf{W}_p \mathbf{h}_{ij}^t \quad (4-2)$$

其中 $\mathbf{m}_j^t \in \mathbb{R}^h$ 表示一元信息，用来表征第 j 个智能体本身的属性（如：单个物体的局部视觉特征）， $\mathbf{m}_{ij}^t \in \mathbb{R}^h$ 表示二元信息，用来表征两个智能体之间的交互信息（如：两个智能体的相对位置信息）。 $\mathbf{h}_{ij}^t \in \mathbb{R}^d$ 表示第 i 个智能体与第 j 个智能体之间的共同特征，它的初始化是两个智能体物体框合并之后的视觉特征。对于第 i 个智能体，所有来自其他智能体的通信信息 $\{M_{i*}^t\}$ 和其内部状态 \mathbf{h}_i^t 都输入到信息更新模块，更新其内部状态。

(c) 状态更新模块：在每轮智能体通信过程中，对于每个智能体，我们使用注意力机制^[148]来融合不同智能体的通信信息：

$$\begin{aligned}
 u_j^t &= \mathbf{w}_u[\mathbf{h}_i^t, \mathbf{h}_j^t], \\
 u_{ij}^t &= \mathbf{w}_p[\mathbf{h}_i^t, \mathbf{h}_{ij}^t], \\
 \alpha_j^t &= \exp(u_j^t) / \sum_k \exp(u_k^t), \\
 \alpha_{ij}^t &= \exp(u_{ij}^t) / \sum_k \exp(u_{ik}^t), \\
 \mathbf{x}_i^{t+1} &= \mathbf{W}_x(\text{ReLU}(\mathbf{h}_i^t + \sum_j \alpha_j^t \mathbf{m}_j^t + \sum_j \alpha_{ij}^t \mathbf{m}_{ij}^t)) \\
 \mathbf{h}_{ij}^{t+1} &= \text{ReLU}(\mathbf{h}_{ij}^t + \mathbf{W}_s \mathbf{h}_i^{t+1} + \mathbf{W}_e \mathbf{h}_j^{t+1})
 \end{aligned} \tag{4-3}$$

其中 α_j^t 和 α_{ij}^t 是不同信息融合的权重， $\mathbf{w}_u \in \mathbb{R}^{2h}$ 、 $\mathbf{w}_p \in \mathbb{R}^{h+d}$ 、 $\mathbf{W}_x \in \mathbb{R}^{h \times d}$ 、 $\mathbf{W}_s \in \mathbb{R}^{h \times d}$ 和 $\mathbf{W}_e \in \mathbb{R}^{h \times d}$ 这些都是需要学习的映射矩阵。

视觉关系预测：在 T 轮智能体通信之后，所有的智能体都完成了状态更新。在测试阶段，对于所有的智能体，我们直接根据预测的分数 \mathbf{s}_i^T 来选取所有智能体的物体类别 v_i^T 。之后，我们再利用视觉关系预测模型对任意两个智能体之间进行视觉关系分类：

$$\begin{aligned}
 \mathbf{z}_i &= \mathbf{W}_o[\mathbf{h}_i^T, \mathbf{E}[v_i^T]], \\
 \mathbf{z}_j &= \mathbf{W}_o[\mathbf{h}_j^T, \mathbf{E}[v_j^T]], \\
 \mathbf{p}_{ij} &= \text{softmax}([\mathbf{z}_i, \mathbf{z}_j] \odot \mathbf{W}_r \mathbf{z}_{ij} + \mathbf{w}_{v_i^T, v_j^T}), \\
 r_{ij} &= \arg \max_{r \in \mathcal{R}} \mathbf{p}_{ij}(r),
 \end{aligned} \tag{4-4}$$

其中 $\mathbf{W}_o \in \mathbb{R}^{(h+e) \times z}$ 、 $\mathbf{W}_r \in \mathbb{R}^{z \times 2z}$ 都是需要学习的映射矩阵， $\mathbf{z}_{ij} \in \mathbb{R}^z$ 是第 i 个智能体与第 j 个智能体之间预测的视觉关系特征， \odot 是特征融合函数^[216]： $\mathbf{x} \odot \mathbf{y} = \text{ReLU}(\mathbf{W}_x \mathbf{x} + \mathbf{W}_y \mathbf{y}) - (\mathbf{W}_x \mathbf{x} - \mathbf{W}_y \mathbf{y})^2$ ， $\mathbf{w}_{v_i^T, v_j^T} \in \mathbb{R}^{|\mathcal{C}|}$ 是基于 VG 数据集统计的视觉关系类别偏置^[82]。

4.2.2 反事实多智能体学习

在本节，我们将详细介绍 CMAT 中优化目标的细节，具体包括：(1) 符合整体一致性的多智能体策略梯度算法；(2) 符合局部敏感性的反事实评论家模型。

优化目标的整体一致性：目前，几乎所有的场景图生成算法都是将物体和视觉关系分类的交叉熵之和作为模型的优化目标。对于一个预测的场景图 $(\hat{\mathcal{V}}, \hat{\mathcal{E}})$ ，如果人工标注的场景图为 $(\mathcal{V}^{gt}, \mathcal{E}^{gt})$ ，根据交叉熵之和（cross-entropy, XE）的优化目标，整个模型的损失函数则为：

$$L(\theta) = \sum_{ij} (\text{XE}(\hat{v}_i, v_i^{gt}) + \text{XE}(\hat{r}_{ij}, r_{ij}^{gt})). \tag{4-5}$$

由公式(4-5)可以看出，交叉熵之和的优化目标本质上将所有的节点预测都看成相互独立的。

为了解决上述问题，我们提出将交叉熵之和的优化目标替换成以下两种具备整体一致性的优化目标：(1) **Recall@K**^[73] 在预测分数最高的前 K 个视觉三元组（“主语 → 谓语 → 宾语”）中，预测正确的视觉三元组占所有标注视觉三元组的百分比。(2) **SPICE**^[211]：所有视觉三元组预测的准确率和召回率之间的 F 值。与交叉熵之和不同的是，Recall@K 和 SPICE 都是不可导的。因此，CMAT 模型借助多智能体策略梯度算法对模型参数进行优化。

多智能体策略梯度：我们首先定义模型 CMAT 中智能体的动作 (action)、策略函数 (policy) 和状态 (state)。然后，我们推导出模型参数的梯度计算公式。

(a) 动作：每个智能体的动作空间是所有可选择的物体类别的总和，即第 i 的智能体的动作是 v_i^t 。我们用集合 $V^t = \{v_i^t\}$ 来表示所有智能体的动作集合。

(b) 状态：我们参考 Hausknecht 等人^[217] 使用递归神经网络 LSTM（信息提取模块）来编码智能体与环境之间的交互历史。LSTM 的隐含状态 h_i^t 可以看成是第 i 个智能体对局部可见环境状态的近似。我们用集合 $H^t = \{h_i^t\}$ 来表示所有智能体的状态集合。

(c) 策略函数：每个智能体的策略函数就是物体类别分类模型。在训练阶段，每个智能体通过对分类概率进行采样得到智能体类别，即： $\mathbf{p}_i^T = \text{softmax}(\mathbf{s}_i^T)$ 。因为 CMAT 只在 T 轮智能体通信之后进行动作采样，根据策略梯度的理论^[212]，CMAT 中梯度计算公式为：

$$\nabla_{\theta} J \approx \sum_{i=1}^n \nabla_{\theta} \log \mathbf{p}_i^T(v_i^T | h_i^T; \theta) Q(H^T, V^T), \quad (4-6)$$

其中， $Q(H^T, V^T)$ 为状态-动作函数 (state-action value function)。不同于现有的演员-评论家^[214,218,219] (actor-critic) 算法通常用一个独立的网络来近似 Q ，在 CMAT 中，我们参考 Rao 等人^[220] 直接使用实际的奖励来代替 Q 。这样做的原因主要有两个：(1) 在图像场景图生成任务中，智能体的数量（如：SGDet 中通常检测 64 个物体框）和动作空间的采样大小（如：VG 数据集中有 150 个物体类别）都明显大于现有的多智能体策略梯度的工作。这容易导致训练样本不充足，难以训练一个准确的状态-动作函数。(2) 直接使用实际奖励可以大大减小模型的复杂度，提升训练速度。因此，CMAT 中梯度计算公式变为：

$$\nabla_{\theta} J \approx \sum_{i=1}^n \nabla_{\theta} \log \mathbf{p}_i^t(v_i^T | h_i^T; \theta) R(H^T, V^T), \quad (4-7)$$

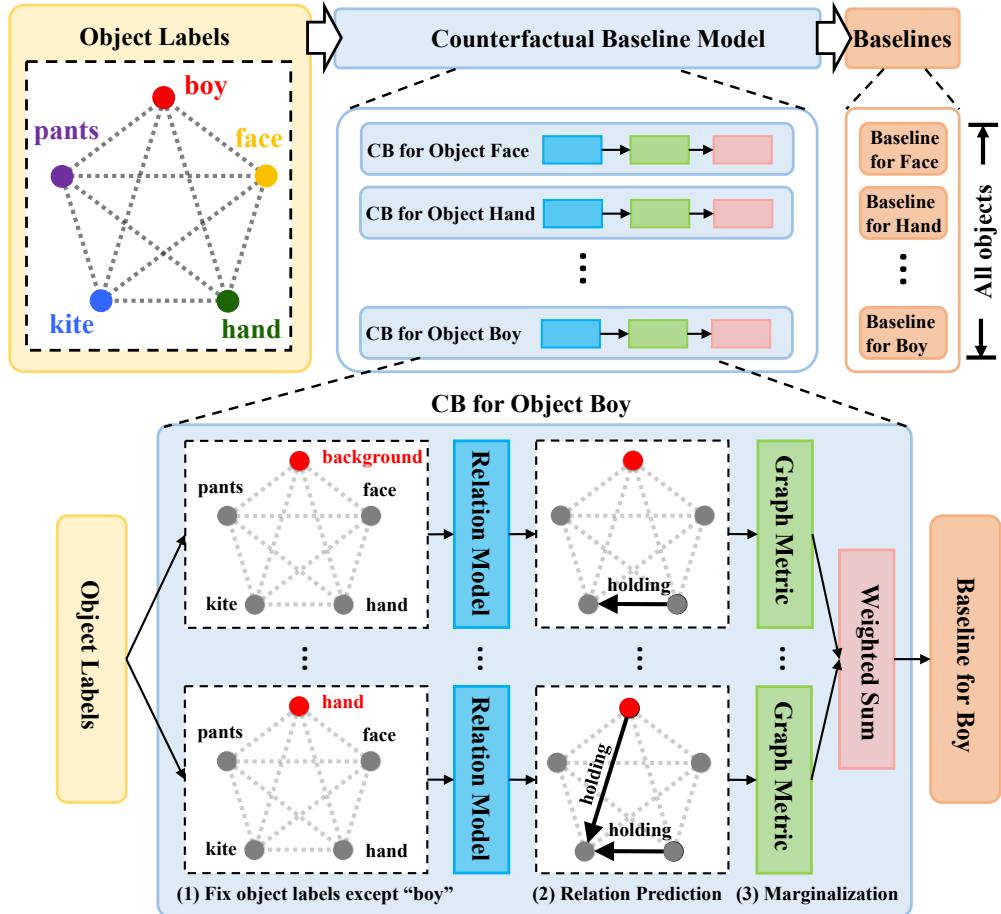


图 4-5 CMAT 中反事实基准模型

其中， $R(H^T, V^T)$ 是实际的全局奖励（即：Recall@K 或 SPICE）。另外，值得注意的是， $R(H^T, V^T)$ 里包含一个可以学习的视觉关系分类模型。

优化目标的局部敏感性：从上述公式(4-7)可以看出，全局的奖励是综合考虑了所有智能体预测类别的总贡献，即对每个单独的智能体而言，总贡献都是完全相同的。在图 4-6 中，我们举一个简单的例子来说明这种相同的总贡献对场景图生成任务的副作用。如图 4-6 所示，绿色和红色分别表示正确的预测和错误的预测（节点和边），假定全局奖励定义为预测正确的视觉三元组数减去预测错误的视觉三元组数，且在 (1) (2) 两

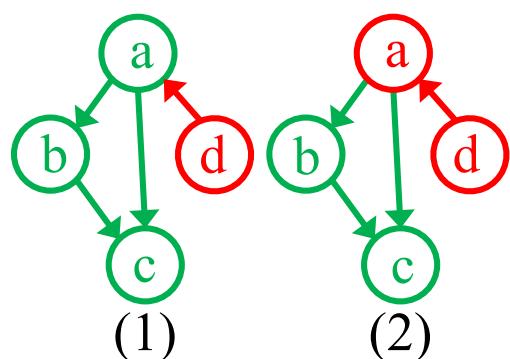


图 4-6 优化目标局部敏感性的重要性示例图

个场景图中只有节点“a”预测不同，其他预测都完全相同。

根据公式(4-7)，场景图(1)中所有的节点都得到一个正的全局奖励($3-1=+2$)，而场景图(2)中所有的节点都得到一个负的全局奖励($1-3=-2$)。在这两种情况下，虽然节点“b”、“c”、“d”的预测完全相同，但是对于它们计算得到的梯度方向却完全相反，容易造成模型多次无效的优化迭代。因此，通过让优化目标满足局部敏感性，即通过计算每个智能体各自的局部奖励，有利于提供更有效的优化信号，提升训练效率。

反事实评论家：一个直观地计算某个智能体动作的局部奖励的方法就是将目标智能体的动作替换成其他动作，然后利用总体贡献的变化来近似。因此， $R(H^T, V^T) - R(H^T, (V_{-i}^T, \tilde{v}_i^T))$ 可以表示第 i 个智能体的动作 v_i^T 的局部贡献，其中 V_{-i}^T 表示所有除第 i 个智能体以外其他所有智能体都保持初始的动作，而第 i 个智能体采取新的动作 \tilde{v}_i^T 。由于新的动作 \tilde{v}_i^T 有 $|\mathcal{C}|$ 种可能性，为了准确地计算其他所有智能体（即： V_{-i}^T ）的贡献，我们对所有可能的动作进行平均： $\text{CB}^i(H^T, V^T) = \sum \mathbf{p}_i^T(\tilde{v}_i^T)R(H^T, (V_{-i}^T, \tilde{v}_i^T))$ ，其中 $\text{CB}^i(H^T, V^T)$ 称为第 i 个智能体的**反事实基准**。这个反事实基准表示的是无论第 i 个智能体采用的动作，而其他所有智能体采用默认动作时模型能够得到的平均全局奖励。**CMAT** 的反事实基准模型展示在图 4-5 中。

给定一个全局的奖励 $R(H^T, V^T)$ 和第 i 个智能体动作的反事实基准 $\text{CB}^i(H^T, V^T)$ ，我们可以得到第 i 个智能体动作的局部贡献为：

$$A^i(H^T, V^T) = R(H^T, V^T) - \text{CB}^i(H^T, V^T) \quad (4-8)$$

其中， $A^i(H^T, V^T)$ 在演员-评论家算法[221,222]中常被称为“优势”(advantage)， $\text{CB}^i(H^T, V^T)$ 在策略梯度算法中常被称为“基准”(baseline)，用来减少梯度计算时的方差。整个计算 $A^i(H^T, V^T)$ 的网络结构可以合称为“反事实评论家”(counterfactual critic)。因此，根据公式(4-7)，CMAT 的梯度计算公式变为：

$$\nabla_{\theta} J \approx \sum_{i=1}^n \nabla_{\theta} \log \mathbf{p}_i^T(v_i^T | h_i^T; \theta) A^i(H^T, V^T) \quad (4-9)$$

最后，我们在加上交叉熵之和损失一起训练，最终参数训练的梯度为：

$$\begin{aligned} \nabla_{\theta} J \approx & \underbrace{\sum_{i=1}^n \nabla_{\theta} \log \mathbf{p}_i^T(v_i^T | h_i^T; \theta) A^i(H^T, V^T)}_{\text{CMAT}} + \\ & \underbrace{\alpha \sum_{i=1}^n \sum_{j=1}^n \nabla_{\theta} \log \mathbf{p}_{ij}(r_{ij})}_{\text{视觉关系交叉熵之和}} + \underbrace{\beta \sum_{i=1}^n \nabla_{\theta} \log \mathbf{p}_i^T(v_i^T)}_{\text{物体类别交叉熵之和}}, \end{aligned} \quad (4-10)$$

其中， α 和 β 是不同损失函数之间权衡的权重，额外的交叉熵之和是为了增加模型训练的稳定性^[220]。同样，我们还增加了一个熵正则项^[100,118]来约束 $\{\mathbf{p}_i^T\}_i$ 。

4.3 实验设置与性能对比

4.3.1 图像场景图生成数据集与实验设定

图像场景图生成数据集: 我们使用目前最大的图像场景图数据集 Visual Genome (VG)^[3] 对模型性能进行评估。为了能够与现有的工作公平地进行比较，我们采用与它们相同的数据集划分和预处理^[80,82,89,223,224]。处理后的数据集共包含 150 个物体类别和 50 个视觉关系类别。每张图像平均包含 11.5 个物体和 6.2 个视觉关系。在整个数据集中，70% 的图像作为训练集，剩余 30% 的图像作为测试集。

实验设定: 参考现有的图像场景图生成文献^[80,82,88]，我们在三种实验设定下评估场景图生成质量：

1. **视觉关系分类 (PredCls):** 给定图像中所有的物体框位置和物体类别，模型只需要预测两两物体间的视觉关系；
2. **场景图分类 (SGCls):** 给定图像中所有的物体框位置，模型需要预测所有物体的类别以及两两物体间的视觉关系；
3. **场景图检测 (SGDet):** 给定图像，模型需要检测物体框位置、预测物体类别以及两两物体间的视觉关系。

一个视觉三元组预测正确，不仅需要主语 (subject)、谓语 (predicate) 和宾语 (object) 的类别都预测正确，同时需要主语和宾语的物体框与真实准确物体框的交并比 (IoU) 均大于 0.5。按照图像场景图生成任务惯例，我们使用 Recall@20 (R@20)、Recall (R@50) 和 Recall (R@100) 作为场景图生成质量的评价指标。

4.3.2 实验细节

物体检测器: 为了公平地与现有的工作进行对比，我们采用了与^[82]相同的物体检测器。具体来说，它是以 VGG 网络^[9]为主干网络，然后锚框的大小和长宽比与 YOLO-9000^[225] 设置一样，然后用 RoIAlign^[17] 代替 RoIPooling^[226,227]。

训练细节: 我们参照之前的策略梯度的工作，将整个训练过程分成两个阶段，并且先使用监督训练对模型进行参数初始化。在监督训练过程中，我们将 RoIAlign 层之前的参数都固定住，然后使用物体分类和视觉关系分类的交叉熵之和作为优化

目标。批处理的大小和初始的学习率分别设为 6 和 10^{-3} 。在策略梯度优化过程中，初始的学习率设为 3×5^{-5} 。对于场景图检测任务，因为所有可能的物体组合非常多（例如：64 个物体就存在大约 4000 中组合），我们参照 Zellers 等人^[82] 只考虑当物体有重叠时的视觉关系，这样可以将每张图预测的视觉关系数减少到 1000 左右。

速度与正确率的权衡：在策略梯度的训练过程中，完整的反事实评论家的计算需要对所有可能的物体类别进行平均加权，通常需要非常多的时间（如：对于 64 个智能体，每个智能体共有 151 种物体类别选择，则需要超过 9600 次 ($\approx 151 \times 64$) 的评估计算评价指标 Recall@K）。幸运的是，初始的目标检测器可以对物体类别有个初步的预测概率，而只有极少数的类别才有较大的预测概率，绝大多数类别的预测概率都趋近于 0。为了速度与正确率之间的权衡，我们只对背景 (background) 和预测概率最高的两种类别进行平均求和来近似对所有的类别的期望。在我们的实验中，这样的实验简化可以减少 70 倍的评估计算，同时维持几乎相同的实验效果。

SGDet 的后处理：对于场景图检测任务 (SGDet)，为了与之前的工作^[82,228] 公平地进行对比，我们采用相同的后处理操作。具体来说，在对每个 RoI 预测出物体所有类别的概率分布之后，我们对每个类别使用一次非极大值抑制来确定最终的物体类别，以及选择对应类别的位移偏置。在我们的实验中，非极大值抑制的 IoU 阈值设置为 0.5。

4.3.3 场景图生成性能分析

在本节，我们通过大量的对比实验来分析 CMAT 模型中的不同设计选择对总体性能的影响，包括全局奖励函数的选择、基准模型的选择、以及多智能体通信步数的选择等。

全局奖励函数的选择：为了验证不同的全局奖励对最终场景图生成性能的影响，我们对比了两种全局奖励函数：Recall@K 和 SPICE。在这个实验中，我们使用预测前 20 个视觉三元组来计算对应的 Recall@K 和 SPICE 值。实验结果展示在表 4-1 中，其中 XE 表示以交叉熵之和为损失函数的预训练结果。从表 4-1 可以看出，全局奖励函数 Recall@K 和 SPICE 都能在交叉熵之和为优化目标的预训练基础上进一步提升场景图生成性能，这主要是因为将全局奖励函数作为场景图生成优化目标满足整体一致性。另外，使用 Recall@K 可以得到比 SPICE 稍微好一点的结果，可能的原因是因为目前的场景图生成数据集都是不完全标注的，而评价指标 SPICE 不适合用来评估这类数据集。因此，在后续的实验中，我们都使用 Recall@K 作为全局奖励函数。

		XE	R@20	SPICE
SGCls	R@20	34.08	35.93	35.27
	SPICE	15.39	16.01	15.90
SGDet	R@20	16.23	16.53	16.51
	SPICE	7.48	7.66	7.64

表 4-1 不同全局奖励函数的选择对性能的影响

		XE	MA	SC	CF
SGCls	R@20	34.08	34.76	34.68	35.93
	R@50	36.90	37.58	37.54	39.00
	R@100	37.61	38.29	38.25	39.75
SGDet	R@20	16.23	16.07	16.37	16.53
	R@50	20.62	20.41	20.82	20.95
	R@100	23.24	23.02	23.41	23.62

表 4-2 不同基准模型对性能的影响

		2-step	3-step	4-step	5-step
SGCls	R@20	35.09	35.25	35.40	35.93
	R@50	37.95	38.19	38.37	39.00
	R@100	38.67	38.91	39.09	39.75
SGDet	R@20	16.35	16.43	16.47	16.53
	R@50	20.89	20.88	20.92	20.95
	R@100	23.49	23.50	23.54	23.62

表 4-3 不同多智能体通信步数对性能的影响

基准模型的选择：为了验证不同的基准模型对最终场景图生成性能的影响，我们将本章提出的反事实基准（CounterFactual baseline, CF）与其他两种流行的基准模型进行对比：“移动平均”（Moving Average, MA）^[229] 和 “自评论”（Self-Critical, SC）^[97]。MA 是一个对总体奖励进行动态平均得到的常数^[100,118]，而 SC 是对所有

的动作都采取贪婪选择时得到的全局奖励。实验结果展示在表 4-2 中，其中 XE 表示以交叉熵之和为损失函数的预训练结果。从表 4-2 可以看出，反事实基准 CF 可以在交叉熵之和作为优化目标的预训练基础上显著提升实验性能；而 MA 和 SC 只能提升细微的实验性能。这主要是因为反事实基准符合局部敏感性，可以对所有的智能体提供更加有效的训练梯度；而 MA、SC 都只能提供全局的奖励，不具备局部敏感性。

多智能体通信步数的选择：为了验证不同的多智能体通信步数对最终场景图生成性能的影响，我们将通信步数分别从 2 变化到 5。从表 4-3 可以看出，随着通信步数的增加，模型的性能可以持续提升，同时会造成计算资源和训练时间的增加。由于 GPU 的限制，我们将最大步数设为 5。通过与现有的信息传递机制相比，我们的多智能体通信模型可以避免过早饱和的问题^[76,80]。这主要的原因是我们没有将视觉关系也看成节点进行信息传递。

4.3.4 场景图生成性能对比

在本节，我们将本章提出 CMAT 模型与目前最先进的图像场景图生成算法进行对比。这些方法主要可以分为两大类：(1) **VRD**^[73]、**AED**^[223]、**FREQ**^[82]。这类方法都是将图像场景图生成任务分解成物体分类和视觉关系分类两个独立的子任务。(2) **MSDN**^[86]、**IMP**^[80]、**TFR**^[88]、**MOTIFS**^[82]、**G-RCNN**^[89]、**GPI**^[224]、**KER**^[230]。这类方法都是利用信息传递机制来编码每个物体周围的视觉元素和物体间的内在联系。特别地，所有的这些方法都是利用物体和视觉关系分类的交叉熵之和作为模型的优化目标。

定量性能分析：表 4-4 展示了不同场景图生成方法在 VG 数据集上的实验结果，其中图限制 (Graph Constraint)^[82] 表示当主语和宾语确定时，两物体间只能存在一种视觉关系；而没有图限制 (No Constraint) 表示当主语和宾语确定时，两物体间可以存在多种视觉关系。由表 4-4 可以看出，CMAT 在所有的评估指标下都达到了最好的性能。尤其值得注意的是，CMAT 在场景图分类 (SGCls) 任务中可以显著提升实验效果 (即：在有图限制和没有图限制的条件下可以分别提升 3.4% 和 4.3%)。这也刚好符合我们的设计动机，通过将预测物体类别看成智能体的动作选择，较好地提升物体的类别预测准确率。另一方面，实验结果也表明反事实多智能体学习可以显著地提升场景图生成任务的性能。对于视觉关系分类任务 (PredCls)，即使我们使用的视觉关系分类模型非常简单，我们仍然可以达到最好的实验性能。这说明在 CMAT 模型中，视觉关系分类模型的输入已经更好地编码了智能体的内部状态。

