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: xx Professor ZJU

Committeeman 1: xx Professor ZJU

Committeeman 2: xx Professor ZJU

Committeeman 3: xx Professor ZJU

Committeeman 4: xx Professor ZJU

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

xx

1. xx

关键词：视觉感知，目标检测，目标跟踪，目标识别，深度学习

Abstract

xx

1. xx

Keywords: Object perception, object detection, object tracking, object recognition, deep learning

目　　次

插 图

表 格

1 绪论

1.1 研究背景

人类对外界客观世界的感知信息绝大部分来自视觉，同样，计算机视觉是机器感知世界的基础。计算机视觉技术的发展，是迈向真正人工智能至关重要的一步。

随着互联网技术及数字媒体设备的快速发展，图像、动态图、视频等包含视觉内容的视觉媒体数据呈现指数级增长。例如，截止到 2019 年 3 月[1]，图像社区 Flickr 每日上传图像的记录高达 2500 万张。截止到 2019 年 10 月[2]，视频分享网站 YouTube 每分钟上传 300 小时视频数据。海量的视觉媒体数据蕴含丰富的视觉场景，通过对视觉场景进行感知、理解和知识推理，可以帮助人类在日常生活中作出重要的决策，推动社会的发展和进步。

然而，日常媒体数据的视觉场景通常包含大量的物体以及物体间的交互，导致对复杂视觉场景的理解与推理存在巨大挑战。因此，利用计算机视觉技术对海量的复杂视觉场景进行感知和理解，在“大数据时代”的背景下，具有十分重大的意义和应用价值。计算机视觉技术的终极目标就是构建一个计算机模型，使其能够像人一样感知和理解复杂的视觉场景。具体来说，这个模型应该具备以下三种能力：

1. 识别和检测视觉场景中所有的组成元素，例如：规则物体（object），不规则物体（stuff），和视觉关系（visual relationship）等；

2. 对视觉场景中感知的视觉信息进行知识推理；

3. 将推理之后得到的知识与人类通过自然语言进行交互。

总的来说，对复杂视觉场景进行感知和理解主要包含对个体层次的识别、场景层次的识别、理解以及推理等关键技术。本文采取的具体技术路线为逐步感知和理解复杂视觉场景的递进研究，如图 1-1所示，包括物体分类、场景图生成、视觉描述生成、视觉检索与视觉问答：

[1]https://expandedramblings.com/index.php/flickr-stats/

[2]https://merchdope.com/youtube-stats/

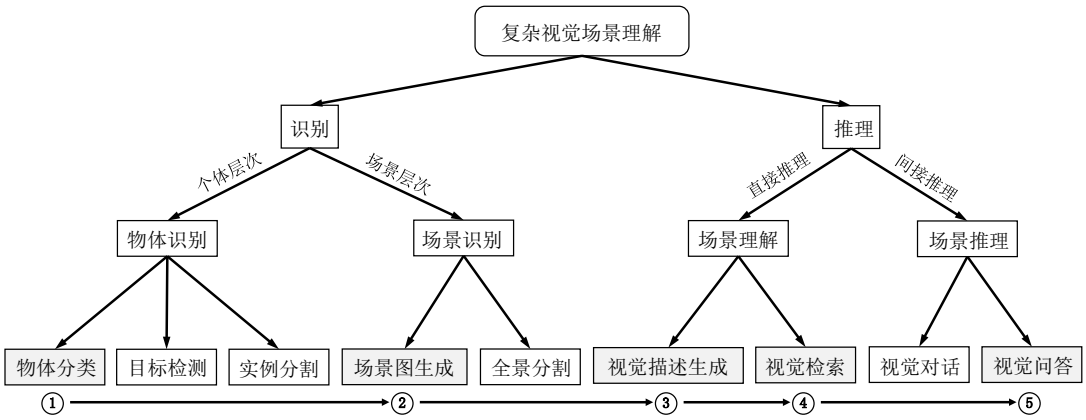

图 1-1 复杂场景感知和理解的关键技术路线

1. **物体识别**：对复杂视觉场景理解的首要步骤是对场景中包含的所有规则物体进行个体层次的识别，它是后续对整个复杂视觉场景进行场景层次的识别、理解和推理的基础。物体识别通常包括其个体的类别（物体分类）、大致位置（目标检测）或者精确位置（实例分割）等任务。具体来说，物体分类[1]任务的目标是对物体进行多类别分类，这也是计算机视觉研究中最基本的任务。随着卷积神经网络[2]在大规模图像分类数据集 ImageNet 上的成功，理想情况下（即每个类别样本数量均衡）的物体分类技术已经日趋成熟，少样本[3]或零样本[4]等更加接近实际应用条件下的物体分类问题已成为近年来的研究热点。目标检测[5]与实例分割[6]任务的目标是在物体分类的基础上，同时对物体位置（矩形框或像素）进行定位。通常，这些问题都被转化为多个候选框（或候选形状）的多类别分类和排序问题。本文主要聚焦在零样本条件下的物体分类任务，研究如何保持物体分类所需的属性特征。