Model	SGDet	SGCls			PredCls					
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Graph Constraint	VRD	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0
	IMP	-	3.4	4.2	-	21.7	24.4	-	44.8	53.0
	MSDN	-	7.0	9.1	-	27.6	29.9	-	53.2	57.9
	AED	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4
	FREQ+	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
	IMP+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
	TFR	3.4	4.8	6.0	19.6	24.3	26.6	40.1	51.9	58.3
	MOTIFS	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
	G-RCNN	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
	GPI	-	-	-	-	36.5	38.8	-	65.1	66.9
	KER	-	27.1	29.8	-	36.7	37.4	-	65.8	67.6
	CMAT	22.1	27.9	31.2	35.9	39.0	39.8	60.2	66.4	68.1
No Constraint	AED	-	9.7	11.3	-	26.5	30.0	-	68.0	75.2
	IMP+	-	22.0	27.4	-	43.4	47.2	-	75.2	83.6
	FREQ+	-	28.6	34.4	-	39.0	43.4	-	75.7	82.9
	MOTIFS	22.8	30.5	35.8	37.6	44.5	47.7	66.6	81.1	88.3
	KER	-	30.9	35.8	-	45.9	49.0	-	81.9	88.9
	CMAT	23.7	31.6	36.8	41.0	48.6	52.0	68.9	83.2	90.1

表 4-4 不同场景图生成方法在 VG 数据集上的性能对比

另外，CMAT 模型可以兼容任何效果更好的视觉关系分类模型。对于场景图检测任务 (SGDet)，CMAT 的提升没有场景图分类任务明显。我们猜测可能的原因来自于物体检测框的准确率还不是特别高，导致部分智能体本身是背景信息。

定性性能对比：图 4-7展示了模型 CMAT 和模型 MOTIFS 在数据集 VG 上的场景图生成结果。其中绿色框表示与真实物体框交叉比大于 0.5 的物体框，蓝色框表示模型的检测框但数据集中没有标注，红色框表示遗漏的真实物体框。绿色的边表示真阳性 (true positive) 的视觉关系预测，红色边表示假阴性 (false negative) 的视觉关系预测，以及蓝色边表示假阳性 (false positive) 的视觉关系预测。

由图 4-7前两排结果看出可以看出，CMAT 模型很少遗漏一些重要的物体节点，如“laptop”、“surfboard”等。这主要原因是 CMAT 的优化目标满足整体一致性，往往对重要节点的重要性考虑更多。从第三排结果可以看出，CMAT 的错误主要是检测出比 MOTIFS 更多的未标注的视觉关系 (蓝色边)。由于目前使用的评价指标主

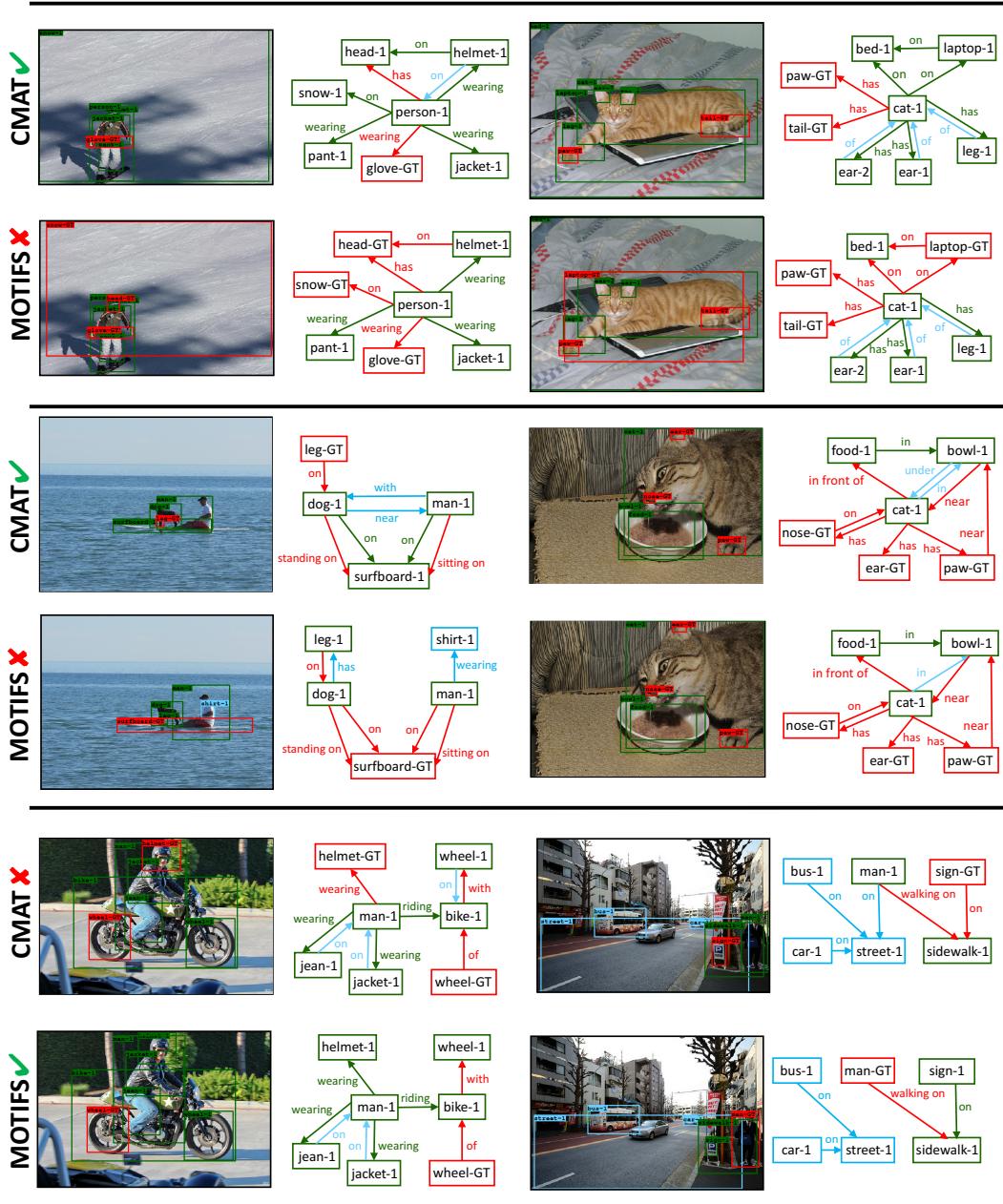


图 4-7 模型 CMAT 和模型 MOTIFS 在数据集 VG 上的场景图生成结果对比

要是基于 Recall@K，它只依据所有标注的视觉三元组的排序结果。因此，如果检测出更多的未标注的正样本，反而得到更低的评价分数。

4.4 本章小结

在本章，我们提出了图像场景图生成模型的优化目标应该同时具备整体一致性和局部敏感性。现有的场景图生成模型基本都是使用物体和视觉关系分类的交叉

熵之和作为模型的优化目标，缺乏整体一致性。为了解决这一问题，本章提出全新的反事实多智能体学习模型：CMAT。CMAT 首次将图像场景图生成任务转化成一个多智能体协同合作的决策任务。然后使用场景图生成的评价指标（如 Recall@K）作为模型的优化目标，满足整体一致性要求。其次，CMAT 中包含一个反事实基准模型，通过固定其他智能体的预测同时改变目标智能体的预测，来近似目标智能体局部贡献，进而将每个智能体的局部贡献分离出来，得到更加有效的训练信号。在大规模图像场景图生成数据集 VG 上也证明，CMAT 可以通过提升物体类别的预测，显著提升场景图生成质量。

5 基于多层空间和通道注意力网络的图像描述生成方法

注意力机制已经被广泛地运用在视觉场景理解任务中，如图像描述生成、视觉问答等。现有的注意力机制都是属于空间注意力机制，即只对卷积神经网络的最后一个特征图（feature map）在空间维度上进行加权。然而，卷积神经网络的特征图除了空间维度以外，还有通道和层级两个维度。因此，目前的注意力机制并没有充分利用卷积神经网络特征图的特性。在本章，我们提出一种全新的注意力机制网络：多层次空间和通道注意力网络（Spatial and Channel-wise Attention in CNN, SCA-CNN）。对于图像描述生成任务，SCA-CNN 在生成每个单词的过程中，动态地对不同层级下的特征图中所有的空间位置和通道进行加权，提升图像编码网络的表达能力。我们在图像描述生成的三个标准数据集（Flickr8K、Flickr30K、MSCOCO）对模型 SCA-CNN 进行评估，大量的对比实验结果都表明我们提出的多层次空间和通道注意力网络（SCA-CNN）可以显著地提升图像描述生成质量。

5.1 问题描述

注意力机制已经被广泛地证明可以用来提升视觉场景理解任务的性能，如图像或视频的描述生成^[118,231]、视觉问答^[120,121,152]等。在图像描述生成任务中，注意力机制的使用主要是基于一个合理的设想：人类在生成图像描述过程中，往往不是一次性直接记住整个图像，而是在生成语句的过程中不断地去调整关注的图像区域^[232]。具体来说，与之前的图像描述生成工作直接将整个图像编码成一个固定的向量表达不同^[22,112]，注意力机制让模型在语句生成的过程中不断地调整图像的特征表达，从而生成更加丰富和准确的描述语句。因此，注意力机制也可以被看成是一种动态的特征调节机制^[233,234]。

目前主要的图像视觉特征都是通过卷积神经网络进行编码^[9,11]。给定一个大小为 $W \times H \times 3$ 的彩色图像，通常卷积层用一个通道数为 C 的卷积核对输入图像进行卷积，得到一个大小为 $W' \times H' \times C$ 的特征图，然后这个特征图又输入到后续

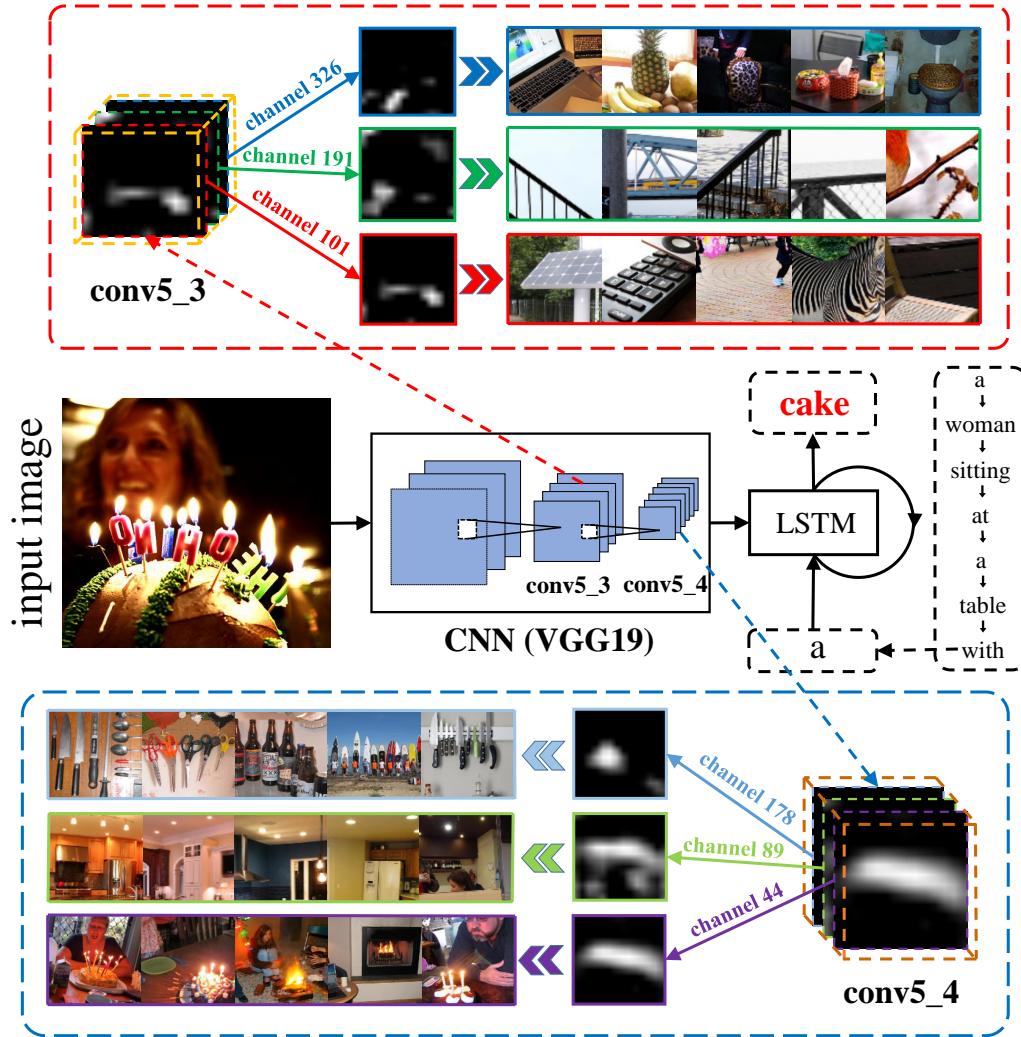


图 5-1 VGG19 网络中 conv5_4 层和 conv5_3 层的通道注意力机制示意图

的网络结构中。在三维的特征图中，每个通道本质上对应的是不同卷积核通道的响应。另一方面，卷积核也可以被看成是一种模式检测器：底层的卷积核往往倾向于检测一些底层的视觉特征（如：边、角等），而高层的卷积核往往倾向于检测一些高层的语义特征（如：物体等）^[129]。通过叠加多个卷积层，卷积神经网络实现对图像的多层次语义特征提取。因此，卷积神经网络的图像特征本质上具有三个维度：空间维度、通道维度、和层级维度。然而，目前所有注意力模型都只考虑了空间维度^[118]，即：只利用语句信息对卷积神经网络的最后一个特征图在空间维度上进行加权。

在本章，我们扩展现有的空间注意力模型，将注意力机制使用在卷积神经网络特征图的三个维度上。具体来说，我们提出了一种全新的网络结构：多层次空间和通道注意力网络（SCA-CNN），对多个卷积层的所有元素都进行加权。如图 5-1 所示，特征图的每一个通道本质上可以认为是一种特定属性或者物体检测器的响应结果，

即通道注意力也可以看成是基于生成语句对不同的属性特征进行选择。例如，当模型要预测单词“cake”时，通道注意力机制（如图 5-1 中的 conv5_4 层和 conv5_3 层）会对部分属性特征赋予更大的权重，如“火”、“光”、“蜡烛形状”等属性。另外，由于每个卷积层本质上都是底层卷积的输出结果，因此，可以对多个不同的卷积层同时在空间维度和通道维度使用注意力机制。例如，对于低层的特征图（如图 5-1 中 conv5_3 层）往往关注更加底层的属性特征。

我们在三个标准的图像描述生成数据集（Flickr8K、Flickr30K 和 MSCOCO）对模型 SCA-CNN 的性能进行评估。相比于现有的空间注意力模型^[118]，SCA-CNN 可以在评价指标 BLEU4 上提升 4.8%。总而言之，我们提出了一种全新的注意力机制网络，对卷积神经网络特征层在空间上、通道上、和层级上三个维度使用注意力机制。SCA-CNN 是一种通用的结构，可以运用在任意的卷积神经网络结构和网络层上，如 VGG^[9]、ResNet^[11] 等。SCA-CNN 也帮助我们更好的理解卷积神经网络特征在描述语句生成过程中的变化过程。

5.2 空间和通道注意力机制

5.2.1 概述

我们采用流行的编码-解码框架（encoder-decoder framework）对图像生成描述语句，即先用编码器（如卷积神经网络）将图像编码成一个向量表达，然后再使用解码器（如递归神经网络）将图像编码向量解码成描述语句。如图 5-2 所示，模型 SCA-CNN 利用语句信息对不同层级的特征图分别在空间维度和通道维度使用注意力机制。

假设当模型在生成第 t 个单词时，LSTM 的隐含状态为 $\mathbf{h}_{t-1} \in \mathbb{R}^d$ ，其中 d 是隐含状态的维度。对于第 l 层的特征，SCA-CNN 根据 \mathbf{h}_{t-1} 和目前的卷积特征 \mathbf{V}^l ，可以得到新的卷积特征 \mathbf{X}^l ：

$$\begin{aligned}\mathbf{V}^l &= \text{CNN}(\mathbf{X}^{l-1}), \\ \gamma^l &= \Phi(\mathbf{h}_{t-1}, \mathbf{V}^l), \\ \mathbf{X}^l &= f(\mathbf{V}^l, \gamma^l).\end{aligned}\tag{5-1}$$

其中， $\Phi(\cdot)$ 是空间和通道注意力函数， \mathbf{V}^l 是上一个卷积层的输出 \mathbf{X}^{l-1} 之后再接卷积层或池化层^[9,11]， $f(\cdot)$ 是线性加权函数。当卷积特征达到最后一层（第 L 层）时，

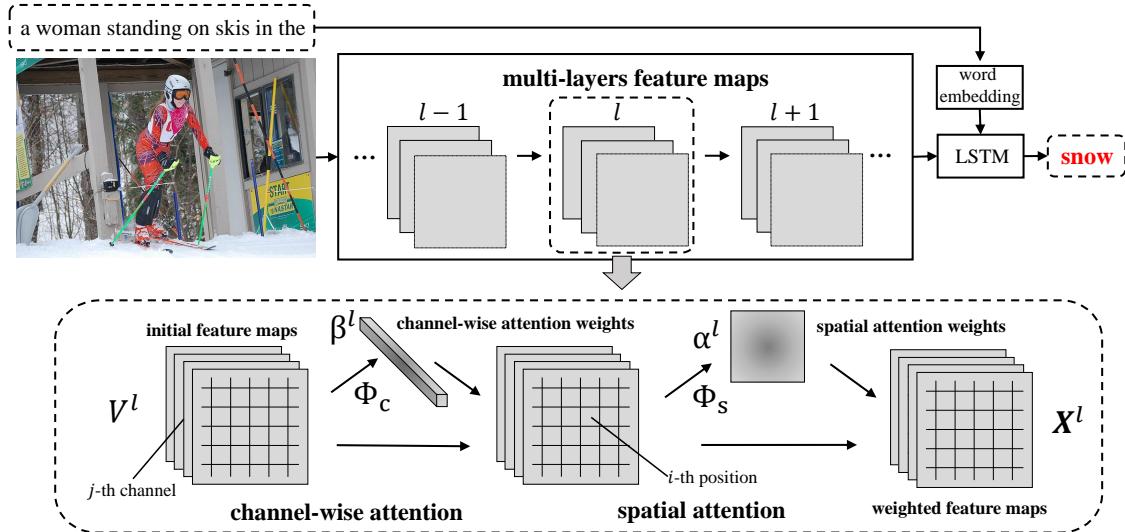


图 5-2 空间和通道注意力卷积神经网络流程图

我们使用 \mathbf{X}^L 来生成第 t 个单词：

$$\begin{aligned}\mathbf{h}_t &= \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{X}^L, y_{t-1}), \\ y_t &\sim p_t = \text{softmax}(\mathbf{h}_t, y_{t-1}).\end{aligned}\quad (5-2)$$

其中， $p_t \in \mathbb{R}^{|\mathcal{D}|}$ 是字典中所有单词的预测概率， \mathcal{D} 是预定义的字典，包含训练集所有语句中出现的所有单词。

当注意力参数 γ^l 与特征 \mathbf{V}^l 或 \mathbf{X}^l 具有相同的尺寸时（即 $W^l \times H^l \times C^l$ ），网络的计算量大小为 $\mathcal{O}(W^l H^l C^l k)$ ，其中 k 是将卷积网络特征 \mathbf{V}^l 和 LSTM 的隐含状态 \mathbf{h}_{t-1} 映射到同一空间中的维度大小。当特征图尺寸非常大时，对 GPU 的显存需求比较大。因此，我们提出将三维的 γ^l 分解成空间注意力参数 α^l 和通道注意力参数 β^l ：

$$\alpha^l = \Phi_s(\mathbf{h}_{t-1}, \mathbf{V}^l), \quad (5-3)$$

$$\beta^l = \Phi_c(\mathbf{h}_{t-1}, \mathbf{V}^l). \quad (5-4)$$

其中 Φ_c 和 Φ_s 分别表示通道注意力模型和空间注意力模型。这种简化将极大地减小计算空间到 $\mathcal{O}(C^l k + W^l H^l k)$ 。

5.2.2 空间注意力机制

通常，语句中的每个单词只与图像中的部分区域有关，如图 5-1 所示，当预测单词“cake”时，只有包含“蛋糕”的图像区域对于单词“cake”的预测有用。因此，在生成每个单词时，如果直接使用同一个全局图像特征，容易使模型陷入局部最优解。

空间注意力机制就是对空间维度上不同的图像区域特征赋予不同的权重。为了不失一般性，我们省略层数上角标 l 。我们先将视觉特征 \mathbf{V} 改写成 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$ ，其中 $\mathbf{v}_i \in \mathbb{R}^C$ ， $m = W \cdot H$ ， $\mathbf{v}_i \in \mathbb{R}^C$ 可以看成是第 i 个位置的图像特征。给定上一个时刻的 LSTM 的隐含状态 \mathbf{h}_{t-1} ，我们使用单层神经网络来生成空间注意力权重 α ：

$$\begin{aligned}\mathbf{a} &= \tanh((\mathbf{W}_s \mathbf{V} + \mathbf{b}_s) \oplus \mathbf{W}_{hs} \mathbf{h}_{t-1}), \\ \alpha &= \text{softmax}(\mathbf{W}_i \mathbf{a} + b_i).\end{aligned}\quad (5-5)$$

其中， $\mathbf{W}_s \in \mathbb{R}^{k \times C}$ 、 $\mathbf{W}_{hs} \in \mathbb{R}^{k \times d}$ 、 $\mathbf{W}_i \in \mathbb{R}^k$ 都是需要学习的映射矩阵，其中 \mathbf{W}_s 和 \mathbf{W}_{hs} 分别将视觉特征、隐含状态映射到同一个维度。符号 \oplus 表示矩阵和向量之间相加，即对矩阵中的每一个列向量都加上该向量。 $\mathbf{b}_s \in \mathbb{R}^k$, $b_i \in \mathbb{R}^1$ 是模型中可学习的偏置。

5.2.3 通道注意力机制

从公式 5-3 中可以看出，空间注意力机制需要使用视觉特征 \mathbf{V} 计算空间注意力参数，我们也同时可以使用通道注意力机制对 \mathbf{V} 在通道维度上进行加权。因为卷积神经网络的特征图的每一个通道本质上可以认为是一种特定属性或者物体检测器的响应结果，所以对特征图使用通道注意力机制可以看成是对不同的语义特征进行筛选的过程。

对于通道注意力机制，我们先将特征图 \mathbf{V} 改写成 \mathbf{U} ，其中 $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_C]$ ， $\mathbf{u}_i \in \mathbb{R}^{W \times H}$ 表示特征图 \mathbf{V} 的第 i 个通道， C 是总的通道数。然后，我们对每个通道使用平均池化，得到通道特征 \mathbf{v} ：

$$\mathbf{v} = [v_1, v_2, \dots, v_C], \mathbf{v} \in \mathbb{R}^C, \quad (5-6)$$

其中标量 v_i 是向量 \mathbf{u}_i 的平均，表示第 i 个通道的特征。于是，通道注意力模型 Φ_c 可以定义为：

$$\begin{aligned}\mathbf{b} &= \tanh((\mathbf{W}_c \otimes \mathbf{v} + \mathbf{b}_c) \oplus \mathbf{W}_{hc} \mathbf{h}_{t-1}), \\ \beta &= \text{softmax}(\mathbf{W}'_i \mathbf{b} + b'_i).\end{aligned}\quad (5-7)$$

其中 $\mathbf{W}_c \in \mathbb{R}^k$ 、 $\mathbf{W}_{hc} \in \mathbb{R}^{k \times d}$ 和 $\mathbf{W}'_i \in \mathbb{R}^k$ 是需要学习的映射矩阵， \otimes 表示向量间的外积， $\mathbf{b}_c \in \mathbb{R}^k$, $b'_i \in \mathbb{R}^1$ 是模型可学习的偏置。

根据空间注意力模型和通道注意力模型的使用顺序，总共有两种不同的组合形式：

通道-空间模型 (Channel-Spatial)：第一类是先使用通道注意力机制再使用空间注意力机制。对于这类，如图 5-2 所示，给定一个初始的视觉特征图 \mathbf{V} ，我们先使用通道注意力模型 Φ_c 得到通道注意力权重 β ，然后利用 β 对 \mathbf{V} 进行线性加权。之后，我们将加权的特征图输入到空间注意力模型 Φ_s 得到空间注意力权重 α 。在得到注意力权重 α 和 β 之后，我们可以得到最终的特征图 \mathbf{X} ：

$$\begin{aligned}\beta &= \Phi_c(\mathbf{h}_{t-1}, \mathbf{V}), \\ \alpha &= \Phi_s(\mathbf{h}_{t-1}, f_c(\mathbf{V}, \beta)), \\ \mathbf{X} &= f(\mathbf{V}, \alpha, \beta).\end{aligned}\tag{5-8}$$

其中， $f_c(\cdot)$ 表示特征图通道与通道注意力权重在通道维度上进行乘积。

空间-通道模型 (Spatial-Channel)：第二类是先使用空间注意力机制再使用通道注意力机制。对于这类，给定一个初始的视觉特征图 \mathbf{V} ，我们先使用空间注意力模型 Φ_s 得到空间注意力权重 α 。通过注意力权重、线性函数 $f_s(\cdot)$ 和通道注意力模型 Φ_c ，我们可以得到最终的特征图 \mathbf{X} ：

$$\begin{aligned}\alpha &= \Phi_s(\mathbf{h}_{t-1}, \mathbf{V}), \\ \beta &= \Phi_c(\mathbf{h}_{t-1}, f_s(\mathbf{V}, \alpha)), \\ \mathbf{X} &= f(\mathbf{V}, \alpha, \beta).\end{aligned}\tag{5-9}$$

其中， $f_s(\cdot)$ 表示特征图空间不同区域特征与空间注意力权重进行乘积。

5.3 实验设置与性能对比

5.3.1 图像描述生成任务的数据集和评价指标

图像描述生成数据集：我们在三个通用的图像描述生成数据集对模型 SCA-CNN 的性能进行评估。各个数据集具体的细节如下：

Flickr8k^[235]：它一共包含 8000 张图像。按照官方的数据集划分，其中 6000 张图像作为训练集，1000 张图像为验证集，1000 张图像为测试集。

Flickr30k^[236]：它一共包含 31000 张图像。由于这个数据集缺少官方的数据集划分，我们参考 Karpathy 等人^[112] 的数据集划分，将其中 29000 张图像作为训练集，1000 张图像作为验证集，1000 张图像作为测试集。

MSCOCO^[1]：根据该数据集的官方划分，训练集包含 82783 张图像，验证集包含 40504 张图像，以及测试集包含 40775 张图像。由于官方测试集中所有的图像

都没有公开其中的人工标注信息。我们同样参考 Karpathy 等人^[112] 的数据集划分，其中验证集和测试集各包含 5000 张图像。

图像描述生成评价指标：我们在四个常用的图像描述生成评价指标对模型 SCA-CNN 的性能进行评估。各个评价指标具体的细节如下：

BLEU^[237] (B@1, B@2, B@3, B@4): BLEU (Bilingual evaluation understudy) 通过比较生成语句和所有人工标注语句中的 n 元词组 (n-gram) 的准确率，然后对所有的 n 元词组准确率计算几何平均数。其中，四元词组 (即 $n = 4$) 时计算得到的评估结果与人工的评估结果最接近。

METEOR^[238] (MT): METEOR (Metric for Evaluation of Translation with Explicit ORdering) 是基于生成语句和人工标注语句中一元词组的对齐关系。先对所有的一元词组进行精确匹配、近义词匹配、词干匹配之后，计算所有一元词组匹配的 F 分数 (F-measure)。同时，METEOR 引入一个权重系数，对生成过长的语句减小权重系数，通过对 F 分数进行加权得到每个生成语句的分数。对于多个人工标注语句，METEOR 分别用生成语句和每个标注语句单独计算分数，然后选取其中的最高分作为最终的分数。

CIDEr^[239] (CD): CIDEr (Consensus-based Image Description Evaluation) 先将所有单词进行词干提取 (stemming)，然后对所有预处理后的语句中的 n 元词组利用 TF-IDF^[240] 进行加权，最后利用所有 n 元词组的余弦相似度得到分数。

ROUGE-L^[241] (RG): ROUGE(Recall-Oriented Understudy for Gisting Evaluation) 先找到生成语句和人工标注语句之间最长的相同序列，然后基于最长的相同序列计算匹配的 F 分数。同样，对于多个人工标注语句，ROUGE-L 分别单独计算与每个标注语句的分数，然后选取最高分作为最终分数。

总之，这四种评价指标都是通过比较生成语句和人工标注语句中 n 元词组的匹配程度。所有实验中这些评价指标的计算都是直接采用 MSCOCO 官方的测评工具：<https://github/tylin/coco-caption>。

5.3.2 实验细节设定

对于图像编码部分，我们采用两种流行的卷积神经网络：VGG-19^[9] 和 ResNet-152^[11]。对于文本解码部分，我们使用递归神经网络 LSTM^[116] 来生成描述语句。单词编码向量的维度和 LSTM 的隐含状态的维度分别设定为 100 和 1000。用于计算注意力权重的共同空间维度设置为 512。对于 Flickr8k 数据集，批处理大小设置为 16；对于 Flickr30k 和 MSCOCO，批处理大小设置为 64。为了避免模型过拟合，我

们采用 dropout 和 early stopping 机制。整个 SCA-CNN 模型直接采用端到端的训练方式，用优化算法 Adadelta^[242] 进行参数优化。当模型刚好预测一个特定的“END”字符或者达到了预先设定的句子最长的长度时，整个语句的生成过程将会终止。在测试阶段，我们采用 BeamSearch^[22] 的方法，在每个时刻选择 5 个语句作为候选答案。

5.3.3 通道注意力机制的性能分析

实验设定：本小节主要探讨通道注意力机制对现有空间注意力模型的影响。我们总共比较了五种实验设定：(1) **Spatial**: 它是一个空间注意力模型，对卷积网络最后一层特征图先计算空间注意力权重，然后对最后一层特征图的不同空间区域特征进行加权。对于 VGG19 和 ResNet152 网络，最后一层分别为 conv5_4 和 res5c。在对最后一层特征图进行加权之后，我们将加权后的特征图输入的原本的网络结构中。对于 VGG19 网络而言，conv5_4 层之后还有两个全连接层；对于 ResNet152 网络而言，res5c 层之后是一个平均池化层。(2) **Channel**: 它是一个通道注意力模型，和空间注意力模型 Spatial 基本相同，除了将空间注意力模型（公式 (5-3)）替换成通道注意力模型（公式 (5-4)）。(3) **Channel-Spatial**: 第一种融合方式，先使用通道注意力模型，再使用空间注意力模型（公式 (5-8)）。(4) **Spatial-Channel**: 第二种融合方式，先使用空间注意力模型，再使用通道注意力模型（公式 (5-9)）。(5) **HAT**: 它是 Xu 等人^[118] 提出的“硬注意力”模型 (Hard-ATtention, HAT)。HAT 模型和 Spatial 模型一样，都属于空间注意力模型。但是与 Spatial 模型主要有两个区别：第一个是注意力权重和特征图的融合方式，第二个是是否将加权的特征图输入到后续的网络结构中。所有的实验结果都展示在表 5-1 中。

实验结果：从表 5-1 中可以得到以下发现：(1) 对于 VGG19 网络而言，Spatail 模型的结果比 HAT 的好；但对于 ResNet152 网络，实验结果相反。这主要的原因在于 VGG19 网络中有全连接层，可以保持空间语义信息，而 ResNet152 网络中只有平均池化层，将会破坏原有的空间语义信息。(2) 相比于 VGG19 网络，Channel 模型在 ResNet152 网络中可以显著提升性能（与 Spatial 模型相比）。这主要的原因在于 ResNet152 网络的最后一个卷积特征图拥有更多的通道数（如：ResNet152 有 2048 个通道，而 VGG19 网络只有 512 个通道）。(3) 在 ResNet152 网络中，相比于 Spatial 模型，Channel-Spatial 和 Spatial-Channel 都能显著提升性能。这说明通道注意力机制在特征图通道数足够大时能显著提升性能。(4) 在 VGG19 网络和 ResNet152 网络中，Channel-Spatial 和 Spaital-Channel 两个模型的效果都非常接近，

Dataset	Network	Method	B@4	MT	RG	CD
Flickr8k	VGG	Spatial	23.0	21.0	49.1	60.6
		HAT	21.3	20.3	—	—
		Channel	22.6	20.3	48.7	58.7
		Spatial-Channel	22.6	20.9	48.7	60.6
		Channel-Spatial	23.5	21.1	49.2	60.3
	ResNet	Spatial	20.5	19.6	47.4	49.9
		HAT	21.7	20.1	48.4	55.5
		Channel	24.4	21.5	50.0	65.5
		Spatial-Channel	24.8	22.2	50.5	65.1
		Channel-Spatial	25.7	22.1	50.9	66.5
Flickr30k	VGG	Spatial	21.1	18.4	43.1	39.5
		HAT	19.9	18.5	—	—
		Channel	20.1	18.0	42.7	38.0
		Spatial-Channel	20.8	17.8	42.9	38.2
		Channel-Spatial	21.0	18.0	43.3	38.5
	ResNet	Spatial	20.5	17.4	42.8	35.3
		HAT	20.1	17.8	42.9	36.3
		Channel	21.5	18.4	43.8	42.2
		Spatial-Channel	21.9	18.5	44.0	43.1
		Channel-Spatial	22.1	19.0	44.6	42.5
MS COCO	VGG	Spatial	28.2	23.3	51.0	85.7
		HAT	25.0	23.0	—	—
		Channel	27.3	22.7	50.1	83.4
		Spatial-Channel	28.0	23.0	50.6	84.9
		Channel-Spatial	28.1	23.5	50.9	84.7
	ResNet	Spatial	28.3	23.1	51.2	84.0
		HAT	28.4	23.2	51.2	84.9
		Channel	29.5	23.7	51.8	91.0
		Spatial-Channel	29.8	23.9	52.0	91.2
		Channel-Spatial	30.4	24.5	52.5	91.7

表 5-1 VGG-19 网络和 ResNet-152 网络中单层注意力机制的性能对比

其中 Channel-Spatial 稍微高一点。在之后的实验中，我们使用 Channel-Spatial 作为空间注意力机制和通道注意力机制的融合方式。

Dataset	Network	Method	B@4	MT	RG	CD
Flickr8k	VGG	1-layer	23.0	21.0	49.1	60.6
		2-layer	22.8	21.2	49.0	60.4
		3-layer	21.6	20.9	48.4	54.5
	ResNet	1-layer	20.5	19.6	47.4	49.9
		2-layer	22.9	21.2	48.8	58.8
		3-layer	23.9	21.3	49.7	61.7
Flickr30k	VGG	1-layer	21.1	18.4	43.1	39.5
		2-layer	21.9	18.5	44.3	39.5
		3-layer	20.8	18.0	43.0	38.5
	ResNet	1-layer	20.5	17.4	42.8	35.3
		2-layer	20.6	18.6	43.2	39.7
		3-layer	21.0	19.2	43.4	43.5
MS COCO	VGG	1-layer	28.2	23.3	51.0	85.7
		2-layer	29.0	23.6	51.4	87.4
		3-layer	27.4	22.9	50.4	80.8
	ResNet	1-layer	28.3	23.1	51.2	84.0
		2-layer	29.7	24.1	52.2	91.1
		3-layer	29.6	24.2	52.1	90.3

表 5-2 空间注意力模型在 VGG-19 网络和 ResNet-152 网络下不同层的性能对比

5.3.4 多层注意力机制的性能分析

实验设定：本小节主要探讨在卷积神经网络的不同层级中使用空间注意力机制或通道注意力机制对模型性能的影响。我们分别对 Spatial 模型和 Channel-Spatial 模型进行了不同层级的实验：“一层”(1-layer)、“两层”(2-layer)、“三层”(3-layer) 分别表示在一个层级、两个层级和三个层级上使用注意力机制。对于 VGG19 网络，这三个层级分别表示：conv5_4、conv5_3 和 conv5_2；对于 ResNet152 网络，这三个层级分别表示 res5c、res5c_branch2b 和 res5c_branch2a。

实验结果：从表 5-2 的结果和表 5-3 结果，我们有以下发现：(1) 在绝大多数的实验设定中，模型都可以通过在更多的卷积层上使用注意力机制来提升实验结果。这主要是因为在多层特征上使用注意力机制可以帮助模型对不同层次的语义特征上进行选择性关注。(2) 在太多的卷积层上使用注意力机制也容易造成过拟合。例如：当注意力机制改变的层数增加时，Flickr8K 更容易造成性能下降。这主要原因

Dataset	Network	Method	B@4	MT	RG	CD
Flickr8k	VGG	1-layer	23.5	21.1	49.2	60.3
		2-layers	22.8	21.6	49.5	62.1
		3-layers	22.7	21.3	49.3	62.3
	ResNet	1-layer	25.7	22.1	50.9	66.5
		2-layers	25.8	22.4	51.3	67.1
		3-layers	25.3	22.9	51.2	67.5
Flickr30k	VGG	1-layer	21.0	18.0	43.3	38.5
		2-layers	21.8	18.8	43.7	41.4
		3-layers	20.7	18.3	43.6	39.2
	ResNet	1-layer	22.1	19.0	44.6	42.5
		2-layers	22.3	19.5	44.9	44.7
		3-layers	22.0	19.2	44.7	42.8
MS COCO	VGG	1-layer	28.1	23.5	50.9	84.7
		2-layers	29.8	24.2	51.9	89.7
		3-layers	29.4	24.0	51.7	88.4
	ResNet	1-layer	30.4	24.5	52.5	91.7
		2-layers	31.1	25.0	53.1	95.2
		3-layers	30.9	24.8	53.0	94.7

表 5-3 空间和通道注意力模型在 VGG-19 网络和 ResNet-152 网络下不同层的性能对比

是因为 Flickr8K 有 6000 张训练图像，而 MSCOCO 有 82783 张训练图像。

5.3.5 空间和通道注意力卷积神经网络的性能比较

实验设定：我们将本章提出的模型 SCA-CNN 与目前最好的图像描述生成方法进行对比，这些方法主要可以分为三类：(1) **Deep VS**^[112]、**NIC**^[22]、**m-RNN**^[114]。这些方法都是端到端的编码-解码框架，并且模型中没有包含注意力机制。(2) **SAT**^[118] 和 **HAT**^[118] 是空间注意力模型。其中模型 SAT (Soft-ATTention) 是在每步生成单词的过程中对所有的空间区域进行线性加权，而模型 HAT (Hard-ATTention) 是在每步生成单词的过程中对所有的区域只采样其中一个区域特征。(3) **gLSTM**^[128] 和 **ATT**^[125] 是属性注意力模型。其中模型 gLSTM 使用图像和生成的描述语句作为全局的属性信息，而模型 ATT 使用图像额外检测的属性作为属性信息。表 5-4 的结果“ours(V)”和“ours(R)”分别表示 VGG19 网络和 ResNet152 网络中两层 Channel-Spatial 模型。另外，我们还将模型上传到 MSCOCO 在线服务器对官方测试集进行



图 5-3 空间注意力和通道注意力权重的可视化结果

预测，结果展示在表 5-5 中。

实验结果：从表 5-4 和表 5-5 中可以看出，模型 SCA-CNN 可以超过目前现有的模型，这是因为 SCA-CNN 分别考虑了卷积神经网络特征层的三个维度：空间、通道和层级，而其他的方法只考虑了其中一个维度。在 MSCOCO 在线服务器上，模型 SCA-CNN 性能比模型 ATT 和模型 NIC 低的主要原因来自两个方面：(1) ATT 和 NIC 的结果都是基于集成模型 (ensemble model) 的结果，而 SCA-CNN 是单个模型的结果。(2) 它们使用更强的卷积神经网络对输入图像提取特征，如模型 NIC 使用 Inception-V3 网络^[243] 而 SCA-CNN 只使用 ResNet152 网络。

Model	Flickr8k				Flickr30k				MS COCO			
	B@2	B@3	B@4	MT	B@2	B@3	B@4	MT	B@2	B@3	B@4	MT
Deep VS	38.3	24.5	16.0	—	36.9	24.0	15.7	—	45.0	32.1	23.0	19.5
NIC	41.0	27.0	—	—	42.3	27.7	18.3	—	46.1	32.9	24.6	—
m-RNN	—	—	—	—	41.0	28.0	19.0	—	49.0	35.0	25.0	—
SAT	44.8	29.9	19.5	18.9	43.4	28.8	19.1	18.5	49.2	34.4	24.3	23.9
HAT	45.7	31.4	21.3	20.3	43.9	29.6	19.9	18.5	50.4	35.7	25.0	23.0
gLSTM	45.9	31.8	21.2	20.6	44.6	30.5	20.6	17.9	49.1	35.8	26.4	22.7
ATT	—	—	—	—	46.0	32.4	23.0	18.9	53.7	40.2	30.4	24.3
ours(V)	46.6	32.6	22.8	21.6	45.3	31.7	21.8	18.8	53.3	39.7	29.8	24.2
ours(R)	49.6	35.9	25.8	22.4	46.8	32.5	22.3	19.5	54.8	41.1	31.1	25.0

表 5-4 不同描述语句生成算法在数据集 Flickr8k、Flickr30k 和 MSCOCO 上的性能对比

Model	B@1		B@2		B@3		B@4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40								
ours	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
HAT	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
ATT	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	95.3	95.8
NIC	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6

表 5-5 不同图像描述语句生成算法在数据集 MSCOCO 的在线服务器上的性能对比

5.3.6 空间注意力和通道注意力权重的可视化

如图 5-3 所示，我们还提供了模型 SCA-CNN 的注意力权重在 VGG19 网络的可视化示例。其中“Layer-1”和“Layer-2”分别表示 conv5_4 和 conv5_3 层。每个示例包含三句描述语句：“Ours”(SCA-CNN)、“SAT”(Soft-ATTention) 和“GT”(Ground Truth)。第二列的上部分是空间注意力权重，白色部分表示空间注意力权重较大，而灰色部分表示空间注意力权重较小。第二列下部分为所有通道权重的直方统计图。第三列的数字表示通道注意力权重最大的两个通道编号。后面的五张图像是数据集 MSCOCO 训练集中对同样通道编号响应最大的图像。为了简洁，我们只展示了预测语句中的其中一步。例如第一个例子，当模型 SCA-CNN 准备预测单词“umbrella”时，我们的通道注意力机制往往对例如“伞”、“棍状”以及“圆形”等语义信息响应强烈的通道赋予更大的权重。为了表示每个通道表示的语义信息，我们使用和 Zeiler 等人^[129]相同的方法，其中红色框表示对应通道最强响应的感受野。

5.4 本章小结

在本章，我们提出了一种全新的基于注意力机制的卷积神经网络：多层次空间和通道注意力网络（SCA-CNN）对图像生成描述语句。SCA-CNN 充分利用了卷积神经网络特征图的三个维度信息（空间维度、通道维度和层级维度），大大提升了卷积网络的编码能力，并在图像描述生成任务中达到了目前最好的性能。本章的贡献不仅仅是提出了一个更强的注意力机制（通道注意力机制），同时帮助人们理解在语句生成过程中，注意力机制在卷积网络特征图中的变化过程。

6 基于密集型自底向上网络的视频片段检索方法

在本章，我们主要解决视频片段检索任务。具体来说，给定一个查询（如：自然语句或者视频片段）和一段未裁剪的长视频序列，视频片段检索模型需要在视频序列中定位出和查询内容相匹配的视频片段。目前，现有的视频片段检索方法可以分为两大类：(1) 自顶向下 (Top-down) 的方法：它们先将整个视频序列切分成若干个候选视频片段，然后对每个候选片段分别进行分类和回归。其中分类主要是计算候选片段与查询的相似度，回归主要是计算视频片段微调的偏移大小。(2) 自底向上 (Bottom-up) 的方法：先将视频序列和查询进行特征融合，然后对融合后的特征序列中的每一帧分别预测其属于视频序列定位边界的概率（即起始时刻和终止时刻）。然而，这两类方法都各自具有明显的缺点：自顶向下的方法需要人为地预先设定许多切分的规则（如：候选片段的大小、候选片段的数量等），同时自顶向下模型定位速度也相对较慢，而自底向上的方法的目前在性能还低于自顶向上的方法。在本章，我们重点分析了现有自底向上模型的设计缺陷，提出了一种全新的密集型自底向上网络。我们将位于起始时刻和终止时刻之间的每一帧都看成是正样本，然后对每一个正样本帧都回归其各自到两个定位边界的距离。与此同时，为了更好的适应密集型自底向上的框架，我们还提出了一种基于图结构的特征金字塔网络，来强化目前的骨干网络 (backbone) 的输出特征帧序列。我们先将多尺度的特征帧序列映射到同一个语义空间中，然后利用图卷积来学习语义空间中不同特征间的内在联系。我们在常用的四个视频片段检索数据集 (TACoS^[244]、Charades-STA^[23]、ActivityNet Captions^[245]、Activity-VRL^[142]) 中验证了我们方法的有效性。我们提出的密集型自底向上网络不仅可以在性能上超过目前所有的方法，同时还可以保持和其他自底向上的模型相同的定位速度。

6.1 问题描述

视频片段检索是视觉场景理解领域一个重要的研究问题。它不仅仅需要准确地把握输入查询的语义内容，同时需要对视频内容有正确的理解。随着大规模视

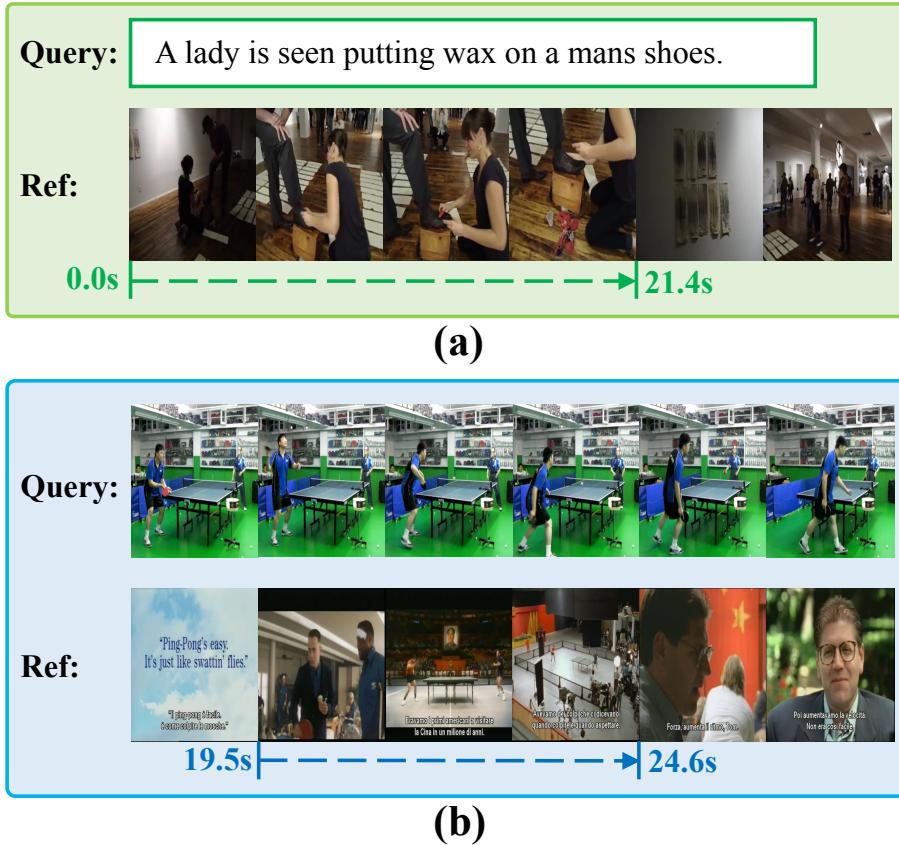


图 6-1 两种不同的视频片段检索任务

频数据集的出现^[5,246] 和视频特征学习的发展^[247]，目前主要有两种视频片段检索任务：(1) 基于语句查询的视频片段检索，即查询内容是一个自然语言描述语句（如图 6-1 (a) 所示）。(2) 基于视频查询的视频片段检索，即查询内容是包含一个动作的短视频片段（如图 6-1 (b) 所示）。这两种视频片段检索任务的目标完全相同：在视频序列中定位出两个边界时刻（起始时刻和终止时刻），使得从起始时刻到终止时刻之间的视频片段内容刚好与查询内容一致。另外，视频片段检索已经成为众多重要的视频应用技术的基础。如：基于内容的精彩片段检索、行人重识别等。

到目前为止，绝大多数的视频片段检索方法都属于**自顶向下的**方法：它们将视频序列切分为众多的候选视频片段，然后对每个候选片段进行分类和回归。具体来说，这些自顶向下的方法又可以细分为两类：(1) 滑窗型^[23,135,136,248–252]：它们预先定义一系列不同大小的滑窗，然后利用滑窗密集地将视频“显式”地切分成若干候选视频片段，最后分别对查询和候选片段提取特征。这样，视频片段检索任务就转变为一个相似度匹配问题。但是这类方法忽略了视频中每个视频帧与其他周围帧之间的内在联系，并且通常这些周围帧往往对视频理解有着巨大的帮助^[253]。(2) 锚框型 (anchor-based)^[137,254]：它们不像滑窗型直接对视频预先切分，而是对于每个

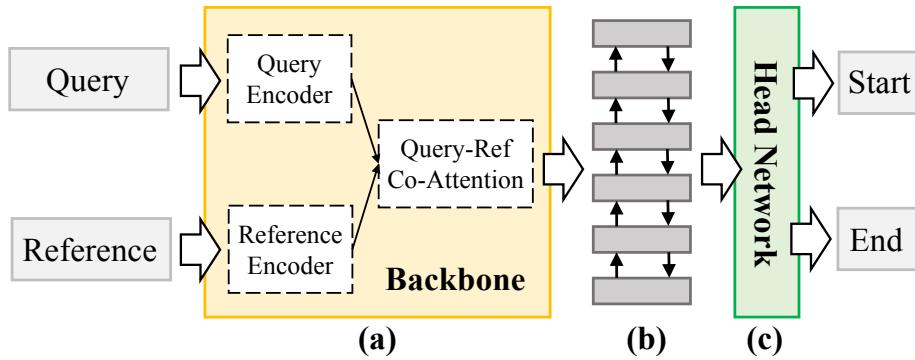


图 6-2 典型的稀疏型自底向上视频片段检索模型

视频帧都定义若干个锚框，然后对每个锚框内的视频进行分类和回归。为了充分地利用周围的视频帧，它们通常采用递归神经网络将所有的视频帧进行串接。这类方法可以看成是基于锚框型的目标检测方法^[14]在视频领域的拓展。

尽管这些自顶向下的模型（包括滑窗型和锚框型）可以在多个视频片段检索数据集上达到目前最好的性能，但是这类方法本质上仍然有许多一些不可避免的缺点：(1) 最终的视频片段检索结果受预先定义的规则影响较大（如滑窗或锚框的大小、数量等）；(2) 为了尽可能的提高召回率，模型必须增大滑窗或锚框数量，这将导致整个计算量增大、定位速度慢。

为了消除上述这些缺点，一些视频片段检索方法^[138,139,142]开始借鉴自然语言处理领域中阅读理解任务（reading comprehension）的方法^[255–257]，用一种**稀疏型自底向上的**网络直接预测两个边界的概率。如图 6-2 所示，一个典型的稀疏型自底向上模型通常包含两个组成部分：骨干网络（图 6-2 (a)）和头网络（图 6-2 (c)）。骨干网络通常会使用协同注意力机制（co-attention mechanism）来融合查询特征和每个视频帧的特征，它的输出是融合后的特征帧序列（图 6-2 (b)）。为了后续的头网络能够直接对视频帧序列中每一帧预测边界概率，融合后的特征帧序列往往需要保持和输入视频帧相同的长度。尽管这种稀疏型自底向上的方法可以避免自顶向下方法的缺点，但是它们的性能目前却仍然低于自顶向下方法。尤其是对于长视频（如：数据集 TACoS），这种差距往往更加明显。在本章，我们认为自底向上方法的性能低于自顶向下方法的主要原因来自于目前骨干网络和头网络的不合理设计：

骨干网络（Backbone Network）：对于骨干网络的设计，目前的稀疏型自底向上模型主要有两个缺点：(1) 每个视频通常包含丰富的场景变化，即不同的视频场景分布在视频的不同位置中。因此，理解视频中不同场景的变化以及场景之间的关系对于充分理解视频内容十分重要。然而，目前的方法通常使用递归神经网络 RNN