2. **场景识别**：整个视觉场景的识别，除了对所有规则物体进行个体层次识别之外，还需要对物体之间的视觉关系（场景图生成[7]）和所有的不规则物体（全景分割[8]）进行识别。对于复杂场景（即场景中存在大量的物体以及物体间的交互），如果忽略了视觉关系或非规则物体，其后续的场景理解和推理的难度将大大增加。对于前者，场景图生成通过对两两物体间的视觉关系进行检测，将非结构化的视觉场景转化成结构化的场景图，简化对整个场景的理解和推理过程。对于后者，全景分割通过对所有非规则物体进行像素级别的分割，实现对整个场景的语义解析。本文主要聚焦在场景图生成任务的研究上，研究如何设计更加鲁棒的优化目标函数。

3. **场景理解**：在对视觉场景中的所有元素都进行识别之后，计算机模型就可以开始对场景的内容进行理解和推理。对于场景理解而言，一个重要的问题是缺乏统一和标准的衡量指标来判断模型对视觉场景的理解程度。随着递归神经网络（如：

LSTM[9]，GRU[10] 等）在自然语言处理领域（Natural Language Processing, NLP）的成功，在计算机视觉领域开始出现众多视觉与文本融合的多模态研究任务作为场景理解的代理任务，如视觉描述生成[11]、视觉检索[12] 等。对于视觉描述生成而言，模型需要生成描述语句来描述整个视觉场景的内容，并且生成的描述语句是人类能够理解的自然语言。通过对描述语句生成的质量，来判断模型对场景的理解程度。对于视觉检索而言，给定模型一个自然描述语句和一段视频序列，通过对视频序列内容进行理解，从中检索出其视觉场景内容与描述语句一致的视频片段，进而判断模型对场景的理解程度。本文将同时聚焦到视觉描述生成和视觉检索两个任务，研究多模态设定下的视觉场景理解。

4. **场景推理**：人工智能的最终目标，除了对已经存在的物体进行感知和理解外，还希望像人类一样能够做进一步的场景推理。视觉问答[13] 或视觉对话[14] 等任务，通常被看作是一种视觉图灵测试，用来判断模型的推理能力。由于测试问题的自由和开放性，理论上一个理想的模型需要具备物体识别、场景识别、空间推理、常识推理等多方面的能力。在理论上，通过对这类问题的求解，可以进一步思考和理解人类对外界世界的感知和推理过程。在实际应用上，可以帮助人类更好的与机器完成互动，推动社会的进步。本文将主要聚焦到视觉问答任务的研究上，研究如何突破近年来视觉问答研究的瓶颈（如：模型受文本偏置影响较大等）。

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感知和理解。

本文分别针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：

1.2.1 基于属性保持对抗学习的零样本物体分类

1.2.2 基于反事实多智能体学习的场景图生成

1.2.3 基于通道注意力机制的视觉描述生成

1.2.4 基于密集型自底向上框架的视觉检索

1.2.5 基于反事实样本生成的视觉问答

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，提出了多个新算法。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、场景图生成、视觉描述生成、视觉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学习的零样本物体分类方法。为了解决零样本物体分类中的属性丢失（semantic loss）问题，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[3]。算法通过利用对抗学习的思想，对图像分类语义特征与图像重建语义特征进行对抗学习，让图像分类语义特征能够像图像重建语义特征保持尽量多的属性，进而提升零样本物体分类的结果。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场景图生成方法。由于目前绝大多数的图像场景图生成方法都是将所有物体和视觉关系的交叉熵之和作为模型最终的优化目标，这往往忽略了图像中不同物体的重要性。本章首次提出反事实多智能体的训练方法，直接将整个场景图的生成质量当成优化目标，有效地对不同重要性的物体赋予不同的梯度，得到准确的物体和视觉关系分类结果。此项工作发表在国际顶级计算机视觉会议 ICCV 上。

• 第五章介绍了基于通道注意力机制的视觉描述生成方法。现有的视觉描述生成的方法往往借助空间注意力机制，让模型在生成不同单词时关注到不同的空间区域。本章首次提出通道注意力机制，通过对图像卷积网络特征的通道维度进行加权，让模型能关注到不同的语义信息。将通道注意力机制和空间注意力机制进行融合，不仅极大地提升了描述语句的生成质量，也加深了人们对卷积网络特征的理解。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框架的视觉检索方法。本章首先对现有的视觉检索两种框架（自顶向下和稀疏型自底向上）进行了分析，首次提出了一种密集型自底向上框架。通过将视频检索序列中的每一帧看成是正样本，极大地增加了训练过程中正样本的数量，解决了自底向上框架下正负样本不均的问题。利用同一帧视频特征对视频序列两端（起始点、终止点）进行预测，也解决了自底向上框架下两端预测相互独立的问题。此项工作发表在国际顶级自然语言处理会议 EMNLP 上。在密集型自底向上框架的基础上，我们进一步创新，对查询（query）和参考视频（reference video）的交互网络进行设计，提出一种基于图结构的金字塔模型。此项工作发表在国际顶级人工智能会议 AAAI 上。