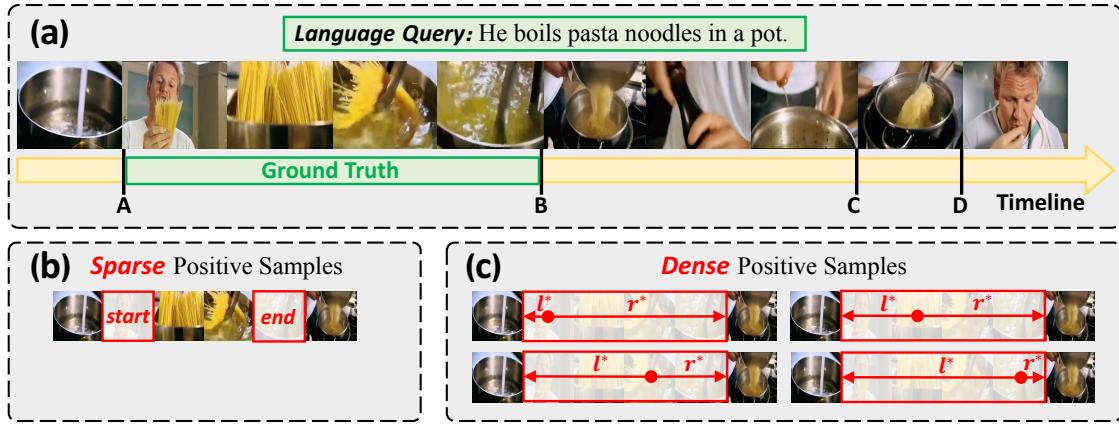


图 6-3 一个基于语句的视频片段检索示例

来编码视频帧特征，忽略了场景级别的关系。(2) 为了方便头网络的帧预测，骨干网络需要让融合后的特征帧序列保持和原始视频阵相同的长度，这容易导致每个视频帧特征只编码局部的语义信息，而忽略了全局的视频语义信息^[258,259]。

头网络 (Head Network)：对于头网络的设计，目前的稀疏型自底向上模型主要有三个缺点：(1) 两个边界时刻（起始时刻和终止时刻）的预测是相互独立的，即模型预测边界时忽略了两个边界内部视频内容的一致性。如图 6-3 (a)，在时刻 B 和时刻 D 时的视频帧有非常相似的场景内容。因此，模型很容易将结果预测为 (A→D)，即使在时刻 (B→C) 之间存在有明显的视频场景内容变化。(2) 在训练过程中，正样本和负样本的数量极度不均。因为视频的长度通常较长（如：数据集 TACoS 中每个视频的平均长度为 9000 帧），而只有其中两个边界帧为正样本（如图 6-3 (b))。(3) 即使没有查询的约束，对视频中发生动作的边界进行预测本身仍然是目前尚未解决的开放性难题^[260]。因为它不仅需要判断每个视频帧内容和查询内容是否相关，同时需要判断该视频帧是否为动作边界。

在本章，我们针对稀疏型自底向上模型的缺点，提出了一种全新的密集型自底向上网络：基于图特征金字塔的密集型预测 (Graph-FPN with Dense Prediction, GDP)。对于骨干网络，GDP 引入一个图特征金字塔层来增强骨干网络输出的特征帧序列（图 6-2 (b))。GDP 首先构建一个金字塔多尺度特征，然后将不同尺度的特征序列映射到一个高语义的场景空间中，通过利用图卷积对场景空间中的节点进行特征融合。图卷积不仅可以充分地利用不同语义场景之间的内在联系，同时可以消除不同尺度下特征的语义差。最后，这些场景空间的特征组成为新的特征序列。对于头网络，我们将起始时刻到终止时刻中间的每一帧都看成是正样本。对于每个正样本，GDP 包含一个回归网络来预测从当前帧到两个边界时刻各自的距离。这

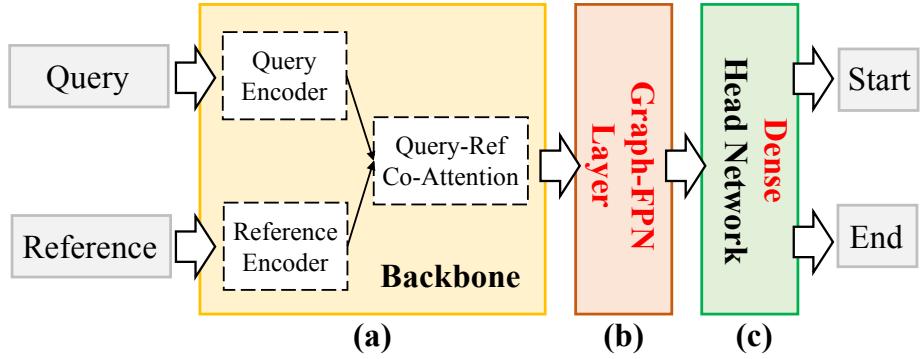


图 6-4 典型的稀疏型自底向上视频片段检索模型

样的设计一方面可以缓解训练过程中正负样本极度不均的问题，另一方面由于两个边界预测来自于同一个特征，也可以避免陷入独立预测的局部最优。同时，GDP 包含一个置信网络分支来预测当前帧与查询的关联度，可以将边界帧预测任务分离成关联度判断和边界回归两个任务，减少了视频片段检索任务的难度。

我们在四个常用的视频片段检索数据集 (TACoS^[244]、Charades-STA^[23]、ActivityNet Captions^[245] 和 Activity-VRL^[142]) 中验证了模型 GDP 的有效性。在多个不同的评估指标下，模型 GDP 都超过了目前所有的自顶向下的方法，同时保持了稀疏型自底向上方法的定位速度。

6.2 基于图特征金字塔的密集型预测

给定一个视频序列 \mathcal{V} 和查询 Q ，视频片段检索任务需要预测出两个边界时刻 (t_s, t_e) ，满足从 t_s 到 t_e 之间的视频片段内容与查询内容一致。在本节，我们将首先介绍 GDP 模型中的组合部分 6-4，包括骨干网络 (a)、图特征金字塔层 (b) 和密集型头网络 (c)。然后，我们再介绍 GDP 的训练和测试过程。

6.2.1 骨干网络

GDP 的骨干网络使用模型 QANet^[257] 来融合查询和视频的特征。如图 6-5 所示，QANet 共有两个输入：查询特征 $Q = \{\mathbf{q}_n\}_{n=1}^N$ 和视频序列特征 $V = \{\mathbf{v}_i\}_{i=1}^T$ ，其中 N 和 T 分别表示查询和视频的长度。具体来说，QANet 包含四个主要部分：

(1) 查询特征编码器：查询特征编码器包含多个特征编码层对输入的查询特征进行编码。如图 6-5 (b) 所示，每个特征编码层由多个卷积层、层归一化层、自注意力层和全连接层组成。查询特征编码器的输出是 $\tilde{Q} = \{\tilde{\mathbf{q}}_n\}_{n=1}^N$ 。

(2) 视频特征编码器：视频特征编码器对输入的视频特征进行编码。其结构和

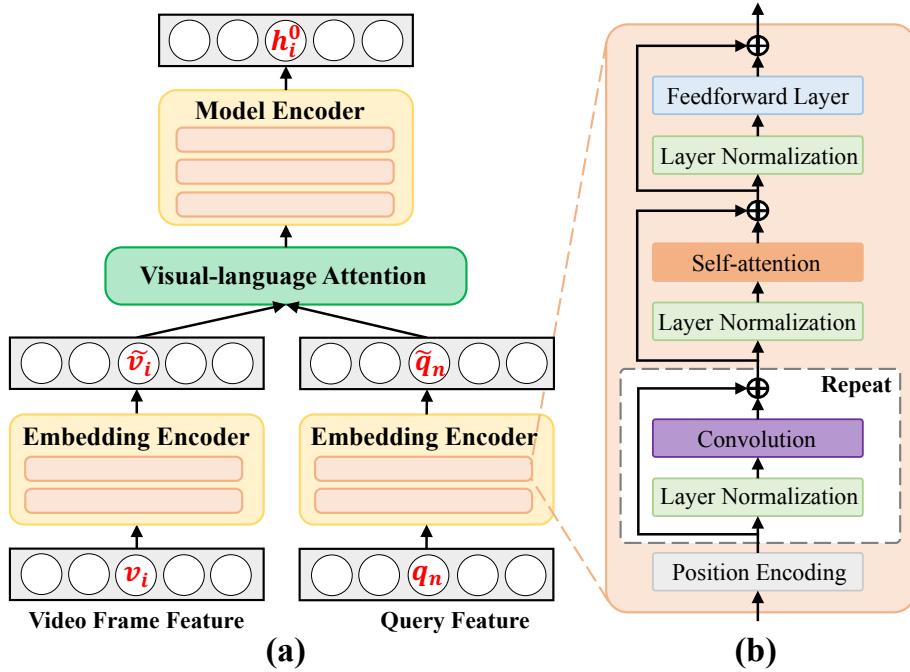


图 6-5 QANet 的模型结构

查询特征编码器完全相同，即由多个特征编码层组成（如图 6-5 (b)）。视频特征编码器的输出是 $\tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_i\}_{i=1}^T$ 。

(3) 查询-视频协同注意力层：查询-视频协同注意力层包含一个协同注意力机制来融合查询特征 $\tilde{\mathbf{Q}} = \{\tilde{\mathbf{q}}_n\}_{n=1}^N$ 和视频特征 $\tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_i\}_{i=1}^T$ 。具体来说，它先计算一个相似矩阵 $\mathbf{S} \in \mathbb{R}^{T \times N}$ ，其中每个元素 S_{ij} 表示 $\tilde{\mathbf{v}}_i$ 和 $\tilde{\mathbf{q}}_j$ 之间的相似度。然后可以得到两个加权特征：

$$\mathbf{A} = \bar{\mathbf{S}} \cdot \tilde{\mathbf{Q}}, \quad \mathbf{B} = \bar{\mathbf{S}} \cdot \bar{\mathbf{S}}^T \cdot \tilde{\mathbf{V}}, \quad (6-1)$$

其中 $\bar{\mathbf{S}}$ 和 $\bar{\mathbf{S}}$ 分别是对 \mathbf{S} 按行和按列进行归一化。

(4) 融合特征编码器：给定两个注意力权重矩阵 \mathbf{A} 和 \mathbf{B} ，融合特征编码器开始对融合后的特征进行编码。融合特征编码器同样由多层特征编码层（图 6-5 (b)）组成。融合特征编码器的输入是一个特征序列，它的第 i 个特征为 $[\mathbf{v}_i, \mathbf{a}_i, \mathbf{v}_i \odot \mathbf{a}_i, \mathbf{v}_i \odot \mathbf{b}_i]$ ，其中 \mathbf{a}_i 和 \mathbf{b}_i 分别是矩阵 \mathbf{A} 和 \mathbf{B} 的第 i 行， \odot 是元素积， $[,]$ 是向量连接符。融合特征编码器的输出为 $\mathbf{H}_0 = \{\mathbf{h}_i^0\}_{i=1}^T$ ， $\mathbf{H}_0 \in \mathbb{R}^{T \times D}$ ，其中每个特征 $\mathbf{h}_i^0 \in \mathbb{R}^D$ 都编码了查询信息。现有的稀疏型自底向上模型通常直接将 \mathbf{H}_0 作为头网络的输入，相反，模型 GDP 包含一个图特征金字塔层对特征 \mathbf{H}_0 进行增强。值得注意的是，模型 GDP 可以对任意的骨干网络兼容。

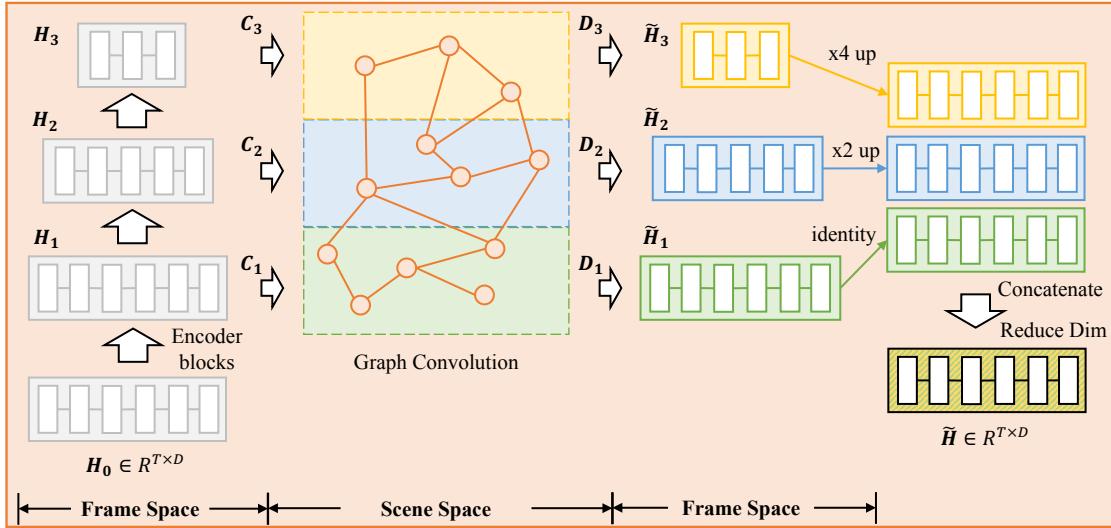


图 6-6 图特征金字塔层

6.2.2 图特征金字塔层

如图 6-6 所示，图特征金字塔层主要包含四个步骤来增强骨干网络输出 H_0 ：

(1) 特征金字塔的构建：给定特征帧序列 H_0 ，我们首先通过逐渐减少特征序列长度来构建特征金字塔 $\{H_i \in \mathbb{R}^{T_i \times D}, H_1 \in \mathbb{R}^{T_1 \times D}, H_2 \in \mathbb{R}^{T_2 \times D}, H_3 \in \mathbb{R}^{T_3 \times D}\}$ ，其中 $T_{i+1} = T_i/2$ 。我们同样使用相同的多个特征编码层（如图 6-5 (b)）和一个额外的步长为 2 的卷积层将特征 H_i 转换为 H_{i+1} 。

(2) 帧空间到场景空间：在得到多个尺度的特征 $\{H_1, H_2, H_3\}$ 之后，我们将这些特征从原始的帧空间映射到场景空间。以 $H_2 = \{h_i^2\}_{i=1}^{T_2}$ 为例，我们希望得到一系列场景空间特征 $X_2 = f_2(H_2) \in \mathbb{R}^{N_2 \times D}$ ，其中 N_2 表示场景空间在该尺度特征的节点数量，映射函数 $f_2(\cdot)$ 是对原始输入特征的线性组合：

$$x_i^2 = c_i^2 H_2 = \sum_j c_{ij}^2 h_j^2, \quad (6-2)$$

其中 $C_2 = [c_1^2, \dots, c_{N_2}^2]$ ， $C_2 \in \mathbb{R}^{N_2 \times T_2}$ 。 C_2 由 H_2 经过一个 1×1 卷积层得到。相似地，我们可以通过 H_1 、 H_3 得到 $X_1 \in \mathbb{R}^{N_1 \times D}$ 、 $X_3 \in \mathbb{R}^{N_3 \times D}$ 。

(3) 场景空间图卷积：当把不同尺度的特征都从帧空间映射到场景空间之后，我们使用图卷积（graph convolution）^[261] 里编码不同场景间的关系。具体来说，我们将所有的 N_{total} 节点 ($N_{total} = N_1 + N_2 + N_3$) 看成一个全连接图，然后利用图卷积进行参数更新：

$$\mathbf{Y} = ((\mathbf{I} - \mathbf{A}_{adj}) \mathbf{X}) \mathbf{W}, \quad (6-3)$$

其中， $\mathbf{X} = [X_1; X_2; X_3] \in \mathbb{R}^{N_{total} \times D}$ 是场景空间所有节点特征的集合，[:] 表示矩

阵按行连接符, $\mathbf{W} \in \mathbb{R}^{D \times D}$ 是可学习的映射矩阵, $\mathbf{A}_{adj} \in \mathbb{R}^{N_{total} \times N_{total}}$ 是可学习的邻接矩阵, \mathbf{I} 是大小与 \mathbf{A}_{adj} 相同的单位矩阵。

(4) 场景空间到帧空间: 给定场景空间特征 $\mathbf{Y} = [\mathbf{Y}_1; \mathbf{Y}_2; \mathbf{Y}_3]$, 我们重新将特征从场景空间映射回帧空间。以 \mathbf{Y}_2 为例:

$$\tilde{\mathbf{h}}_i^2 = \mathbf{d}_i^2 \mathbf{Y}_2 = \sum_j d_{ij}^2 y_j^2, \quad (6-4)$$

其中 $\mathbf{D}_2 = [\mathbf{d}_1^2, \dots, \mathbf{d}_{T_2}^2]$, $\mathbf{D}_2 \in \mathbb{R}^{T_2 \times N_2}$ 。为了减少模型参数量, 我们设定 $\mathbf{C}_i = \mathbf{D}_i^T$ 。因此, 我们可以得到增强的特征序列 $\{\tilde{\mathbf{H}}_1, \tilde{\mathbf{H}}_2, \tilde{\mathbf{H}}_3\}$ 。最后, 通过增大 $\tilde{\mathbf{H}}_1$ 和 $\tilde{\mathbf{H}}_2$ 的长度, 并将所有的特征连接起来, 得到输出特征 $\tilde{\mathbf{H}} \in \mathbb{R}^{T_1 \times D}$ 。

6.2.3 密集型头网络

与稀疏型自底向上模型不同, GDP 将起始时刻和终止时刻之间的每一帧都看成是正样本。对于每一帧, GDP 分别使用两个分支网络 (如图 6-7) :

(1) 回归分支网络 (regression subnet): 对于每一帧, 回归分支网络分别预测当前帧位置到两侧边界时刻的距离。给定图特征金字塔层的输出 $\tilde{\mathbf{H}}$, 回归分支网络使用四个通道数为 D 的 1×3 的卷积层和一个通道数为 1 的 1×3 卷积层。最后用非线性激活函数 sigmoid 输出预测的两侧边界距离。对于回归分支网络, 我们只对正样本计算预测损失。对于第 i 帧, 假设人工标注的结果为 (t_s, t_e) (即: $t_s \leq i \leq t_e$), 回归分支网络的预测目标为:

$$l_i^* = i - t_s, \quad r_i^* = t_e - i, \quad (6-5)$$

其中 l_i^* 和 r_i^* 分别表示从第 i 帧到左右两侧边界的距离。

(2) 置信分支网络 (confidence subnet): 虽然每一帧都有独立的边界预测, 但是不同帧的预测置信度往往不同。例如, 离起始时刻比较近的帧预测起始时刻的置信度通常比预测终止时刻的置信度要高。基于这一观察, 我们认为中心帧对两个边界预测的综合置信度最高。因此, 我们将置信分支的目标设为:

$$s_i^* = \begin{cases} \frac{\min(l_i^*, r_i^*)}{\max(l_i^*, r_i^*)}, & t_s \leq i \leq t_e \\ 0, & i < t_s \text{ or } i > t_e \end{cases} \quad (6-6)$$

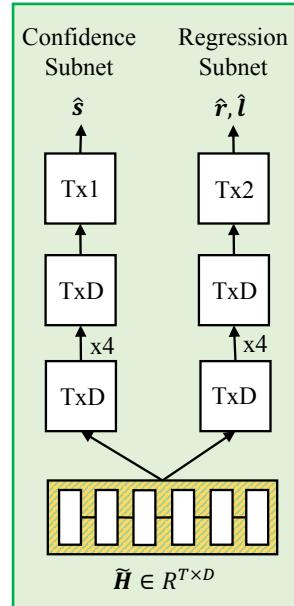


图 6-7 密集型头网络

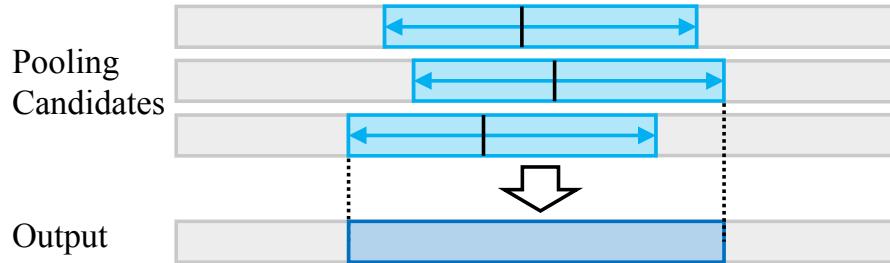


图 6-8 时域池化示意图

6.2.4 训练阶段和测试阶段

损失函数: 给定所有特征序列帧两个分支网络的预测 $\{(\hat{t}_i, \hat{s}_i)\}_{i=1}^T$ 和相应人工标注的目标 $\{(t_i^*, s_i^*)\}_{i=1}^T$, 整个 GDP 模型的训练损失函数为:

$$L = \frac{1}{T} L_{conf}(\hat{s}_i, s_i^*) + \frac{1}{T_p} \mathbf{1}_{\{s_i^* > 0\}} L_{reg}(\hat{t}_i, t_i^*), \quad (6-7)$$

其中, T 和 T_p 分别表示总样本和正样本的数量, $\mathbf{1}_{\{s_i^* > 0\}}$ 为指示函数, 当 $s_i^* > 0$ 时值为 1, 否则值为 0。置信分支网络的分类损失函数 L_{conf} 为是二值化交叉熵, 回归分支网络的回归损失函数 $L_{reg}(\hat{t}_i, t_i^*) = L_{l1}(\hat{t}_i, t_i^*) + L_{IoU}(\hat{t}_i, t_i^*)$, 其中 L_{l1} 是平滑的 l_1 损失函数, L_{IoU} 是 IoU 损失函数 (即: $-\ln \frac{\min(\hat{r}_i, r_i^*) - \max(\hat{l}_i, l_i^*)}{\max(\hat{r}_i, r_i^*) - \min(\hat{l}_i, l_i^*)}$)。

测试阶段: 对于每一帧, 我们可以得到单独的置信分数和边界预测结果。一种简单的方法就是直接选择置信分数最高的帧的边界预测结果作为最终视频片段检索结果。然而, 我们从实验中发现, 这种直接选择容易造成预测结果存在较大的方差。为了缓解这种问题, 我们使用了一种简单的时域池化来融合多个预测结果。如图 6-8 所示, 我们使用所有候选帧的最左侧预测帧和最右侧预测帧分别作为最终的起始时刻和终止时刻预测。至于候选帧的选择, 需要同时满足两个条件: (1) 预测的视频片段和置信度最高的视频片段有重叠部分; (2) 片段的置信度大于最高置信度乘以一个预先设定的阈值 δ (其中超参数 $\delta \in \{0.1, 0.2, \dots, 0.9\}$ 通过实验对不同的数据集进行不同选择)。

6.3 实验设置与性能对比

6.3.1 视频片段检索数据集和评价指标

基于语句查询的视频片段检索数据集: 我们在以下三个基于语句查询的视频片段检索数据集上对模型 GDP 的性能进行评估:

TACoS^[244]: 它一共包含 127 个视频和 17344 个文本与视频序列对。我们参考 Gao 等人^[23] 的数据集划分，将其中 50% 的样本作为训练集，25% 的样本作为验证集，25% 的样本作为测试集。在数据集 TACoS 中，每个样本中视频的平均长度为 5 分钟。

Charades-STA^[23]: 它一共包含 12408 个文本与视频序列对作为训练集，3720 个文本与视频序列对作为测试集。在数据集 Charades-STA 中，每个样本中视频的平均长度为 30 秒。

ActivityNet Captions^[245]: 它是目前为止数据集最大、最丰富的数据集，一共包含 19209 个视频。我们参考 Yuan 等人^[139] 的数据集划分，将其中 37421 个文本与视频序列作为训练集，17505 个文本与视频序列作为测试集。在数据集 ActivityNet Captions 中，每个样本中视频的平均长度为 2 分钟。

基于视频查询的视频片段检索数据集: 我们在一个基于视频查询的视频片段检索数据集上对模型 GDP 的性能进行评估：

ActivityNet-VRL^[142]: 它是目前唯一公开发布的基于视频查询的视频片段检索数据集。通过对动作识别数据集 ActivityNet^[262] 中 200 个类别的视频进行了重组，任意选取 160 个类别对应的视频作为训练集，20 个类别对应的视频作为验证集，和剩余 20 个类别对应的视频作为测试集。这种零样本式的数据集划分能够充分评估模型的泛化能力。在训练阶段，查询和视频对是随机组合的。在测试阶段，查询和视频对是固定的。

评价指标: 参考现有的工作，对于上述两种视频片段检索任务，我们分别使用下列三种通用的评价指标：

R@N, IoU@θ: 在测试集中，每个样本中预测分数最高的 n 个视频片段的交并比 (IoU) 大于阈值 θ 的百分比。基于自底向上模型的特性，我们仅对比 $N = 1$ 时的实验结果。

mIoU: 测试集中所有测试样本预测分数最高的视频片段的平均交并比。

mAP@1: 在不同阈值下，测试集中所有测试样本预测分数最高的视频片段的平均精度均值 (mAP)。

6.3.2 实验设定

给定一个视频序列 \mathcal{V} ，我们首先对视频进行下采样，并且用在 Sports-1M 数据集^[4] 上预训练好的 C3D 网络提取特征^[263]，然后利用 PCA 将特征维度降低到 500 维作为初始的视频特征。当查询为自然语句时，我们先将语句的最大长度设为 15，

Query: A bald man is shaving the back of another man's head.

Ref:



GT:

0.0s |-----> 13.6s

Score \hat{s} :

0.0s |-----> 14.9 s

Query: A little girl is standing in a kitchen doing dishes.