[3]图像分类需要丢失部分属性特征，而图像重建需要保持所有的属性特征

• 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，让视觉问答模型能够更加关注图像或者文本中的所有重要内容（如：图像中的区域或问题中的单词）。通过将图像或者文本中重要的区域或者单词替换，同时更改成不同的标准答案，组成新的训练样本。将合成的训练样本和原有的训练样本一起训练，迫使模型关注到被替换的区域或者单词，进而让模型能够关注到正确的视觉区域和问题，提升视觉问答准确率和模型的鲁棒性。此项工作已经投稿至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，并提出了对今后的研究展望。

1.4 本章小结

本章对复杂视觉场景理解研究进行了叙述，分别介绍了该问题的研究背景、本文主要的研究内容以及全文的组织结构。

2 相关研究综述

本章将就感兴趣零样本物体识别、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 零样本物体识别

2.1.1 零样本学习

零样本物体识别的主流方法是基于类别属性的物体识别[4,15–19]：这类方法通常将类别属性看成是一个共同语义空间的中间特征，从而实现对不同类别之间的语义迁移。为了实现零样本，流行的类别属性的方法[20?] 都是通过

2.1.2 域偏移问题

2.1.3 对抗生成网络

2.2 图像场景图生成

2.2.1 场景图生成

2.2.2 多智能体梯度策略

2.3 图像描述语句生成

空间注意力机制

语义注意力机制

多层注意力机制

2.4 视频片段检索

2.4.1 基于文本的视频片段检索

2.4.2 基于视频的视频片段检索

2.4.3 自上向下框架与自底向上框架

2.5 图像视觉问答

2.5.1 视觉问答模型的文本偏差

2.5.2 视觉问答模型的特性

视觉可解释性

文本敏感性

3 基于属性保持的零样本物体分类方法

3.1 问题描述

3.2 属性保持的对抗网络学习

3.3 实验设置与性能分析

3.3.1 零样本物体分类数据集

CUB[21]：全称是 Caltech-UCSD-Birds 200-2011 数据集。它是一个细粒度鸟类别分类数据集，总共包含 11788 张来自 200 个细粒度类别的鸟图像，并且每张图像有 312 个语义属性标注。其中训练集包含 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[22]：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，并且每张图像有 102 个语义属性标注。其中训练集包含 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[4]：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，并且每张图像有 85 个语义属性标注。其中训练集包含 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[23].

aPY[15]：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，并且每张图像有 64 个语义属性标注。其中训练集包含 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。

为了公平地和其他模型进行比较，我们使用 Xian 等人[23] 提供的类别嵌入映射

向量，其中每个嵌入映射向量都经过 l_2 范数进行归一化。

3.3.2 实验设定与零样本物体分类评价指标

实验设定：为了评估模型对零样本物体分类的结果，我们采用三种实验设定：

1. U→U: 测试图像的类别和可以预测的类别都只是未见类别；

2. S→T: 测试图像的类别是未见类别，但是可以预测的类别是未见类别和已见类别的总和；

3. U→T: 测试图像的类别和可以预测的类别都是未见类别和已见类别的总和。

通常，U→U 被称为传统型零样本分类，而 U→T 被称为通用型零样本分类。

评价指标: 我们参考现有的文献[23]，常用的每类平均准确率作为评价指标。对于通用型零样本分类，我们另外使用常用的 H 作为主要的评价指标，其中 H 是已见类别 L_s 的准确率（$Acc_{S\to T}$）和未见类别 L_u 的准确率（$Acc_{U\to T}$）的调和平均数：

$$H = 2 \times Acc_{S\to T} \times Acc_{U\to T} / (Acc_{S\to T} + Acc_{U\to T}) \tag{3-1}$$

3.3.3 网络模型与参数设置

网络结构：整个网络结构都是端到端地直接进行训练。其中映射网络 E 是基于 ResNet-101[24]，输入图像的大小是 $224 \times 224 \times 3$。映射网络 F 是基于 AlexNet[2]，然后附加上两层额外的全连接层。重建网络 G 采用类似于生成器[25] 的结构，通过五个连续的反卷积和非线性操作（leaky ReLU）将向量特征转换成三维卷积特征。