Ref:



GT:

3.4s |-----> 43.9s

Score \hat{s} :

2.1s |-----> 42.8 s

Query:



Ref:



GT:

46.8s |-----> 48.7s

Score \hat{s} :

46.6s |-----> 49.2s

Query:



Ref:



GT:

128.7s |-----> 138.7s

Score \hat{s} :

127.6s |-----> 139.2s

图 6-9 GDP 模型分别在 ActivityNet Captions 和 Activity-VRL 的检索结果

然后每个单词使用 300 维的 GloVe 向量^[50]作为编码向量的初始化，然后利用一个可学习的映射矩阵将维度也映射到 500 维。当查询为视频片段时，我们使用和之前参考视频同样的预处理。中间所有层的维度都设为 128，并且节点数 N_1 、 N_2 和 N_3 分别设为 10。这个网络利用优化算法 Adam^[264] 对模型进行优化。整个模型在数据集上训练 100 个周期，初始的学习率设为 0.0001，训练的批处理大小设为 16。

6.3.3 视频片段检索性能对比

基于语句查询的视频片段检索：在本节，我们将本章提出的 GDP 模型与目前最先进的基于语句查询的视频片段检索方法进行对比。按照自顶向下和自底向上框架划分，这些方法可以分为三类：（1）自顶向下模型：**VSA-RNN**^[23]、**VSA-**

	Method	TACoS			Charades-STA			ActivityNet Captions		
		IoU@0.1	IoU@0.3	mIoU	IoU@0.3	IoU@0.5	IoU@0.7	IoU@0.3	IoU@0.5	mIoU
TD	VSA-RNN	8.84	6.91	-	-	10.50	4.32	-	-	-
	VSA-STV	15.01	10.77	-	-	16.91	5.81	-	-	-
	CTRL	24.32	18.32	-	-	23.63	8.89	-	-	-
	ROLE	-	-	-	25.26	12.12	-	-	-	-
	ACRN	24.22	19.52	-	-	-	-	-	-	-
	MCF	25.84	18.64	-	-	-	-	-	-	-
	TGN	-	-	-	-	-	-	43.81	27.93	-
	ACL	28.31	22.07	-	-	26.47	11.23	-	-	-
	SAP	31.15	-	-	-	27.42	13.36	-	-	-
RL	QSPN	-	-	-	54.70	35.60	15.80	45.30	27.70	-
	RWM	-	-	-	-	36.70	-	-	36.90	-
	SM-RL	26.51	20.25	-	-	24.36	11.17	-	-	-
BU	L-NET	-	-	13.41	-	-	-	-	-	-
	ABLR-aw	31.60	18.90	12.50	-	-	-	53.65	34.91	35.72
	ABLR-af	34.70	19.50	13.40	-	-	-	55.67	36.79	36.99
	GDP	39.68	24.14	16.18	54.54	39.47	18.49	56.17	39.27	39.80

表 6-1 不同基于语句查询的视频片段检索方法的性能对比

mAP@1	0.5	0.6	0.7	0.8	0.9	Avg
Frame-level	18.8	13.9	9.6	5.0	2.3	9.9
Video-level	24.3	17.4	12.0	5.9	2.2	12.4
SST	33.2	24.7	17.2	7.8	2.7	17.1
CGBM	43.5	35.1	27.3	16.2	6.5	25.7
GDP	44.0	35.4	27.7	20.0	12.1	27.8

表 6-2 不同基于视频查询的视频片段检索方法的性能对比

STV^[23]、**CTRL**^[23]、**ROLE**^[136]、**ACRN**^[135]、**MCF**^[265]、**TGN**^[137]、**ACL**^[249]、**SAP**^[250]和**QSPN**^[251]。(2) 基于强化学习的模型：**RWM**^[140]、**SM-RL**^[141]。(3) 自底向上模型：**L-Net**^[138]、**ABLR-af**、**ABLR-aw**^[139]。

表 6-1 展示了 GDP 模型的定量实验结果。从表 6-1 可以看出，GDP 在所有的数据集下都可以达到目前最好的性能。尤其对于更加严格的评价指标，性能增益往往更加明显。例如，在数据集 TACoS 和 ActivityNet Captions 中，mIoU 可以分别提升 2.77% 和 2.81%；在数据集 Charades-STA 中，IoU@0.7 可以提升 2.69%。

图 6-9 展示了 GDP 模型的定性实验结果。由图 6-9 可以看出，置信度最高的视频帧往往都接近于人工标注的检索片段中心位置，这也符合我们的模型设计初衷，

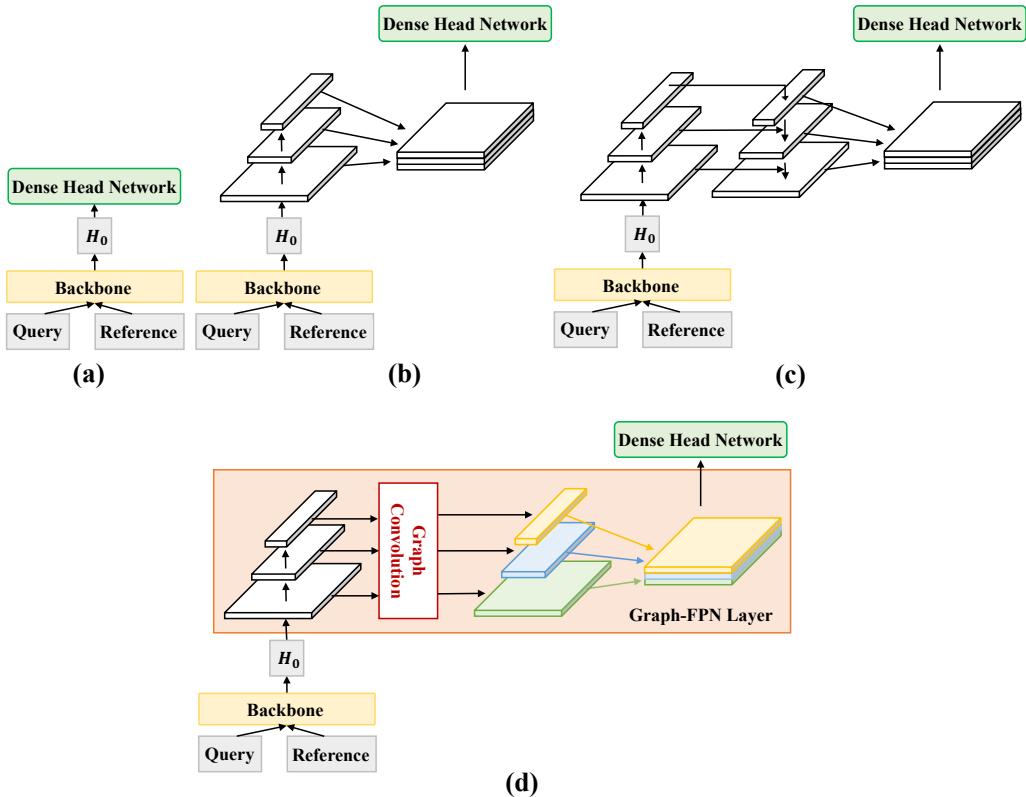


图 6-10 同一骨干网络不同特征优化层的性能对比

将检索片段中心帧的置信分数设为最高。

基于视频查询的视频片段检索：在本节，我们将本章提出的 GDP 模型与目前最先进的基于视频查询的视频片段检索方法进行对比。同样，我们可以将这些方法分为两类：(1) 自顶向下模型：**Frame-level**^[142]、**video-level**^[142]、**SST**^[266]。(2) 自底向上模型：**CGBM**^[142]。

从表 6-2 可以看出，GDP 模型在所有的评价指标下都超过目前最好的模型。尤其对于高质量的预测，性能增益更加明显。例如：在 tIoU 阈值为 0.9 时，GDP 的性能相比提升接近 100%。

6.3.4 视频片段检索性能分析

图特征金字塔层的有效性：为了验证图特征金字塔层的有效性，我们设计了三种基准模型进行对比。如图 6-10 所示，模型 A (a) 包含一个骨干网络和一个密集头网络；模型 B (b) 用骨干网络的输出特征序列构建一个特征金字塔，然后直接将不同尺度的特征直接进行拼接融合；模型 C (c) 参考 FPN (Feature Pyramid Network)^[259] 使用一个自顶向下的支路网络将不同尺度的特征进行融合；模型 D (d) 为 GDP 模型。其中，四个对比模型中的骨干网络和密集头网络结构都完全相同。

Model	TACoS			Charades-STA			ActivityNet Captions		
	IoU@		mIoU	IoU@		mIoU	IoU@		mIoU
	0.1	0.3		0.3	0.5		0.1	0.3	
A	37.4	23.3	11.5	15.3	51.8	38.3	17.8	35.1	72.1
B	37.3	23.1	13.9	15.8	53.8	38.6	18.4	36.0	56.0
C	36.8	23.1	13.8	15.7	52.6	38.9	18.3	35.8	40.3
D	39.7	24.1	13.5	16.2	54.5	39.5	18.5	36.6	39.3
							75.0	56.2	39.8

表 6-3 基于语句查询的视频片段检索任务中模型 A、B、C、D 的性能对比

mAP@1	0.5	0.6	0.7	0.8	0.9
A	41.1	34.2	27.7	20.3	6.8
B	43.3	35.0	27.9	18.2	9.6
C	42.9	34.5	26.9	18.8	8.4
D	44.0	35.4	27.7	20.0	12.1

表 6-4 基于视频查询的视频片段检索任务中模型 A、B、C、D 的性能对比

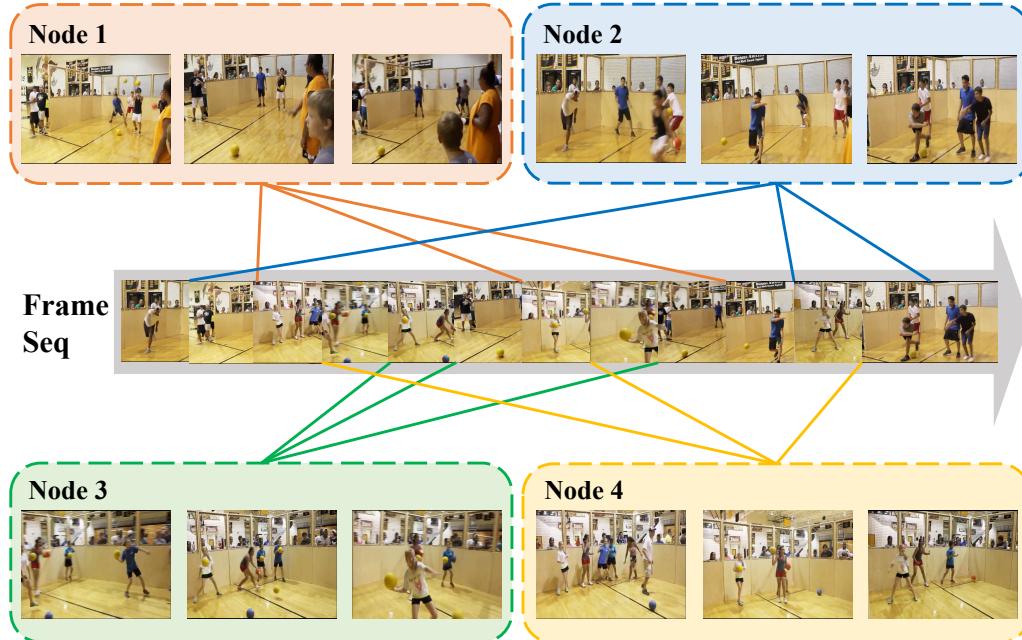


图 6-11 场景空间中的节点可视化

基于语句查询和视频查询的视频片段检索的结果分别展示在表 6-3 和表 6-4 中。根据实验结果，我们可以发现：(1) 特征金字塔结构可以显著地提升视频片段检索任务的性能（即：模型 B、C、D 的性能明显好于模型 A 的性能）。(2) 利用自顶向下的支路将连续两个尺度的特征进行融合并不能有效地融合多尺度特征。例如，模

Dataset	Metric	Sparse	Dense*	Dense
TACoS	IoU@0.1	32.3	36.5	39.7
	IoU@0.3	18.7	22.9	24.1
	IoU@0.5	9.6	13.0	13.5
	mIoU	12.9	15.2	16.2
Charades-STA	IoU@0.3	52.9	53.9	54.5
	IoU@0.5	31.4	39.0	39.5
	IoU@0.7	14.7	18.3	18.5
	mIoU	35.1	36.1	36.6
ActivityNet Captions	IoU@0.1	72.4	73.5	75.0
	IoU@0.3	53.0	55.9	56.2
	IoU@0.5	37.5	39.8	39.3
	mIoU	39.0	39.3	39.8
ActivityNet-VRL	tIoU@0.5	41.6	42.3	44.0
	tIoU@0.6	30.5	35.3	35.4
	tIoU@0.7	25.7	27.6	27.7
	tIoU@0.8	19.8	20.6	20.0
	tIoU@0.9	8.5	12.5	12.1
	Average	25.2	27.7	27.8

表 6-5 密集型头网络和稀疏型头网络的性能对比

型 B 和模型 C 的性能十分接近，即自顶向下支路和直接拼接效果接近。(3) GDP 模型（模型 D）在绝大多数的数据集和评价指标下都能取得最好的结果，说明了图特征金字塔层对视频检索任务的有效性。

密集头网络的有效性：为了评估密集头网络的有效性，我们设计了一个基准模型：它使用和 GDP 完全相同的骨干网络和图卷积特征层，然后只是将密集头网络替换成稀疏型头网络，即直接预测两端边界的概率。

基于语句查询和视频查询的视频片段检索的结果展示在表 6-5 中，其中 Sparse 表示基准模型（稀疏型头网络），Dense* 表示模型在测试阶段没有使用时域池化。由表 6-5 可以看出，密集头网络可以显著提升模型性能。尤其对于长视频（如：数据集 TACoS），性能提升更加明显。这也间接说明密集头网络能够缓解稀疏头网络中存在的正负训练样本极度不均等问题。

场景空间节点的可视化：在图 6-11 中，我们随机选取了同一个尺度下的四个节

点，然后每个节点选取了三个视频帧来表示节点信息。如图 6-11 所示，同一个节点中的视频帧通常包含一个特定的视觉场景或相似的场景内容。

6.4 本章小结

在本章，我们深入分析了目前视频片段检索方法的优缺点，并针对稀疏型自底向上框架的设计缺点，提出了一种全新的密集型自底向上网络：基于图特征金字塔的密集型预测（GDP）。GDP 通过引入一个图特征金字塔层来充分编码视频中多个场景之间的内在联系，提升特征帧序列的表达能力。同时，GDP 将目前的稀疏头网络替换成密集头网络，不仅可以避免训练过程中正负样本极度不均的问题，同时可以将动作边界预测问题分解成相关性预测和动作边界回归两个子问题，简化视频片段检索的难度。在两种不同的查询形式下（基于自然语句和视频片段）的视频片段检索任务中，GDP 模型都能取得目前最好的性能。

7 基于反事实样本生成的视觉问答方法

随着计算机视觉和自然语言处理等技术的发展，视觉问答技术在过去的十年间取得了长足的进步。然而，目前的视觉问答模型仍然过于依赖问题与答案之间的联系（即文本偏置）。为了缓解这一问题，近期的一些视觉问答模型开始通过引入一个辅助网络来约束视觉问答模型的训练。这种复合模型在标准视觉问答数据集 VQA-CP 上取得了目前最好的实验性能。但是，由于这类复合模型设计的复杂性，它们往往难以满足一个理想视觉问答模型应该具备的两个特性：(1) 视觉可解释性 (visual-explainable ability)：在预测答案时，模型应该基于正确的视觉图像区域。(2) 问题敏感性 (question-sensitive ability)：在改变问题时，模型应该能够感知问题的变化，调整相应的预测结果。因此，在本章，我们提出了一种通用的反事实样本生成机制 (Counterfactual Samples Synthesizing, CSS)。CSS 通过遮盖图像中的重要区域或问题中的重要单词来生成新的反事实训练样本，并且对这些反事实样本赋予不同的标准答案。通过使用原始训练样本和新生成的反事实样本一起对模型进行训练，迫使模型能够关注被遮盖的重要区域和重要单词，提升模型的视觉可解释性和问题敏感性。另一方面，由于模型的视觉可解释性和问题敏感性得到了提升，模型的视觉问答性能可以被进一步提升。在视觉问答数据集 VQA-CP 中，我们通过大量的对比实验证明了 CSS 的有效性。值得注意的是，当对目前最好的视觉问答模型 LMH^[173] 使用 CSS 机制后，模型在数据集 VQA-CP v2 上可以提升 6.5% 的准确率，达到破纪录的 58.95%。

7.1 问题描述

视觉问答 (Visual Question Answering, VQA) 是视觉场景推理中一个重要的任务，也是人类迈向真正人工智能时代至关重要的一步。视觉问答任务是给定一个图像和一个关于图像场景的问题，模型需要在充分理解图像和问题内容之后给出答案预测。随着大量视觉问答数据集的出现（如：VQA v1^[24]、VQA v2^[167] 等），视觉问答任务在近年来取得了前所未有的关注。然而，由于在真实图像数据集的收集

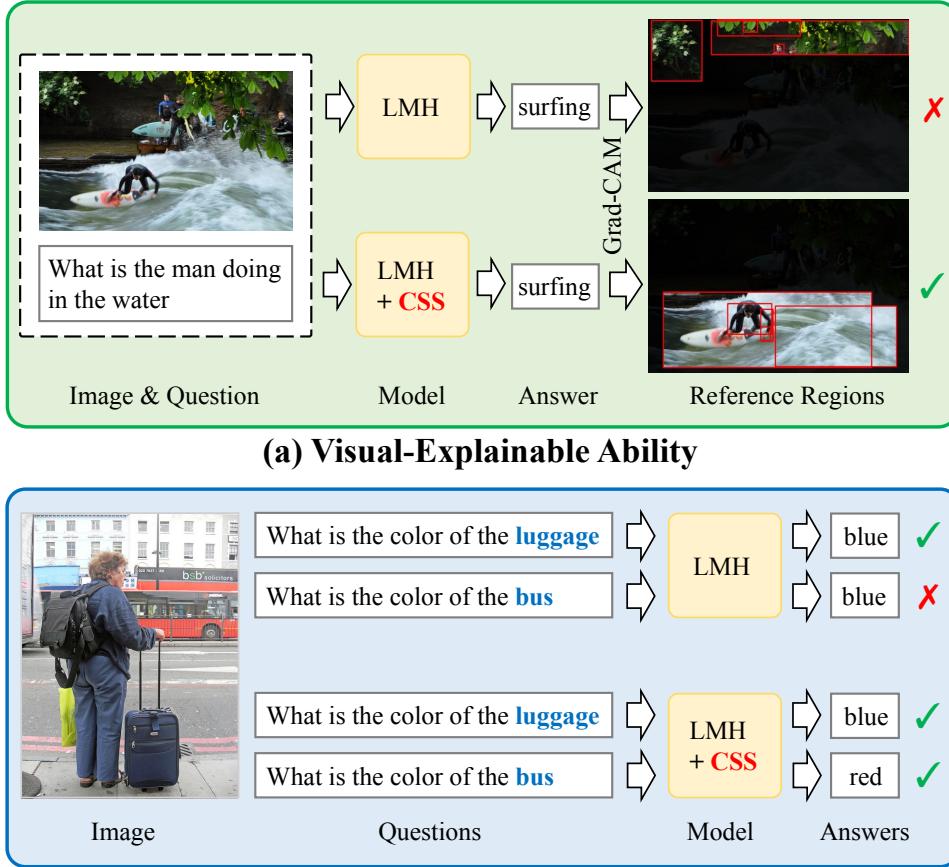


图 7-1 VQA 模型的视觉可解释性和问题敏感性

和标注过程中会不可避免地会引入文本偏置，并且目前的视觉问答模型往往过于依赖这种文本偏置（即问题和答案之间的直接联系）^[165–167,267]。例如，对于所有以“*How many*”开头的问题，模型都回答“2”就可以在数据集评估中得到较好的实验结果。为了尽可能地去除这种文本偏置对视觉问答模型的评估，Agrawal 等人^[168]提出了一个新的数据集 VQA-CP (VQA under Changing Priors)。VQA-CP 通过故意让训练集和测试集中问题和答案的分布不同，让过于依赖文本偏置的视觉问答模型无法在测试集中得到较好的实验结果。大量的视觉问答模型在 VQA-CP 数据集上都出现了明显的性能下降^[120,122,153,268]。

目前，对于数据集 VQA-CP，性能最好的方法是基于复合模型的设计：通过引入一个辅助网络来约束视觉问答网络的训练，其中的辅助网络只使用问题作为输入信号。具体来说，这些复合模型可以细分为两小类：(1) 基于对抗学习的方法^[169–171]：它们使用对抗学习^[65]的方式来训练这两个网络（视觉问答网络和辅助网络），即减少视觉问答网络的损失函数，同时增大辅助网络的损失函数。因为这两个网络共享同一个文本编码器，所以这类基于对抗学习的方法本质上希望通过

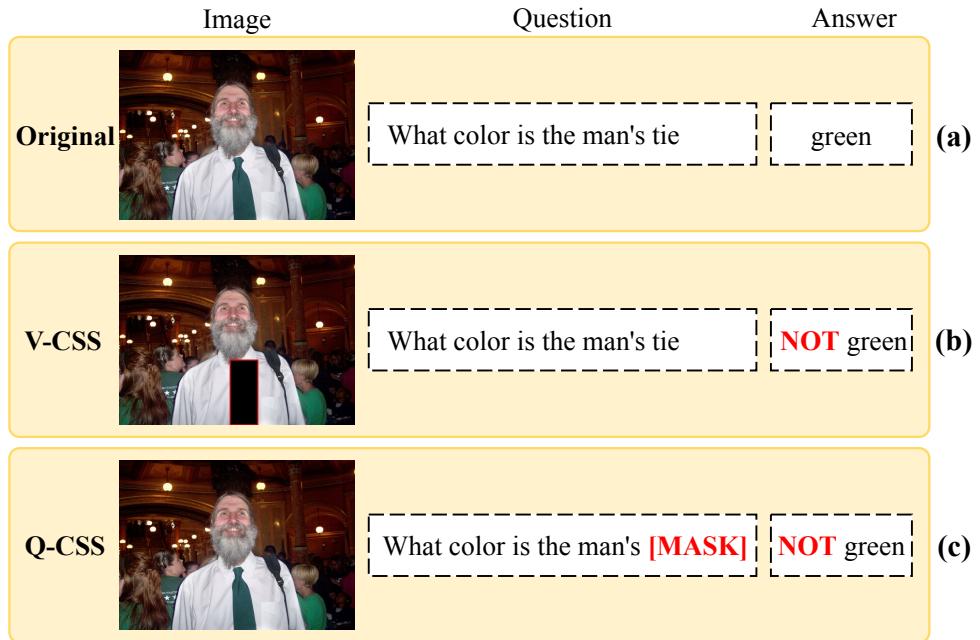


图 7-2 V-CSS 和 Q-CSS 示意图

学习到不含偏置信息的问题编码向量来缓解模型过于依赖文本偏置的问题。然而，Grand 等人^[170]通过实验表明，这类基于对抗学习的方法在计算梯度过程中引入大量噪声，容易造成训练过程的不稳定。（2）基于融合的方法^[172-174]：它们将两个网络预测的答案概率分布进行融合，然后根据融合后的概率分布来计算参数优化的梯度。这类方法的设计动机是让视觉问答网络更加关注到辅助网络无法回答的训练样本。

尽管这些复合模型在数据集 VQA-CP 上取得了最好的实验性能，但是由于目前复合模型这些复杂的网络设计，使得这类模型难以满足一个理想视觉问答模型应该具体的两个特性：（1）**视觉可解释性**：在预测答案时，模型应该关注正确的视觉图像区域^[269]。如图 7-1（a）所示，尽管两个模型（LMH 和 LMH+CSS）都可以预测出正确的答案“surfing”，但实际上这两个模型是根据完全不同的视觉区域做的决策。（2）**问题敏感性**：在改变问题时，模型应该能够感知问题的变化，调整相应的预测结果。如图 7-1（b）所示，当两个问题的句式结构十分接近时（如只替换“luggage”成“bus”），因为此时两个问题的意思已经完全不同，模型应该能够感知这两个问题的差异，对预测做出调整。

在本章，我们提出了一个全新的反事实样本生成机制（CSS），来提升视觉问答模型的视觉可解释性和问题敏感性。如图 7-2 所示，CSS 包含两种不同的样本生成机制：V-CSS 和 Q-CSS。对于 V-CSS 而言，我们通过遮盖图像中的重要区域，得

到反事实图像。所谓的“重要区域”是指针对回答该问题时模型需要依据或者关注的图像区域。这些反事实图像和原始的问题就组成了一个新的图像问题对。对于 Q-CSS 而言，我们通过将原始问题中的重要单词替换成一个特定的词 “[MASK]”，得到反事实问题。同样，这些反事实问题和原始的图像也组成了一个新的图像问题对。除了图像问题对，一个标准的视觉问答训练样本还需要相应的标准答案。为了避免大量的人工标注，我们设计了一个动态的答案生成机制，对所有新合成的图像问题对分配合理的“标准”答案（如：图 7-2 中的“not green”）。最后，通过使用原始训练样本和新生成训练样本一起对视觉问答模型训练，迫使模型更加关注被遮盖的重要区域和重要单词。

我们通过大量的定性和定量实验都验证了 CSS 机制的有效性。CSS 可以无缝地用在任何视觉问答模型中，包括复合模型。它不仅提升可以提升模型的视觉可解释性和问题敏感性，同时可以进一步提升模型在 VQA-CP 上的准确率。值得注意的是，当对目前最好的视觉问答模型 LMH^[173] 使用 CSS 机制后，模型在数据集 VQA-CP v2 上可以达到 58.95% 的准确率。

7.2 反事实样本生成

对于视觉问答任务，我们将其看成一个多类别分类任务。为了不失一般性，给定一个视觉问答数据集 $\mathcal{D} = \{I_i, Q_i, a_i\}_i^N$ 包含大量的图像 $I_i \in \mathcal{I}$ 、问题 $Q_i \in \mathcal{Q}$ 和答案 $a_i \in \mathcal{A}$ ，视觉问答任务在于学习一个映射函数 $f_{vqa} : \mathcal{I} \times \mathcal{Q} \rightarrow [0, 1]^{\lvert \mathcal{A} \rvert}$ ，可以对任意的图像问题对预测一个答案概率分布。为了后续表达的简洁性，我们在表述中省略下角标 i 。

在本节，我们先介绍视觉问答中的自底向上和自顶向下模型^[122]，以及复合模型。然后，我们再介绍 CSS 的具体细节。

7.2.1 引言

自底向上和自顶向下模型 (Bottom-Up Top-Down, UpDn)：对于每张图像 I ，UpDn 模型使用图像编码器 e_v 将图像编码成一系列物体特征： $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_{n_v}\}$ ，其中 \mathbf{v}_i 表示第 i 个物体特征。对于每个问题 Q ，UpDn 模型使用问题编码器 e_q 编码成一系列单词特征： $\mathbf{Q} = \{\mathbf{w}_1, \dots, \mathbf{w}_{n_q}\}$ ，其中 \mathbf{w}_j 表示第 j 个单词的特征。然后将特征 \mathbf{V} 和 \mathbf{Q} 同时输入到模型 f_{vqa} 中来预测答案的概率分布：

$$P_{vqa}(\mathbf{a}|I, Q) = f_{vqa}(\mathbf{V}, \mathbf{Q}) \quad (7-1)$$

算法 7.1 复合模型（基于融合的方法）

```

1: function  $\mathcal{VQA}(I, Q, a, cond)$ 
2:    $\mathbf{V} \leftarrow e_v(I)$ 
3:    $\mathbf{Q} \leftarrow e_q(Q)$ 
4:    $P_{vqa}(\mathbf{a}) \leftarrow f_{vqa}(\mathbf{V}, \mathbf{Q})$ 
5:    $P_q(\mathbf{a}) \leftarrow f_q(\mathbf{Q})$                                  $\triangleright$  辅助模型
6:    $\hat{P}_{vqa}(\mathbf{a}) \leftarrow M(P_{vqa}(\mathbf{a}), P_q(\mathbf{a}))$ 
7:    $Loss \leftarrow \text{XE}(\hat{P}_{vqa}(\mathbf{a}), a)$                        $\triangleright$  参数更新
8:   if  $cond$  then
9:     return  $\mathbf{V}, \mathbf{Q}, P_{vqa}(\mathbf{a})$ 
10:  end if
11: end function

```

其中，模型 f_{vqa} 通常包含一个注意力机制，然后利用答案分类的交叉熵作为损失函数对模型进行优化。

复合模型：根据前文（第 7.1 节）讨论，复合模型可以细分为两小类：基于对抗学习的方法和基于融合的方法。因为基于对抗学习的方法^[169–171]训练过程不稳定同时性能相对较低。在本节，我们只介绍基于融合的方法^[172–174]。具体如算法 7.1 所示，它们通过引入一个辅助网络 f_q ，其中 f_q 只使用问题 \mathbf{Q} 做为输入直接预测答案概率：

$$P_q(\mathbf{a}|Q) = f_q(\mathbf{Q}) \quad (7-2)$$

然后，它们通过一个融合函数 M 将两个网络预测的答案概率分布进行融合，得到：

$$\hat{P}_{vqa}(\mathbf{a}|I, Q) = M(P_{vqa}(\mathbf{a}|I, Q), P_q(\mathbf{a}|Q)) \quad (7-3)$$

在训练阶段，它们根据融合后的答案概率分布 $\hat{P}_{vqa}(\mathbf{a})$ 来计算交叉熵损失，进而计算梯度并同时优化网络 f_{vqa} 和网络 f_q 。在测试阶段，它们只使用网络 f_{vqa} 的预测结果作为最终结果。

7.2.2 反事实样本生成

整个反事实样本生成机制 (CSS) 的流程展示在算法 7.2 上。具体来说，对于给定的一个视觉问答模型（即 \mathcal{VQA} ）和训练样本 (I, Q, a) ，CSS 主要包含三个步骤：

1. 用原始的训练样本对 \mathcal{VQA} 模型进行训练；

算法 7.2 反事实样本生成

```

1: function  $\mathcal{CSS}(I, Q, a)$ 
2:    $\mathbf{V}, \mathbf{Q}, P_{vqa}(\mathbf{a}) \leftarrow \mathcal{VQA}(I, Q, a, \text{True})$ 
3:    $cond \sim U[0, 1]$ 
4:   if  $cond \geq \delta$  then                                 $\triangleright$  执行 V-CSS
5:      $\mathcal{I} \leftarrow \text{IO\_Sel}(I, Q)$ 
6:      $s(a, \mathbf{v}_i) \leftarrow \mathcal{S}(P_{vqa}(a), \mathbf{v}_i)$ 
7:      $I^+, I^- \leftarrow \text{CO\_Sel}(\mathcal{I}, \{s(a, \mathbf{v}_i)\})$ 
8:      $a^- \leftarrow \text{DA\_Ass}(I^+, Q, \mathcal{VQA}, a)$ 
9:      $\mathcal{VQA}(I^-, Q, a^-, \text{False})$ 
10:  else                                          $\triangleright$  执行 Q-CSS
11:     $s(a, \mathbf{w}_i) \leftarrow \mathcal{S}(P_{vqa}(a), \mathbf{w}_i)$ 
12:     $Q^+, Q^- \leftarrow \text{CW\_Sel}(\{s(a, \mathbf{w}_i)\})$ 
13:     $a^- \leftarrow \text{DA\_Ass}(I, Q^+, \mathcal{VQA}, a)$ 
14:     $\mathcal{VQA}(I, Q^-, a^-, \text{False})$ 
15:  end if
16: end function

```

2. 用 V-CSS 生成反事实样本 (I^-, Q, a^-) 或用 Q-CSS 生成反事实样本 (I, Q^-, a^-) ;
3. 用新生成的反事实样本对 \mathcal{VQA} 模型进行训练。

在接下来，我们先详细介绍两种反事实样本生成机制：V-CSS 和 Q-CSS。如算法 7.2 所示，对于每个训练样本，我们只使用特定的一种样本生成机制，其中 δ 是一个超参数来权衡两种生成机制的比例。

V-CSS：我们将按照算法 7.2 中程序执行的顺序依次介绍 V-CSS 的主要步骤，包括初始物体选择 (IO_Sel)、物体局部贡献计算、重要物体选择 (CO_Sel) 和动态答案生成 (DA_Ass)：

(1) 初始物体选择 (IO_Sel)：对于任意一个图像问题对 (Q, a) ，通常只有图像中部分物体与当前问题有关。为了缩小重要物体的选择范围，我们首先构建一个小小的物体集 \mathcal{I} ，使得 \mathcal{I} 中的物体都可能与当前问题有关。由于数据集中缺乏人工的标注信息，我们参考 Wu 等人^[179]，认为问题和答案中出现的名词往往与当前问题有关。具体来说，我们先用 spaCy POS tagger^[270] 提取出问题和答案中的所有名词，然后利用这些名词的 GloVe 编码^[50] 与图像中所有物体对应类别的 GloVe 编码计算余

弦相似度，其中相似度记为 SIM 。我们根据相似度 SIM 直接选择分数最高的 $|\mathcal{I}|$ 个物体组成物体集合 \mathcal{I} 。

(2) 物体局部贡献计算：当得到物体集 \mathcal{I} 之后，我们开始计算每个物体对最终的正确答案的贡献。我们参考现有的工作^[178,179,271]，同样使用修改版的 Grad-CAM^[180] 算法来计算局部贡献。对于第 i 个物体来说，它对正确答案 a 的贡献为：

$$s(a, \mathbf{v}_i) = \mathcal{S}(P_{vqa}(a), \mathbf{v}_i) := (\nabla_{\mathbf{v}_i} P_{vqa}(a))^T \mathbf{1} \quad (7-4)$$

其中 $P_{vqa}(a)$ 是对正确答案 a 的预测概率， \mathbf{v}_i 是第 i 个物体特征， $\mathbf{1}$ 是所有元素都为 1 的向量。显然，当分数 $s(a, \mathbf{v}_i)$ 越高时，物体特征 \mathbf{v}_i 对正确答案 a 的贡献也越大。

(3) 重要物体选择 (CO_Sel)：在对物体集 \mathcal{I} 中所有物体都计算局部贡献 $s(a, \mathbf{v}_i)$ 之后，我们将分数最大的前 K 个物体当成重要的物体，即 I^+ 。对于每张图像， K 是满足公式 (7-5) 的最小整数：

$$\sum_{\mathbf{v}_i \in I^+} \exp(s(a, \mathbf{v}_i)) / \sum_{\mathbf{v}_j \in \mathcal{I}} \exp(s(a, \mathbf{v}_j)) > \eta \quad (7-5)$$

其中 η 是一个超参数用于控制 K 的数量。在所有的实验中，我们将 η 设为 0.65。然后，反事实图像就是物体集 I^+ 在所有物体集 I 中的补集，即 $I^- = I \setminus I^+$ 。如图 7-3 所示， I^+ 和 I^- 时互补的两个物体集合。

(4) 动态答案生成 (DA_Ass)：给定反事实图像 I^- 和原始问题 Q ，我们可以组成一个新的图像问题对 (I^-, Q) 。作为一个视觉问答任务的训练样本，新的图像问题对同样需要分配“标准”答案。为了减少大量的人工标注，我们设计了一种动态的答案生成机制，来预测标准答案 a^- 。动态答案生成机制的细节在算法 7.3 中。具体来说，我们将另一个图像问题对 (I^+, Q) 输入到视觉问答模型 VQA 中，然后得到预测的答案概率分布 $P_{vqa}^+(a)$ 。根据概率 $P_{vqa}^+(a)$ ，我们选取前 N 个答案作为 a^+ ，然后定义 $a^- := \{a_i | a_i \in a, a_i \notin a^+\}$ 。它的设计理念就是，图像问题对 (I^-, Q) 的标准答案是 (I^+, Q) 遗漏的答案。在极端情况下，如果模型对于图像问题对 (I^+, Q) 可以预测全部的正确答案，即 $a \subset a^+$ ，则 a^- 为空集 \emptyset 。

Q-CSS：Q-CSS 的主要步骤和 V-CSS 非常接近，除了没有初始物体选择（即 IO_Sel）。按照算法 7.2 的执行顺序，Q-CSS 主要包含三步：单词局部贡献计算、重要单词选择 (CW_Sel) 和动态答案生成 (DA_Ass)：

(1) 单词局部贡献计算：与 V-CSS 中物体局部贡献计算相似（公式 (7-4)），对于第 i 个单词，它对正确答案 a 的贡献为：

$$s(a, \mathbf{w}_i) = \mathcal{S}(P_{vqa}(a), \mathbf{w}_i) := (\nabla_{\mathbf{w}_i} P_{vqa}(a))^T \mathbf{1} \quad (7-6)$$

算法 7.3 动态答案生成

```

1: function DA_Ass( $I^+, Q^+, \mathcal{VQA}, a$ )
2:    $\mathcal{VQA}.eval()$                                       $\triangleright$  不更新参数
3:    $\_, \_, P_{vqa}^+(\mathbf{a}) \leftarrow \mathcal{VQA}(I^+, Q^+, a, \text{True})$ 
4:    $a^+ \leftarrow \text{top-N}(\text{argsort}_{a_i \in \mathcal{A}}(P_{vqa}^+(a_i)))$ 
5:    $a^- := \{a_i | a_i \in a, a_i \notin a^+\}$             $\triangleright a$  是标准答案集
6:   return  $a^-$ 
7: end function

```

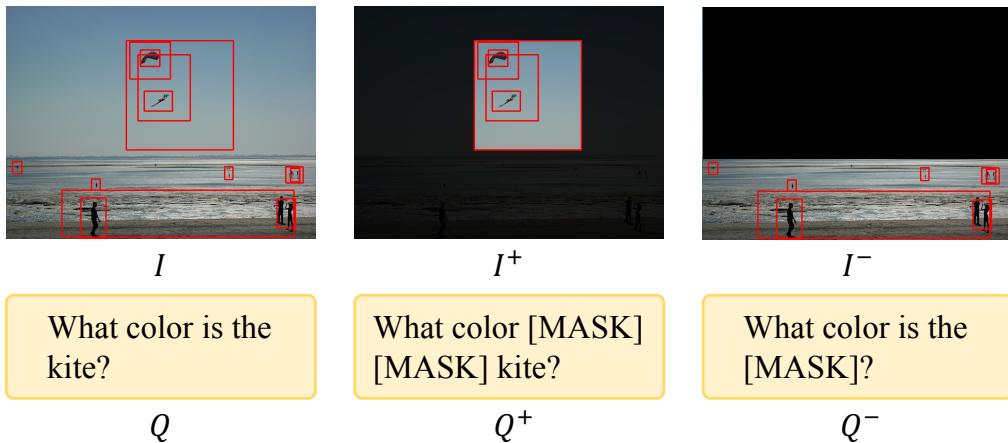


图 7-3 一个 I^+ 、 I^- 、 Q^+ 和 Q^- 的示例图

(2) 重要单词选择 (CW_Sel): 对于这步，我们首先提取每个问题 Q 的问题类别前缀，这些前缀直接使用数据集 VQA-CP 中默认的问题类别前缀。然后对剩余的单词选择贡献最高的 K 个单词作为重要单词，而反事实问句 Q^- 就是将问句 Q 中的重要单词替换成一个特殊字符 “[MASK]”。如图 7-3，当问句是 “what color is the kite” 以及重要单词是 “kite” 时，反事实问句 Q^- 为 “what color is the [MASK]”。

(3) 动态答案生成 (DA_Ass): 这个动态答案生成步骤和 V-CSS 的完全相同，即算法 7.3。唯一不同的是，对于 Q-CSS，DA_Ass 的输入是图像问题对 (I, Q^+) 。

7.3 实验设置与性能对比

我们主要在数据集 VQA-CP^[168] 的测试集上对模型的性能进行验证。为了实验比较的完整性，我们同样在 VQA v2^[167] 的验证集上进行测试。关于模型的准确率，我们使用通用的 VQA 准确率计算方式^[24]。为了实验比较的公平性，所有的实验都采取与 UpDn 模型^[122] 相同的数据预处理。

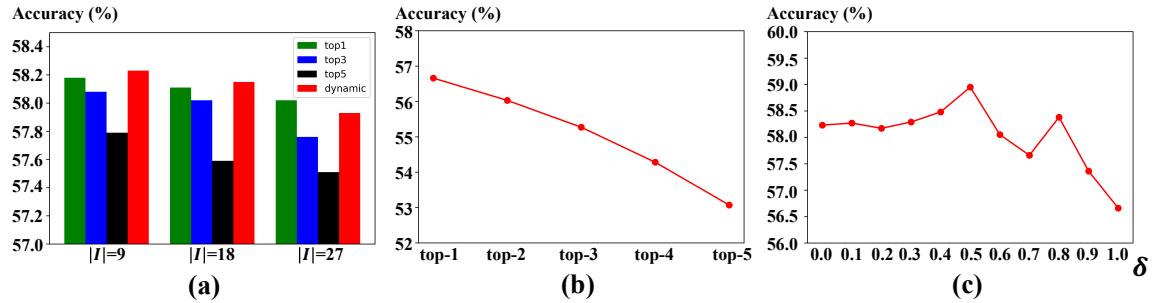


图 7-4 V-CSS 和 Q-CSS 中不同超参数对模型性能的影响

7.3.1 CSS 对视觉问答的性能分析

CSS 机制中不同超参数对模型性能的影响：我们通过大量的对比实验来分析 V-CSS 和 Q-CSS 中不同超参数对模型性能的影响。具体来说，所有的实验都是基于现有的视觉问答模型 LMH^[173]。实验结果展示在图 7-4 中。

(1) V-CSS 中初始物体集 \mathcal{I} 的大小：不同 $|\mathcal{I}|$ 对模型性能的影响如图 7-4 (a) 所示。随着 $|\mathcal{I}|$ 的增加，模型的性能逐渐降低。

(2) V-CSS 中重要物体选择的数量：不同重要物体选择的数量对实验性能的影响如图 7-4 (a) 所示。我们比较了动态选择 K (公式 7-5) 和预先固定一些常数 (如：1、3 或 5)。从实验结果可知，动态选择 K 个重要物体可以得到最佳性能。

(3) Q-CSS 中重要单词选择的数量：不同重要单词选择的数量对实验性能的影响如图 7-4 (b) 所示。从实验结果可知，只选择 1 个单词作为重要单词性能最好。

(4) V-CSS 和 Q-CSS 的比例 δ ：不同 δ 对实验性能的影响如图 7-4 (c) 所示。从实验结果可知， $\delta = 0.5$ 性能最好。

CSS 机制对模型的通用性：由于 CSS 机制的设计没有依赖任何具体的 VQA 模型设计，为了验证 CSS 机制对不同模型结构的通用性，我们将 CSS 机制应用于多种 VQA 模型中：**UpDn**^[122]、**PoE** (Product of Experts)^[173,174]、**RUBi**^[172] 和 **LMH**^[173]。其中模型 PoE、RUBi、LMH 都属于复合模型。所有的实验结果展示在表 7-1 中，[†] 表示我们自己的重现结果。

由表 7-1 中结果可以看出，CSS 可以提升多种不同的视觉问答模型的性能。尤其对于复合模型来说，性能的提升往往更加明显（例如：在 LMH 和 PoE 模型中准确率可以分别提升 6.5% 和 9.79%）。在通常情况下，相比于单独使用一种 CSS 机制 (V-CSS 或 Q-CSS)，同时使用两种 CSS 机制可以得到最佳性能。

		Model	All	Y/N	Num	Other
Plain Models	UpDn ^[122]	Baseline	39.74	42.27	11.93	46.05
		Baseline [†]	39.68	41.93	12.68	45.91
		+Q-CSS	40.05	42.16	12.30	46.56
		+V-CSS	40.98	43.12	12.28	46.86
		+CSS	41.16	43.96	12.78	47.48
Ensemble-Based Models	PoE ^[173,174]	Baseline	39.93	—	—	—
		Baseline [†]	39.86	41.96	12.59	46.25
		+Q-CSS	40.73	42.99	12.49	47.28
		+V-CSS	49.65	74.98	16.41	45.50
		+CSS	48.32	70.44	13.84	46.20
RUBi ^[172]	RUBi ^[172]	Baseline	44.23	—	—	—
		Baseline [†]	45.23	64.85	11.83	44.11
		+Q-CSS	46.31	68.70	12.15	43.95
		+V-CSS	46.00	62.08	11.84	46.95
		+CSS	46.67	67.26	11.62	45.13
LMH ^[173]	LMH ^[173]	Baseline	52.05	—	—	—
		Baseline [†]	52.45	69.81	44.46	45.54
		+Q-CSS	56.66	80.82	45.83	46.98
		+V-CSS	58.23	80.53	52.48	48.13
		+CSS	58.95	84.37	49.42	48.21

表 7-1 CSS 机制对不同 VQA 模型的性能影响

7.3.2 视觉问题方法性能对比

数据集 VQA-CP v2 和 VQA v2 上性能对比：我们将 CSS 机制应用于模型 LMH^[173] 中，然后将模型标记为 **LMH-CSS**。我们将 LMH-CSS 与其他目前性能最好的视觉问答模型在数据集 VQA-CP v2 和 VQA v2 上进行性能对比。根据模型的骨干网络不同，我们可以将模型分为两大类：(1) **AReg**^[169]、**MuRel**^[272]、**GRL**^[170]、**RUBi**^[172]、**SCR**^[179]、**LMH**^[173] 和 **HINT**^[178]。这些模型都是将 UpDn^[122] 作为骨干网络。(2) **HAN**^[273]、**GVQA**^[168]、**ReGAT**^[274]、**NSM**^[275]。这些模型都是使用其他不同的骨干网络。例如：模型 **BLOCK**^[159]、**BAN**^[155] 等。特别地，模型 AReg、GRL、RUBi 和 LMH 都是复合模型。

Model	VQA-CP v2 test↑				VQA v2 val↑				GapΔ↓
	All	Yes/No	Num	Other	All	Yes/No	Num	Other	All
HAN ^[273]	28.65	52.25	13.79	20.33	—	—	—	—	—
GVQA ^[168]	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65	16.94
ReGAT ^[274]	40.42	—	—	—	67.18	—	—	—	26.76
RUBi ^[172]	47.11	68.65	20.28	43.18	61.16	—	—	—	14.05
NSM ^[275]	45.80	—	—	—	—	—	—	—	—
UpDn ^[122]	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	23.74
+AReg ^{†[169]}	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	21.58
+MuRel ^[272]	39.54	42.85	13.17	45.04	—	—	—	—	—
+GRL ^{†[170]}	42.33	59.74	14.78	40.76	51.92	—	—	—	9.59
+RUBi ^{†*[172]}	45.23	64.85	11.83	44.11	50.56	49.45	41.02	53.95	5.33
+SCR ^[179]	48.47	70.41	10.42	47.29	62.30	77.40	40.90	56.50	13.83
+LMH ^{†*[173]}	52.45	69.81	44.46	45.54	61.64	77.85	40.03	55.04	9.19
+LMH-CSS	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11	0.96
+HINT+HAT ^[178]	47.70	70.04	10.68	46.31	62.35	80.49	41.75	54.01	14.65
+SCR+HAT ^[179]	49.17	71.55	10.72	47.49	62.20	78.90	41.40	54.30	13.03
+SCR+VQA-X ^[179]	49.45	72.36	10.93	48.02	62.20	78.80	41.60	54.40	12.75

表 7-2 不同视觉问答模型在 VQA-CP v2 和 VQA v2 上的性能对比

实验结果都展示在表 7-2 中。当在数据集 VQA-CP v2 上进行训练和测试时，LMH-CSS 在所有的问题类别上都可以达到目前最好实验性能。特别地，CSS 可以大幅提升 LMH 模型 6.5% 的准确率（58.95% 相比于 52.45%）。当在数据集 VQA v2 上进行训练和测试时，CSS 造成了细微的性能下降（1.74%）。为了评估不同模型对文本偏置的依赖，我们对比了模型在两个不同数据集中性能的差异。相比于之前的模型过于依赖文本偏置（例如：模型 UpDn 和 LMH 的性能差分别为 23.74% 和 9.19%），LMH-CSS 可以显著地减少模型间的性能差异至 0.96%，说明 CSS 可以显著地缓解模型对文本偏置的依赖。

数据集 VQA-CP v1 上性能对比：我们同时将 LMH-CSS 与目前最好的视觉问答模型在数据集 VQA-CP v1 上进行性能评估。同样，根据骨干网络的不同，我们可以将这些模型分为：(1) **GVQA**。它是使用 SAN^[120] 模型作为骨干网络。(2) **AReg、GRL、RUBi** 和 **LMH**。这些模型都是使用 UpDn 模型作为骨干网络。

所有的实验结果都展示在表 7-3 中。通过和其他视觉问答模型对比，LMH-CSS

Model	All	Yes/No	Num	Other
GVQA ^[168]	39.23	64.72	11.87	24.86
UpDn ^[122]	39.74	42.27	11.93	46.05
+AReg ^{†[169]}	41.17	65.49	15.48	35.48
+GRL ^{†[170]}	45.69	77.64	13.21	26.97
+RUBi ^{†*[172]}	50.90	80.83	13.84	36.02
+LMH ^{†*[173]}	55.27	76.47	26.66	45.68
+LMH-CSS	60.95	85.60	40.57	44.62

表 7-3 不同视觉问答模型在 VQA-CP v1 上的性能对比

在数据集 VQA-CP v1 上可以达到目前最好的实验性能。特别地，CSS 可以提升 LMH 模型 5.68% 的准确率（60.95% 相比于 55.27%）。

7.3.3 CSS 对视觉可解释性的帮助

为了验证 CSS 对提升视觉可解释性的有效性，我们主要回答两个问题：**Q1** 现有的具备视觉可解释性的模型能否嵌入复合模型的框架？**Q2** CSS 如何提升视觉可解释性的有效性？

CSS vs. SCR (Q1)：我们将目前最好的视觉可解释性模型 SCR^[179] 直接嵌入到最好的复合模型 LMH 中，然后和 CSS 机制进行对比。实验结果展示在表 7-4 中。

因为目前最好的视觉可解释性模型（如：SCR、HINT）都不是端到端直接训练的。为了公平地对比，我们使用一个训练好的 LMH 模型（VQA-CP v2 上有 52.45% 的准确率）作为模型的初始化。然而，实验结果发现，随着模型的训练，模型的性能逐渐降低，这表明这些具备视觉可解释性的模型都不能嵌入到复合模型的框架中。相反，CSS 可以无缝地嵌入复合模型并进一步提升性能。

视觉可解释性评估 (Q2)：我们分别定量和定性地评估 CSS 对提升视觉可解释性的有效性。对于定量结果，由于数据集中缺少人工的标注（哪些物体为重要物体），我们直接将 SIM 分数当成人工标注的重要物体。然后我们设计了一个新的评价指标 AI (Average Importance)： $|s(a, \mathbf{v})|$ 分数最高的前 K 个物体中的平均 SIM 分数。定量实验结果展示在表 7-5 中。对于定性结果，我们将结果展示在图 7-5 (a) 中。其中绿色框表示分数 $s(\hat{a}, \mathbf{v})$ 大于 0，即对最终预测的答案有正面的贡献；而红色框表示 $s(\hat{a}, \mathbf{v})$ 小于 0，即对最终预测的答案有反面的贡献。只有与问题答案相关性较大的物体才展示出来（即 SIM 大于等于 0.6）。

Model	All	Yes/No	Num	Other
SCR	48.47	70.41	10.42	47.29
LMH	52.45	69.81	44.46	45.54
LMH+SCR			continued decrease	
LMH+CSS	58.95	84.37	49.42	48.21

表 7-4 VQA-CP v2 测试集的准确率

Model	Top-1	Top-2	Top-3
UpDn	22.70	21.58	20.89
SCR	27.58	26.29	25.38
LMH	29.67	28.06	27.04
LMH+V-CSS	30.24	28.53	27.51
LMH+CSS	33.43	31.27	29.86

表 7-5 VQA-CP v2 测试集的 \mathcal{AI} 分数

Model	k=1	k=2	k=3	k=4	\mathcal{CI}
UpDn	49.94	38.80	31.55	28.08	6.01
LMH	51.68	39.84	33.38	29.11	7.44
LMH+Q-CSS	54.83	42.34	35.48	31.02	9.02
LMH+CSS	55.04	42.78	35.63	31.17	9.03

表 7-6 VQA-CP-Rephrasing 测试集的 $CS(k)$ 和 VQA-CP v2 测试集的 \mathcal{CI} 分数

由表 7-5可以看出，CSS 机制可以显著地提升模型的 \mathcal{AI} 分数，即模型做决策时更加依赖与问题答案更加相关的物体。由图 7-5可以看出，CSS 机制可以帮助模型提升重要物体的影响（绿色框），抑制不相关物体的影响（红色框）。

7.3.4 CSS 对问题敏感性的帮助

为了验证 CSS 对提升问题敏感性的有效性，我们主要回答两个问题：**Q3** CSS 可以提升模型对不同句式表达的鲁棒性吗？**Q4** CSS 如何提升问题敏感性的有效性？

对不同句式表达的鲁棒性（Q3）：根据 Shah^[181] 等人的讨论，对不同句式表达的鲁棒性是问题敏感性的一种表现。为了准确地评估，我们按照 VQA-CP 数据

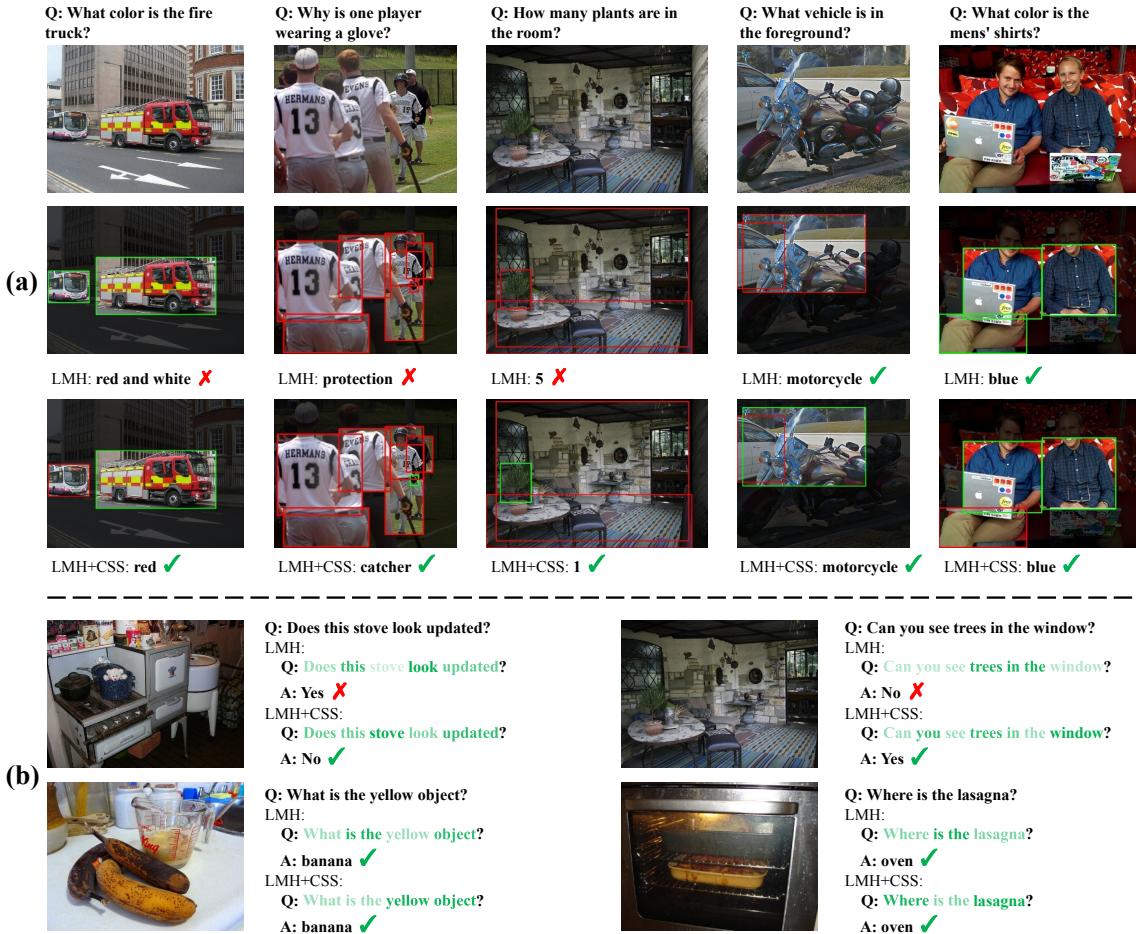


图 7-5 可视化结果

集的划分，将现有的数据集 VQA-Rephrasings^[181] 进行重新划分，记为 VQA-CP-Rephrasings。对于评估，我们使用标准评估指标 $CS(k)$ (Consensus Score)。实验结果展示在表 7-6中。

由表 7-6可以看出，Q-CSS 可以显著提升模型对不同句式结果表达的敏感性。当同时使用两种 CSS 机制时，模型可以进一步提升鲁棒性。

问题敏感性评估 (Q4): 我们分别在定量和定性地评估 CSS 对提升问题敏感性的有效性。对于定量结果，由于缺少统一的评价指标，我们定义一个新的评价指标 \mathcal{CI} (Confidence Improvement)：给定一个样本 (I, Q, a) ，我们通过去除一个重要单词得到另一个样本 (I, Q^*, a) ，然后将这两个样本同时输入到模型中计算标准答案的置信度下降量。我们定义 \mathcal{CI} 为：

$$\mathcal{CI} = \frac{\sum_{(I,Q)} (P_{vqa}(a|I, Q) - P_{vqa}(a|I, Q^*)) \cdot \mathbf{1}(a = \hat{a})}{\sum_{(I,Q)} 1} \quad (7-7)$$

其中 \hat{a} 是模型对样本 (I, Q) 的预测结果， $\mathbf{1}$ 是指示函数。所有的结果展示在表 7-6中。

对于定性结果，我们将结果展示在图 (b) 中，其中不同的绿色程度表示 $s(\hat{a}, \mathbf{w})$ 的相对大小。绿色越深表示该单词对最终模型预测的贡献越大。

由表 7-6可以看出，CSS 机制帮助模型在决策时更加依赖与问题答案更加相关的单词，即去除重要单词之后导致更多的置信度下降。从图 7-5中可以发现，CSS 提升模型对问题的理解，更加关注重要的单词，如“stove”、“lasagna”等。

7.4 本章小结

在本章，我们提出了一种全新的反事实样本生成机制（CSS）来提升视觉问答模型的视觉可解释性和问题敏感性，缓解模型对文本偏置的依赖。CSS 通过遮盖重要的物体和单词，迫使模型更加关注被遮盖的重要区域，同时进一步提升模型预测答案的准确率。因为 CSS 的设计不依赖于任何视觉问答模型，所以 CSS 可以无缝地嵌入任何视觉问答模型中。理论上，CSS 机制可以扩展到其他的多模态任务中，缓解文本偏置问题。

8 总结和展望

8.1 本文工作总结

复杂视觉场景理解是计算机视觉领域的一个研究热点问题。同样，计算机视觉研究的终极目标就是设计一个计算机系统，能够和人类一样感知复杂的外界客观世界。为了能够达到人类级别的视觉场景感知和理解，我们希望该系统模型具备三个基本能力：

1. 模型能够检测和识别场景中所有的组成元素，如规则物体（object）、不规则物体（stuff）和视觉关系（visual relationship）等；
2. 模型可以对视觉场景内容进行理解和推理，并总结和归纳出知识；
3. 模型可以通过自然语言和人类之间进行交互。

对于上述这些能力，我们分别从四个不同的层次对复杂视觉场景进行识别和理解，具体包括：物体识别、场景识别、场景理解和场景推理等。

本文主要的研究内容与贡献如下：

1. 针对目前零样本物体分类模型中普遍存在的属性丢失的问题，本文提出一种全新的零样本学习网络：基于属性保持的对抗网络。本文首次提出图像分类和图像重建是相互冲突的任务，并首次利用对抗学习实现两者之间的知识迁移，减缓语义丢失的问题。本文提出的零样本物体分类模型不仅可以逼真地重建回原始图像，同时可以大幅度提升零样本分类的准确率。
2. 本文首次分析流行的图像场景图生成任务的优化目标的缺陷，并提出图像场景图生成任务优化目标应当具备的两个能力：整体一致性和局部敏感性。本文首次提出将场景图生成任务看成是一个多智能体协同决策问题，并设计一种反事实基准模型，使得模型训练的优化目标同时满足整体一致性和局部敏感性。本文提出的图像场景图生成模型不仅可以显著提升物体的类别预测准确率，同时提升场景图整体的生成质量。
3. 本文首次提出通道注意力机制，并结合现有的空间注意力机制，提出一种

全新的多层空间和通道注意力网络。本文首次分析了卷积神经网络特征图中的三个维度（空间维度、通道维度和层级维度）对图像描述生成的影响。本文提出的图像描述生成模型不仅提升了模型生成描述语句的准确性，同时帮助理解卷积神经网络中特征图的变化过程。

4. 本文通过分析目前视频片段检索框架（自顶向下模型和稀疏型自底向上模型）的优缺点，提出一种全新密集型自底向上的框架，可以避免现有框架的所有缺点。同时，我们设计了一个基于图卷积的特征金字塔层，来增强骨干网络的编码能力。本文提出的视频片段检索模型，在两种不同的查询输入（自然语言和视频片段）形式中，都达到了目前最好的实验性能。

5. 针对目前视觉问答模型忽略的两个重要特性（视觉可解释性和问题敏感性），本文首次提出一种通用的反事实样本生成机制。本文提出的反事实样本机制不仅可以无缝地嵌入任何的视觉问答模型中提升视觉可解释性和问题敏感性，同时可以进一步提升模型性能，在多个视觉问答的数据集中达到目前最好的性能。

8.2 未来研究展望

本文主要围绕复杂视觉场景的感知和理解中涉及的多个关键技术展开研究，其中仍然还有许多方向可以进一步探索，帮助计算机系统达到像人类一样的视觉场景理解能力。具体来说，有以下几点：

1. **设计更加有效的目标检测和分割算法**：目标检测和分割（如：语义分割、实例分割和全景分割等）是计算机视觉领域一个经典的研究问题，它不仅仅是物体层次识别的关键，也是后续场景层次的识别和理解的基础。尽管过去十年内目标检测和分割算法已经取得了长足的进步，但是离大规模的民用还需要进一步提升算法的准确率、检测速度以及鲁棒性。

2. **设计更加轻便的网络结构**：尽管目前的视觉场景感知和理解算法已经可以取得较好的性能，但是基本所有的模型都依赖于强大的计算能力（如：GPU）。另一方面，手机、穿戴式设备等便携设备已经成为日常生活中最常见的计算设备，而这些设备只能依靠CPU等轻量化的计算资源。如何设计更加轻便的网络结构，使得这些视觉场景感知和理解算法能够直接应用于便携设备，将极大地便利人们的日常生活，推动社会的进步。

3. **使用更少的监督信息**：“大数据时代”下，每分钟都会有大量的新图像和视频在互联网上出现。然而，这些海量的媒体数据都不含有任何的人工标注或者少量

的标注信息。如何设计算法使用更少的监督信息（如：弱监督学习、无监督学习或自监督学习等），可以帮助模型充分地利用互联网上的海量媒体数据。

4. 将图像、视频等二维视觉场景推广到三维视觉场景：与图像、视频等二维视觉场景相比，三维视觉场景包含更加丰富的视觉信息，可以辅助对整个视觉场景的感知和理解。例如，当两个物体之间存在较大的 IoU 时，在二维场景下我们通常认为两者之间具有很强的相关性。然而，在三维场景中，这两个物体的深度可能完全不同，即物体之间的相关性很小。因此，对三维视觉场景下的感知和理解同样具有重要的研究意义和研究价值。

参考文献

- [1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick. Microsoft coco: Common objects in context[C]. Proc. ECCV. Springer, 2014:740–755.
- [2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge[J]. Int. J. Comput. Vis., 2015, 115(3):211–252.
- [3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. Int. J. Comput. Vis., 2017, 123(1):32–73.
- [4] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei. Large-scale video classification with convolutional neural networks[C]. Proc. IEEE Conf. CVPR. 2014:1725–1732.
- [5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips[C]. Proc. IEEE ICCV. 2019:2630–2640.
- [6] Yann LeCun, Yoshua Bengio, Geoffrey Hinton. Deep learning[J]. nature, 2015, 521(7553):436–444.
- [7] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks[C]. Proc. NeurIPS. 2012:1097–1105.
- [8] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, Quoc V Le. Self-training with noisy student improves imagenet classification[C]. arXiv. 2019.
- [9] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition[C]. Proc. ICLR. 2015.
- [10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going deeper with convolutions[C]. Proc. IEEE Conf. CVPR. 2015:1–9.

- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition[C]. Proc. IEEE Conf. CVPR. 2016:770–778.
- [12] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He. Aggregated residual transformations for deep neural networks[C]. Proc. IEEE Conf. CVPR. 2017:1492–1500.
- [13] Jie Hu, Li Shen, Gang Sun. Squeeze-and-excitation networks[C]. Proc. IEEE Conf. CVPR. 2018:7132–7141.
- [14] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Proc. NeurIPS. 2015:91–99.
- [15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C Berg. Ssd: Single shot multibox detector[C]. Proc. ECCV. Springer, 2016:21–37.
- [16] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You only look once: Unified, real-time object detection[C]. Proc. IEEE Conf. CVPR. 2016:779–788.
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask r-cnn[C]. Proc. IEEE ICCV. 2017:2961–2969.
- [18] Li Fei-Fei, Rob Fergus, Pietro Perona. One-shot learning of object categories[J]. IEEE Trans. Pattern Anal. and Mach. Intell., 2006, 28(4):594–611.
- [19] Christoph H Lampert, Hannes Nickisch, Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer[C]. Proc. IEEE Conf. CVPR. 2009:951–958.
- [20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollár. Panoptic segmentation[C]. Proc. IEEE Conf. CVPR. 2019:9404–9413.
- [21] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, Li Fei-Fei. Image retrieval using scene graphs[C]. Proc. IEEE Conf. CVPR. 2015:3668–3678.
- [22] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and tell: A neural image caption generator[C]. Proc. IEEE Conf. CVPR. 2015:3156–3164.
- [23] Jiyang Gao, Chen Sun, Zhenheng Yang, Ram Nevatia. Tall: Temporal activity localization via language query[C]. Proc. IEEE ICCV. 2017:5267–5275.
- [24] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, Devi Parikh. Vqa: Visual question answering[C]. Proc. IEEE ICCV. 2015:2425–2433.
- [25] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, Dhruv Batra. Visual dialog[C]. Proc. IEEE Conf. CVPR. 2017:326–335.
- [26] Mateusz Malinowski, Mario Fritz. Towards a visual turing challenge[C]. arXiv. 2014.
- [27] Donald Geman, Stuart Geman, Neil Hallonquist, Laurent Younes. Visual turing test for computer vision systems[J]. Proceedings of the National Academy of Sciences, 2015, 112(12):3618–3623.

- [28] Ali Farhadi, Ian Endres, Derek Hoiem, David Forsyth. Describing objects by their attributes[C]. Proc. IEEE Conf. CVPR. 2009:1778–1785.
- [29] Bernardino Romera-Paredes, Philip Torr. An embarrassingly simple approach to zero-shot learning[C]. Proc. ICML. 2015:2152–2161.
- [30] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings[C]. Proc. ICLR. 2014.
- [31] Berkan Demirel, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning[C]. Proc. IEEE ICCV. 2017:1232–1241.
- [32] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, Xilin Chen. Learning discriminative latent attributes for zero-shot classification[C]. Proc. IEEE ICCV. 2017:4223–4232.
- [33] Christoph H Lampert, Hannes Nickisch, Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization[J]. IEEE Trans. Pattern Anal. and Mach. Intell., 2013, 36(3):453–465.
- [34] Ziad Al-Halah, Makarand Tapaswi, Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning[C]. Proc. IEEE Conf. CVPR. 2016:5975–5984.
- [35] Dinesh Jayaraman, Kristen Grauman. Zero-shot recognition with unreliable attributes[C]. Proc. NeurIPS. 2014:3464–3472.
- [36] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, Osamu Hasegawa. Online incremental attribute-based zero-shot learning[C]. Proc. IEEE Conf. CVPR. IEEE, 2012:3657–3664.
- [37] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, Tom M Mitchell. Zero-shot learning with semantic output codes[C]. Proc. NeurIPS. 2009:1410–1418.
- [38] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, Tomas Mikolov. Devise: A deep visual-semantic embedding model[C]. Proc. NeurIPS. 2013:2121–2129.
- [39] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid. Label-embedding for image classification[J]. 2015, 38(7):1425–1438.
- [40] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, Bernt Schiele. Evaluation of output embeddings for fine-grained image classification[C]. Proc. IEEE Conf. CVPR. 2015:2927–2936.
- [41] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, Bernt Schiele. Latent embeddings for zero-shot classification[C]. Proc. IEEE Conf. CVPR. 2016:69–77.
- [42] Richard Socher, Milind Ganjoo, Christopher D Manning, Andrew Ng. Zero-shot learning

- through cross-modal transfer[C]. Proc. NeurIPS. 2013:935–943.
- [43] Elyor Kodirov, Tao Xiang, Shaogang Gong. Semantic autoencoder for zero-shot learning[C]. Proc. IEEE Conf. CVPR. 2017:3174–3183.
- [44] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, Yueling Zhuang. Zero-shot recognition using dual visual-semantic mapping paths[C]. Proc. IEEE Conf. CVPR. 2017:3279–3287.
- [45] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions[C]. Proc. IEEE ICCV. 2015:4247–4255.
- [46] Li Zhang, Tao Xiang, Shaogang Gong. Learning a deep embedding model for zero-shot learning[C]. Proc. IEEE Conf. CVPR. 2017:2021–2030.
- [47] Ziming Zhang, Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding[C]. Proc. IEEE ICCV. 2015:4166–4174.
- [48] Ziming Zhang, Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding[C]. Proc. IEEE Conf. CVPR. 2016:6034–6042.
- [49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean. Distributed representations of words and phrases and their compositionality[C]. Proc. NeurIPS. 2013:3111–3119.
- [50] Jeffrey Pennington, Richard Socher, Christopher D Manning. Glove: Global vectors for word representation[C]. Proc. EMNLP. 2014:1532–1543.
- [51] George A Miller. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38(11):39–41.
- [52] Scott Reed, Zeynep Akata, Honglak Lee, Bernt Schiele. Learning deep representations of fine-grained visual descriptions[C]. Proc. IEEE Conf. CVPR. 2016:49–58.
- [53] Mohamed Elhoseiny, Babak Saleh, Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions[C]. Proc. IEEE ICCV. 2013:2584–2591.
- [54] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, Terrance E Boult. Toward open set recognition[J]. IEEE Trans. Pattern Anal. and Mach. Intell., 2012, 35(7):1757–1772.
- [55] Abhijit Bendale, Terrance E Boult. Towards open set deep networks[C]. Proc. IEEE Conf. CVPR. 2016:1563–1572.
- [56] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild[C]. Proc. ECCV. Springer, 2016:52–68.
- [57] Yongqin Xian, Christoph H Lampert, Bernt Schiele, Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly[J]. IEEE Trans. Pattern Anal. and Mach. Intell., 2018, 41(9):2251–2265.
- [58] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Shaogang Gong. Transductive multi-view zero-

- shot learning[J]. IEEE Trans. Pattern Anal. and Mach. Intell., 2015, 37(11):2332–2345.
- [59] Kate Saenko, Brian Kulis, Mario Fritz, Trevor Darrell. Adapting visual category models to new domains[C]. Proc. ECCV. Springer, 2010:213–226.
- [60] Bharath Hariharan, Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features[C]. Proc. IEEE ICCV. 2017:3018–3027.
- [61] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, Gianfranco Doretto. Unified deep supervised domain adaptation and generalization[C]. Proc. IEEE ICCV. 2017:5715–5725.
- [62] Pau Panareda Busto, Juergen Gall. Open set domain adaptation[C]. Proc. IEEE ICCV. 2017:754–763.
- [63] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks[C]. Proc. ICML. JMLR. org, 2017:1857–1865.
- [64] Pedro Morgado, Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition[C]. Proc. IEEE Conf. CVPR. 2017:6060–6069.
- [65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative adversarial nets[C]. Proc. NeurIPS. 2014:2672–2680.
- [66] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders[C]. Proc. IEEE Conf. CVPR Workshop. 2018:2188–2196.
- [67] Yongqin Xian, Tobias Lorenz, Bernt Schiele, Zeynep Akata. Feature generating networks for zero-shot learning[C]. Proc. IEEE Conf. CVPR. 2018:5542–5551.
- [68] Yongqin Xian, Saurabh Sharma, Bernt Schiele, Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning[C]. Proc. IEEE Conf. CVPR. 2019:10275–10284.
- [69] Augustus Odena, Christopher Olah, Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans[C]. Proc. ICML. JMLR. org, 2017:2642–2651.
- [70] Eric Tzeng, Judy Hoffman, Kate Saenko, Trevor Darrell. Adversarial discriminative domain adaptation[C]. Proc. IEEE Conf. CVPR. 2017:7167–7176.
- [71] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, Brendan Frey. Adversarial autoencoders[C]. arXiv. 2015.
- [72] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, Russell Webb. Learning from simulated and unsupervised images through adversarial training[C]. Proc. IEEE Conf. CVPR. 2017:2107–2116.
- [73] Cewu Lu, Ranjay Krishna, Michael Bernstein, Li Fei-Fei. Visual relationship detection with

- language priors[C]. Proc. ECCV. Springer, 2016:852–869.
- [74] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, Ian Reid. Towards context-aware interaction recognition for visual relationship detection[C]. Proc. IEEE ICCV. 2017:589–598.
- [75] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, Tat-Seng Chua. Visual translation embedding network for visual relation detection[C]. Proc. IEEE Conf. CVPR. 2017:5532–5540.
- [76] Bo Dai, Yuqi Zhang, Dahua Lin. Detecting visual relationships with deep relational networks[C]. Proc. IEEE Conf. CVPR. 2017:3076–3086.
- [77] Proc. ECCV. Shuffle-then-assemble: Learning object-agnostic visual relationship features[C]. Springer, 2018:36–52.
- [78] Ruichi Yu, Ang Li, Vlad I Morariu, Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation[C]. Proc. IEEE ICCV. 2017:1974–1982.
- [79] Yikang Li, Wanli Ouyang, Xiaogang Wang, Xiao’ou Tang. Vip-cnn: Visual phrase guided convolutional neural network[C]. Proc. IEEE Conf. CVPR. 2017:1347–1356.
- [80] Danfei Xu, Yuke Zhu, Christopher B Choy, Li Fei-Fei. Scene graph generation by iterative message passing[C]. Proc. IEEE Conf. CVPR. 2017:5410–5419.
- [81] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition[C]. Proc. ECCV. Springer, 2018:322–338.
- [82] Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi. Neural motifs: Scene graph parsing with global context[C]. Proc. IEEE Conf. CVPR. 2018:5831–5840.
- [83] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, Ahmed Elgammal. Relationship proposal networks[C]. Proc. IEEE Conf. CVPR. 2017:5678–5686.
- [84] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, Mohamed Elhoseiny. Large-scale visual relationship understanding[C]. Proc. AAAI. volume 33. 2019:9185–9194.
- [85] Yaohui Zhu, Shuqiang Jiang. Deep structured learning for visual relationship detection[C]. Proc. AAAI. 2018.
- [86] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, Xiaogang Wang. Scene graph generation from objects, phrases and region captions[C]. Proc. IEEE ICCV. 2017:1261–1270.
- [87] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation[C]. Proc. ECCV. Springer, 2018:335–351.
- [88] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning[C]. Proc. IEEE Conf.

- CVPR. 2018:1014–1023.
- [89] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, Devi Parikh. Graph r-cnn for scene graph generation[C]. Proc. ECCV. Springer, 2018:670–685.
- [90] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu. Learning to compose dynamic tree structures for visual contexts[C]. Proc. IEEE Conf. CVPR. 2019:6619–6628.
- [91] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, Mingyang Ling. Scene graph generation with external knowledge and image reconstruction[C]. Proc. IEEE Conf. CVPR. 2019:1969–1978.
- [92] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, Jiebo Luo. Attentive relational networks for mapping images to scene graphs[C]. Proc. IEEE Conf. CVPR. 2019:3957–3966.
- [93] Wenbin Wang, Ruiping Wang, Shiguang Shan, Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation[C]. Proc. IEEE Conf. CVPR. 2019:8188–8197.
- [94] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, Wojciech Zaremba. Sequence level training with recurrent neural networks[C]. Proc. ICLR. 2016.
- [95] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward[C]. Proc. IEEE Conf. CVPR. 2017:290–298.
- [96] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, Kevin Murphy. Improved image captioning via policy gradient optimization of spider[C]. Proc. IEEE ICCV. 2017:873–881.
- [97] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, Vaibhava Goel. Self-critical sequence training for image captioning[C]. Proc. IEEE Conf. CVPR. 2017:7008–7024.
- [98] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, Timothy M Hospedales. Actor-critic sequence training for image captioning[C]. Proc. NeurIPS Workshop. 2017.
- [99] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, Feng Wu. Context-aware visual policy network for sequence-level image captioning[C]. Proc. ACM Multimedia. 2018:1416–1424.
- [100] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Kate Saenko. Learning to reason: End-to-end module networks for visual question answering[C]. Proc. IEEE ICCV. 2017:804–813.
- [101] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, Ross B Girshick. Inferring and executing programs for visual reasoning.[C]. Proc. IEEE ICCV. 2017:2989–2998.
- [102] Kan Chen, Rama Kovvuri, Ram Nevatia. Query-guided regression network with context policy for phrase grounding[C]. Proc. IEEE ICCV. 2017:824–832.

- [103] Licheng Yu, Hao Tan, Mohit Bansal, Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions[C]. Proc. IEEE Conf. CVPR. 2017:7282–7290.
- [104] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning[C]. Proc. IEEE ICCV. 2017:2951–2960.
- [105] Juan C Caicedo, Svetlana Lazebnik. Active object localization with deep reinforcement learning[C]. Proc. IEEE ICCV. 2015:2488–2496.
- [106] Stefan Mathe, Aleksis Pirinen, Cristian Sminchisescu. Reinforcement learning for visual object detection[C]. Proc. IEEE Conf. CVPR. 2016:2894–2902.
- [107] Zequn Jie, Xiaodan Liang, Jiashi Feng, Xiaojie Jin, Wen Lu, Shuicheng Yan. Tree-structured reinforcement learning for sequential object localization[C]. Proc. NeurIPS. 2016:127–135.
- [108] Xiaodan Liang, Lisa Lee, Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection[C]. Proc. IEEE Conf. CVPR. 2017:848–857.
- [109] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning[C]. Proc. NeurIPS. 2016:2137–2145.
- [110] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability[C]. Proc. ICML. 2017:2681–2690.
- [111] Ilya Sutskever, Oriol Vinyals, Quoc V Le. Sequence to sequence learning with neural networks[C]. Proc. NeurIPS. 2014:3104–3112.
- [112] Andrej Karpathy, Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions[C]. Proc. IEEE Conf. CVPR. 2015:3128–3137.
- [113] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description[C]. Proc. IEEE Conf. CVPR. 2015:2625–2634.
- [114] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn)[C]. Proc. ICLR. 2015.
- [115] Cheng Wang, Haojin Yang, Christian Bartz, Christoph Meinel. Image captioning with deep bidirectional lstms[C]. Proc. ACM Multimedia. 2016:988–997.
- [116] Sepp Hochreiter, Jürgen Schmidhuber. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735–1780.
- [117] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C]. Proc. ICLR. 2014.
- [118] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov,

- Rich Zemel, Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention[C]. Proc. ICML. 2015:2048–2057.
- [119] Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei. Visual7w: Grounded question answering in images[C]. Proc. IEEE Conf. CVPR. 2016:4995–5004.
- [120] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola. Stacked attention networks for image question answering[C]. Proc. IEEE Conf. CVPR. 2016:21–29.
- [121] Huijuan Xu, Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering[C]. Proc. ECCV. Springer, 2016:451–466.
- [122] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering[C]. Proc. IEEE Conf. CVPR. 2018:6077–6086.
- [123] Ruiyu Li, Jiaya Jia. Visual question answering with question representation update (qru)[C]. Proc. NeurIPS. 2016:4655–4663.
- [124] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, Anton van den Hengel. What value do explicit high level concepts have in vision to language problems?[C]. Proc. IEEE Conf. CVPR. June 2016.
- [125] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo. Image captioning with semantic attention[C]. Proc. IEEE Conf. CVPR. 2016:4651–4659.
- [126] Yingwei Pan, Ting Yao, Houqiang Li, Tao Mei. Video captioning with transferred semantic attributes[C]. Proc. IEEE Conf. CVPR. 2017:6504–6512.
- [127] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, Tao Mei. Boosting image captioning with attributes[C]. Proc. IEEE ICCV. Oct 2017.
- [128] Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation[C]. Proc. IEEE ICCV. 2015:2407–2415.
- [129] Matthew D Zeiler, Rob Fergus. Visualizing and understanding convolutional networks[C]. Proc. ECCV. Springer, 2014:818–833.
- [130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need[C]. Proc. NeurIPS. 2017:5998–6008.
- [131] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares. Image captioning: Transforming objects into words[C]. Proc. NeurIPS. 2019:11135–11145.
- [132] Guang Li, Linchao Zhu, Ping Liu, Yi Yang. Entangled transformer for image captioning[C]. Proc. IEEE ICCV. 2019:8928–8937.
- [133] Lun Huang, Wenmin Wang, Jie Chen, Xiao-Yong Wei. Attention on attention for image captioning[C]. Proc. IEEE ICCV. 2019:4634–4643.

- [134] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara. M ²: Meshed-memory transformer for image captioning[C]. Proc. IEEE Conf. CVPR. 2020.
- [135] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, Tat-Seng Chua. Attentive moment retrieval in videos[C]. Proc. SIGIR. 2018:15–24.
- [136] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, Tat-Seng Chua. Cross-modal moment localization in videos[C]. Proc. ACM Multimedia. 2018:843–851.
- [137] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, Tat-Seng Chua. Temporally grounding natural sentence in video[C]. Proc. EMNLP. 2018:162–171.
- [138] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, Jiebo Luo. Localizing natural language in videos[C]. Proc. AAAI. volume 33. 2019:8175–8182.
- [139] Yitian Yuan, Tao Mei, Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression[C]. Proc. AAAI. volume 33. 2019:9159–9166.
- [140] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos[C]. Proc. AAAI. volume 33. 2019:8393–8400.
- [141] Weining Wang, Yan Huang, Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model[C]. Proc. IEEE Conf. CVPR. 2019:334–343.
- [142] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, Jiebo Luo. Video re-localization[C]. Proc. ECCV. Springer, 2018:51–66.
- [143] Hei Law, Jia Deng. Cornernet: Detecting objects as paired keypoints[C]. Proc. ECCV. 2018:734–750.
- [144] Xingyi Zhou, Jiacheng Zhuo, Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points[C]. Proc. IEEE Conf. CVPR. 2019:850–859.
- [145] Xingyi Zhou, Dequan Wang, Philipp Krähenbühl. Objects as points[C]. arXiv. 2019.
- [146] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian. Centernet: Keypoint triplets for object detection[C]. Proc. IEEE ICCV. 2019:6569–6578.
- [147] Zhi Tian, Chunhua Shen, Hao Chen, Tong He. Fcos: Fully convolutional one-stage object detection[C]. Proc. IEEE ICCV. 2019:9627–9636.
- [148] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Chua. Scancnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]. Proc. IEEE Conf. CVPR. 2017:5659–5667.
- [149] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, Yueting Zhuang. Video question answering via attribute-augmented attention network learning[C]. Proc. SIGIR. 2017:829–832.
- [150] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, Rob Fergus. Simple baseline

- for visual question answering[C]. arXiv. 2015.
- [151] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang. Multimodal residual learning for visual qa[C]. Proc. NeurIPS. 2016:361–369.
- [152] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering[C]. Proc. IEEE Conf. CVPR. 2016.
- [153] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding[C]. Proc. EMNLP. 2016.
- [154] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling[C]. Proc. ICLR. 2017.
- [155] Jin-Hwa Kim, Jaehyun Jun, Byoung-Tak Zhang. Bilinear attention networks[C]. Proc. NeurIPS. 2018:1564–1574.
- [156] Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]. Proc. IEEE ICCV. 2017:1821–1830.
- [157] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering[J]. Trans. Neu. Net. and Learn. Sys., 2018, 29(12):5947–5959.
- [158] Hedi Ben-younes, Remi Cadene, Matthieu Cord, Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering[C]. Proc. IEEE ICCV. 2017:2612–2620.
- [159] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection[C]. Proc. AAAI. volume 33. 2019:8102–8109.
- [160] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. Hierarchical question-image co-attention for visual question answering[C]. Proc. NeurIPS. 2016:289–297.
- [161] Duy-Kien Nguyen, Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]. Proc. IEEE Conf. CVPR. 2018:6087–6096.
- [162] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]. Proc. IEEE Conf. CVPR. 2019:6639–6648.
- [163] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian. Deep modular co-attention networks for visual question answering[C]. Proc. IEEE Conf. CVPR. June 2019.

- [164] Allan Jabri, Armand Joulin, Laurens Van Der Maaten. Revisiting visual question answering baselines[C]. Proc. ECCV. Springer, 2016:727–739.
- [165] Aishwarya Agrawal, Dhruv Batra, Devi Parikh. Analyzing the behavior of visual question answering models[C]. Proc. EMNLP. 2016.
- [166] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Yin and yang: Balancing and answering binary visual questions[C]. Proc. IEEE Conf. CVPR. 2016:5014–5022.
- [167] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C]. Proc. IEEE Conf. CVPR. 2017:6904–6913.
- [168] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering[C]. Proc. IEEE Conf. CVPR. 2018:4971–4980.
- [169] Sainandan Ramakrishnan, Aishwarya Agrawal, Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization[C]. Proc. NeurIPS. 2018:1541–1551.
- [170] Gabriel Grand, Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects[C]. Proc. ACL Workshop. 2019.
- [171] Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, Alexander M Rush. Don’t take the premise for granted: Mitigating artifacts in natural language inference[C]. Proc. ACL. 2019.
- [172] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering[C]. Proc. NeurIPS. 2019:839–850.
- [173] Christopher Clark, Mark Yatskar, Luke Zettlemoyer. Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases[C]. Proc. EMNLP. 2019.
- [174] Rabeeh Karimi Mahabadi, James Henderson. simple but effective techniques to reduce biases[C]. arXiv. 2019.
- [175] Tingting Qiao, Jianfeng Dong, Duanqing Xu. Exploring human-like attention supervision in visual question answering[C]. Proc. AAAI. 2018.
- [176] Chenxi Liu, Junhua Mao, Fei Sha, Alan Yuille. Attention correctness in neural image captioning[C]. Proc. AAAI. 2017.
- [177] Yundong Zhang, Juan Carlos Niebles, Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining[C]. Proc. IEEE WACV. IEEE, 2019:349–357.
- [178] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, Devi Parikh. Taking a hint: Leveraging explanations to make vision and language

- models more grounded[C]. Proc. IEEE ICCV. 2019:2591–2600.
- [179] Jialin Wu, Raymond Mooney. Self-critical reasoning for robust visual question answering[C]. Proc. NeurIPS. 2019:8601–8611.
- [180] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]. Proc. IEEE ICCV. 2017:618–626.
- [181] Meet Shah, Xinlei Chen, Marcus Rohrbach, Devi Parikh. Cycle-consistency for robust visual question answering[C]. Proc. IEEE Conf. CVPR. 2019:6649–6658.
- [182] Yongqin Xian, Bernt Schiele, Zeynep Akata. Zero-shot learning-the good, the bad and the ugly[C]. Proc. IEEE Conf. CVPR. 2017:4582–4591.
- [183] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, Tat-Seng Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval[C]. Proc. ACM Multimedia. 2013:33–42.
- [184] Li-Jia Li, Hao Su, Li Fei-Fei, Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification[C]. Proc. NeurIPS:1378–1386.
- [185] Lorenzo Torresani, Martin Szummer, Andrew Fitzgibbon. Efficient object category recognition using classemes[C]. Proc. ECCV. Springer, 2010:776–789.
- [186] Jason Weston, Samy Bengio, Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings[J]. Machine Learning, 2010.
- [187] Angeliki Lazaridou, Georgiana Dinu, Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning[C]. Proc. ACL. 2015:270–280.
- [188] Zili Yi, Hao Zhang, Ping Tan, Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation[C]. Proc. IEEE ICCV. 2017:2849–2857.
- [189] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. Proc. IEEE ICCV. 2017:2223–2232.
- [190] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma. Dual learning for machine translation[C]. Proc. NeurIPS. 2016:820–828.
- [191] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie. The caltech-ucsd birds-200-2011 dataset[J]. 2011.
- [192] Genevieve Patterson, James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes[C]. Proc. IEEE Conf. CVPR. 2012:2751–2758.
- [193] Justin Johnson, Alexandre Alahi, Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution[C]. Proc. ECCV. Springer, 2016:694–711.
- [194] Alexey Dosovitskiy, Thomas Brox. Generating images with perceptual similarity metrics based

- on deep networks[C]. Proc. NeurIPS. 2016:658–666.
- [195] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]. Proc. IEEE Conf. CVPR. 2017:4681–4690.
- [196] Martin Arjovsky, Soumith Chintala, Léon Bottou. Wasserstein gan[C]. arXiv. 2017.
- [197] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]. Proc. IEEE ICCV. 2015:1026–1034.
- [198] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, Fei Sha. Synthesized classifiers for zero-shot learning[C]. Proc. IEEE Conf. CVPR. 2016:5327–5336.
- [199] Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully convolutional networks for semantic segmentation[C]. Proc. IEEE Conf. CVPR. 2015:3431–3440.
- [200] Ting Yao, Yingwei Pan, Yehao Li, Tao Mei. Exploring visual relationship for image captioning[C]. Proc. ECCV. 2018:684–699.
- [201] Xu Yang, Kaihua Tang, Hanwang Zhang, Jianfei Cai. Auto-encoding scene graphs for image captioning[C]. Proc. IEEE Conf. CVPR. 2019:10685–10694.
- [202] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning[C]. Proc. IEEE Conf. CVPR. 2019:6271–6280.
- [203] Will Norcliffe-Brown, Stathis Vafeias, Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering[C]. Proc. NeurIPS. 2018:8334–8343.
- [204] Drew A Hudson, Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering[C]. Proc. IEEE Conf. CVPR. 2019:6700–6709.
- [205] Jiaxin Shi, Hanwang Zhang, Juanzi Li. Explainable and explicit visual reasoning over scene graphs[C]. Proc. IEEE Conf. CVPR. 2019:8376–8384.
- [206] Monica Haurilet, Alina Roitberg, Rainer Stiefelhagen. It’s not about the journey; it’s about the destination: Following soft paths under question-guidance for visual reasoning[C]. Proc. IEEE Conf. CVPR. 2019:1930–1939.
- [207] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, Martial Hebert. An empirical study of context in object detection[C]. Proc. IEEE Conf. CVPR. IEEE, 2009:1271–1278.
- [208] Xufeng Qian, Yueling Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, Jun Xiao. Video relation detection with spatio-temporal graph[C]. Proc. ACM Multimedia. 2019:84–93.
- [209] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, Philip HS Torr. Conditional random fields as recurrent neural net-

- works[C]. Proc. IEEE ICCV. 2015:1529–1537.
- [210] Philipp Krähenbühl, Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials[C]. Proc. NeurIPS. 2011:109–117.
- [211] Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould. Spice: Semantic propositional image caption evaluation[C]. Proc. ECCV. Springer, 2016:382–398.
- [212] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation[C]. Proc. NeurIPS. 2000:1057–1063.
- [213] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning[J]. PloS one, 2017, 12(4).
- [214] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments[C]. Proc. NeurIPS. 2017:6379–6390.
- [215] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, Shimon Whiteson. Counterfactual multi-agent policy gradients[C]. Proc. AAAI. 2018.
- [216] Yan Zhang, Jonathon Hare, Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering[C]. Proc. ICLR. 2018.
- [217] Matthew Hausknecht, Peter Stone. Deep recurrent q-learning for partially observable mdps[C]. Proc. AAAI. 2015.
- [218] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, Yoshua Bengio. An actor-critic algorithm for sequence prediction[C]. Proc. ICLR. 2017.
- [219] Vijay R Konda, John N Tsitsiklis. Actor-critic algorithms[C]. Proc. NeurIPS. 2000:1008–1014.
- [220] Yongming Rao, Dahua Lin, Jiwen Lu, Jie Zhou. Learning globally optimized object detector via policy gradient[C]. Proc. IEEE Conf. CVPR. 2018:6190–6198.
- [221] Richard S Sutton, Andrew G Barto. Reinforcement learning: An introduction[M]. MIT press, 2018.
- [222] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning[C]. Proc. ICML. 2016:1928–1937.
- [223] Alejandro Newell, Jia Deng. Pixels to graphs by associative embedding[C]. Proc. NeurIPS. 2017:2171–2180.
- [224] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, Amir Globerson. Mapping images

- to scene graphs with permutation-invariant structured prediction[C]. Proc. NeurIPS. 2018:7211–7221.
- [225] Joseph Redmon, Ali Farhadi. Yolo9000: better, faster, stronger[C]. Proc. IEEE Conf. CVPR. 2017:7263–7271.
- [226] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proc. IEEE Conf. CVPR. 2014:580–587.
- [227] Ross Girshick. Fast r-cnn[C]. Proc. IEEE ICCV. 2015:1440–1448.
- [228] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, Bryan Catanzaro. Graphical contrastive losses for scene graph parsing[C]. Proc. IEEE Conf. CVPR. 2019:11535–11543.
- [229] Lex Weaver, Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning[C]. arXiv. 2013.
- [230] Tianshui Chen, Weihao Yu, Riquan Chen, Liang Lin. Knowledge-embedded routing network for scene graph generation[C]. Proc. IEEE Conf. CVPR. 2019:6163–6171.
- [231] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville. Describing videos by exploiting temporal structure[C]. Proc. IEEE ICCV. 2015:4507–4515.
- [232] Maurizio Corbetta, Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain[J]. Nature reviews neuroscience, 2002, 3(3):201–215.
- [233] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention[C]. Proc. NeurIPS. 2014:2204–2212.
- [234] Marijn F Stollenga, Jonathan Masci, Faustino Gomez, Jürgen Schmidhuber. Deep networks with internal selective attention through feedback connections[C]. Proc. NeurIPS. 2014:3545–3553.
- [235] Micah Hodosh, Peter Young, Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics[J]. J. Artif. Intel. Res., 2013, 47:853–899.
- [236] Peter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Trans. Assoc. Comp. Lingui., 2014, 2:67–78.
- [237] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation[C]. Proc. ACL. Association for Computational Linguistics, 2002:311–318.
- [238] Satanjeev Banerjee, Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments[C]. Proc. ACL. 2005:65–72.
- [239] Ramakrishna Vedantam, C Lawrence Zitnick, Devi Parikh. Cider: Consensus-based image description evaluation[C]. Proc. IEEE Conf. CVPR. 2015:4566–4575.

- [240] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf[J]. *Journal of documentation*, 2004.
- [241] Chin-Yew Lin, Eduard Hovy. Manual and automatic evaluation of summaries[C]. Proc. ACL. Association for Computational Linguistics, 2002:45–51.
- [242] Matthew D Zeiler. Adadelta: an adaptive learning rate method[C]. arXiv. 2012.
- [243] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna. Rethinking the inception architecture for computer vision[C]. Proc. IEEE Conf. CVPR. 2016:2818–2826.
- [244] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, Manfred Pinkal. Grounding action descriptions in videos[J]. *Trans. Assoc. Comp. Lingui.*, 2013, 1:25–36.
- [245] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, Juan Carlos Niebles. Dense-captioning events in videos[C]. Proc. IEEE ICCV. 2017:706–715.
- [246] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding[C]. arXiv. 2018.
- [247] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, Yueling Zhuang. Self-supervised spatiotemporal learning via video clip order prediction[C]. Proc. IEEE Conf. CVPR. 2019:10334–10343.
- [248] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, Bryan Russell. Localizing moments in video with natural language[C]. Proc. IEEE ICCV. 2017:5803–5812.
- [249] Runzhou Ge, Jiyang Gao, Kan Chen, Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization[C]. Proc. IEEE WACV. IEEE, 2019:245–253.
- [250] Shaoxiang Chen, Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query[C]. Proc. AAAI. volume 33. 2019:8199–8206.
- [251] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval[C]. Proc. AAAI. volume 33. 2019:9062–9069.
- [252] Songyang Zhang, Jinsong Su, Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language[C]. Proc. ACM Multimedia. 2019:1230–1238.
- [253] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, Ross Girshick. Long-term feature banks for detailed video understanding[C]. Proc. IEEE Conf. CVPR. 2019:284–293.
- [254] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment[C]. Proc. IEEE Conf. CVPR. 2019:1247–1257.
- [255] Caiming Xiong, Victor Zhong, Richard Socher. Dynamic coattention networks for question

- answering[C]. Proc. ICLR. 2017.
- [256] Caiming Xiong, Victor Zhong, Richard Socher. Dcn+: Mixed objective and deep residual coattention for question answering[C]. Proc. ICLR. 2018.
- [257] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension[C]. Proc. ICLR. 2018.
- [258] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proc. ECCV. 2018:801–818.
- [259] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature pyramid networks for object detection[C]. Proc. IEEE Conf. CVPR. 2017:2117–2125.
- [260] Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i Nieto, Shih-Fu Chang. Online detection of action start in untrimmed, streaming videos[C]. Proc. ECCV. 2018:534–551.
- [261] Thomas N Kipf, Max Welling. Semi-supervised classification with graph convolutional networks[C]. Proc. ICLR. 2017.
- [262] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding[C]. Proc. IEEE Conf. CVPR. 2015:961–970.
- [263] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks[C]. Proc. IEEE ICCV. 2015:4489–4497.
- [264] Diederik P Kingma, Jimmy Ba. Adam: A method for stochastic optimization[C]. Proc. ICLR. 2015.
- [265] Aming Wu, Yahong Han. Multi-modal circulant fusion for video-to-language and backward.[C]. Proc. IJCAI. volume 3. 2018:8.
- [266] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, Juan Carlos Niebles. Sst: Single-stream temporal action proposals[C]. Proc. IEEE Conf. CVPR. 2017:2911–2920.
- [267] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C]. Proc. IEEE Conf. CVPR. 2017:2901–2910.
- [268] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. Neural module networks[C]. Proc. IEEE Conf. CVPR. 2016:39–48.
- [269] Andrew Slavin Ross, Michael C Hughes, Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations[C]. Proc. IJCAI. 2017.

- [270] Matthew Honnibal, Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing[J]. To appear, 2017, 7(1).
- [271] Sarthak Jain, Byron C Wallace. Attention is not explanation[C]. Proc. NAACL-HLT. 2019.
- [272] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering[C]. Proc. IEEE Conf. CVPR. 2019:1989–1998.
- [273] Mateusz Malinowski, Carl Doersch, Adam Santoro, Peter Battaglia. Learning visual question answering by bootstrapping hard attention[C]. Proc. ECCV. 2018:3–20.
- [274] Linjie Li, Zhe Gan, Yu Cheng, Jingjing Liu. Relation-aware graph attention network for visual question answering[C]. Proc. IEEE ICCV. 2019:10313–10322.
- [275] Drew Hudson, Christopher D Manning. Learning by abstraction: The neural state machine[C]. Proc. NeurIPS. 2019:5901–5914.

攻读博士学位期间主要研究成果

发表论文：

1. Zero-Shot Visual Recognition using Semantics-Preserving Adversarial Embedding Networks[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. (第一作者, CCF A 类)
2. Counterfactual Critic Multi-Agent Training for Scene Graph Generation[C]. IEEE International Conference on Computer Vision (ICCV), 2019. (第一作者, CCF A 类)
3. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. (第一作者, CCF A 类)
4. Rethinking the Bottom-Up Framework for Query-based Video Localization[C]. Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), 2020. (第一作者, CCF A 类)
5. Counterfactual Samples Synthesizing for Robust Visual Question Answering[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. (第一作者, CCF A 类)
6. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization[C]. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019. (第二作者, CCF B 类)
7. Learning Using Privileged Information for Food Recognition[C]. In ACM International Conference on Multimedia (ACM MM), 2019. (第二作者, CCF A 类)

8. Hierarchical Temporal Fusion of Multi-grained Attention Features for Video Question Answering[J]. Neural Processing Letters, 2019. (第四作者, SCI 期刊)

参与项目:

所获荣誉:

- 2016-2017 学年: 优秀研究生、三好研究生
- 2017-2018 学年: 优秀研究生、浙江大学博士研究生学术新星
- 2018-2019 学年: 优秀研究生、三好研究生

致谢

五年前，我从大连理工大学毕业，怀着对博士的无限憧憬，加入浙江大学计算机学院数字媒体计算与设计实验室（Digital media Computing & Design Lab, DCD），成为一名直博生。时间飞逝，一眨眼我来杭城求学已经整整五个年头，回顾过去的五年，真是感慨万千。五年间，我也曾十分迷茫、痛苦甚至绝望。我曾无数次地想过中途放弃，但身边家人、朋友以及老师们对我始终如一的鼓励和支持，让我最终选择坚持了下来。这段难忘的求学时光，将成为我人生中最为宝贵的一笔精神财富。此时此刻，我想感激曾经帮助过我的每一个人。你们是我不断拼搏前进的动力。

首先，我最想感激的人是我的妻子，黄佳。为了支持我顺利完成学业，在我入学第一年就选择了辞职，放弃了自己的工作，重新选择攻读研究生。在上海工作期间，因为怕耽误我的科研时间，总是选择自己乘坐高铁来杭州看我。如今，为了女儿的健康成长，你又选择了全职在家照顾。没有你的爱，我无法走到现在。此份情意，今生无以为报！

其次，我要感谢我的导师肖俊教授。五年前，因为各种机缘巧合，我有幸成为了肖老师的第一个博士生。肖老师对待自己的学生们，都如同自己的孩子一般，不仅关注学生的学习进步，更关心学生的健康成长。在科研上，肖老师给了我无限的自由。从来没有给我施加任何的科研压力，让我可以凭借自己的兴趣选择研究课题，并且尽自己一切的能力帮助我联系和申请国外的交流合作机会。在生活上，肖老师给我了无限的关心。在我压力大时，时常陪我聊天。在大家很久没有户外活动时，带领实验室同学们一起去西湖边跑步。肖老师不仅传授知识，还教育我如何与人相处。短短几年，收获颇丰。衷心感谢肖老师的付出和指导，能够成为您的学生，是我的荣耀。

感谢南洋理工大学张含望（Hanwang Zhang）教授、哥伦比亚大学张世富（Shih-Fu Chang）教授和新加坡国立大学蔡达成（Tat-Seng Chua）教授。张含望教授是我科研的领路人，从读论文、到构想想法、到设计实验方案、到写论文、再到做报告，张老师对我进行全方位细致的指导。感谢张老师的付出和帮助，让我有幸能够顺利

地度过博士阶段。张世富老师和蔡达成老师作为领域资深专家，他们严谨的治学态度和谦逊的品格，以身作则地告诉我一个优秀的学者应当具备的品质。通过与你们的交流和合作，极大地促进了我科研水平的提升，帮助我加深对问题的思考和对科研的敬畏。

感谢课题组的庄越挺老师、吴飞老师、孟黎瑾老师、李玺老师、汤斯亮老师和赵洲老师在我求学期间对我的真诚帮助和耐心指导。感谢实验室的冯银付、齐天、林昌隆、陈刘策、张瀚之、陈铭洲、周桓等师兄师姐，感谢徐得景、李泽、张宋扬、高旭扬、吉炜等同学，感谢叶钰楠、肖少宁、唐作其、李星辰、金韦克、张凤达、李一萌、王禹潼、孟令涛、严薪、钱旭峰、金松、黄成越、黄一峰、邵飞飞、俞鑫、马文博、蒋志宏、高凯锋、洪暖欣等师弟师妹。感谢所有在新加坡和纽约的小伙伴们，在外交流时和你们一起奋斗的这段时光，将成为我人生中最美好的一段回忆。

最后，感谢我的父母和家人对我的养育和照顾。这段时间，你们在背后默默的支持我、帮助我，在我低落时鼓励我、安慰我。你们无私的奉献，始终激励着我前行。

谨以此文献给所有关心我、帮助过我的人。

陈隆

2020年5月于求是园