参数设置：对于本章所有的实验，训练图像都将短边放缩到 256 个像素。参照 AlexNet[2]，我们采用了增大十倍训练图像的数据增强方式。为了提升训练速度，映射网络 E 中 ResNet-101 部分参数始终保持固定，映射网络 F 的参数初始化采用预训练好的 AlexNet 的参数，重建网络 G 的参数初始化用预训练好的生成器[25]。剩余的所有参数都是用 MSRA 的随机初始化[26]。初始的学习率设置为 $1e^{-4}$，然后当 loss 不下降时，学习率降低 10 倍。

3.4 本章小结

4 基于反事实多智能体学习的图像场景图生成方法

4.1 问题描述

4.2 反事实多智能体学习

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场景图数据集 Visual Genome（VG）[27]。为了与现有工作能够公平地进行比较，我们采用与现有工作相同的数据集划分和预处理[28–32]。处理后的图像数据共包含 150 个物体类别和 50 个视觉关系类别。每张图像平均有 11.5 个物体和 6.2 个视觉关系。整个数据集中，70% 的图像数据当成训练集，30% 的图像数据当成测试集。

实验设定：参考现有的文献[28,29,33]，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、所有的物体框和物体类别，模型需要预测所有的物体组合的视觉关系；

2. 场景图分类（SGCls）：给定图像和所有的物体框，模型需要预测所有物体类别以及所有物体组合的视觉关系；

3. 场景图生成（SGDet）：给定图像，模型需要检测物体框、预测所有物体类别以及所有物体组合的视觉关系。

对于视觉关系中物体框检测来说，需要主语（subject）和宾语（object）与真实准确物体框的交并比（IoU）均大于 0.5。按照惯例，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：为了公平地与现有工作进行对比，我们采用了与[29] 相同的物体检测器。具体来说，它是以 VGG 网络[34] 为主干网络，然后锚框的大小和长宽比与

YOLO-9000[35] 设置一样, 然后用 RoIAlign[6] 代替 RoIPooling。

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，

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）。幸运的是，我们注意到只有极少数的类别有较大的预测概率。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求和来对所有的类别进行近似。在我们的实验中，这样的实验设定可以把速度提升 70 倍，同时维持相同的实验效果。

SGDet 的后处理：对于场景图生成任务（SGDet），为了与之前的工作[29,36] 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。

4.4 本章小结

5 基于通道注意力机制的图像描述语句生成方法

5.1 引言

5.2 通道注意力机制

5.3 实验设置与性能对比

5.3.1 图像描述语句生成数据集

Flickr8k[37]：它一共包含 8000 张图像。按照官方划分，我们将其中 6000 张图像作为训练集，1000 张图像为验证集，以及 1000 张图像为测试集。

Flickr30k[38]: 它一共包含 31000 张图像。因为这个数据集缺少官方划分，我们采用与之前工作[39] 同样的划分。在这种划分中，29000 张图像作为训练集，1000 张图像为验证集，以及 1000 张图像为测试集。

MSCOCO[40]：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。对于官方测试集，由于所有的图像都没有公布其中的人工标注信息，我们同样参考之前的工作[39] 将官方验证集划分为验证集和测试集两部分，其中验证集和测试集各包含 5000 张图像。

5.3.2 评价指标

BLEU[41] (B@1, B@2, B@3, B@4):

METEOR[42] (MT):

CIDEr[43] (CD):

ROUGE-L[44] (RG):

对于所有的四种评价指标来说，他们都是通过比较生成语句中的 n 元词组在人工标注的语句中出现的频率。所有的评价指标都是采用 MSCOCO 官方的测评工具[1]。

[1]https://github/tylin/coco-caption.

5.3.3 实验设定

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[34] 和 ResNet-152[24]。对于文本解码部分，我们使用 LSTM[9] 来生成描述语句中的单词。单词编码的维度和 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，batch size 设置为 16；对于 Flickr30k 和 MSCOCO，batch size 设置为 64。为了避免过拟合，我们采用 dropout 和 early stopping。我们整个模型采用端到端的训练方式，用优化算法 Adadelta[45] 进行参数优化。整个语句的生成过程将会终止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。在测试阶段，我们采用 BeamSearch[11] 的方法，在每个时刻选择 5 个句子作为候选答案。

5.4 本章小结

6 基于自底向上框架的视频片段检索方法

6.1 引言

6.2 视频片段检索

6.3 实验设置与性能对比

6.3.1 视频片段检索数据集

基于语句的视频片段检索：我们在以下三个数据集上进行评估：

TACoS[46]：它一共包含 127 个视频和 17344 个文本与视频序列对（样本）。我们参考现有的标准数据集划分[12]，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。每个样本中视频的平均长度为 5 分钟。

Charades-STA[12]：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。每个样本中视频的平均长度为 30 秒。

ActivityNet Captions[47]：它是目前为止最大、最丰富的数据集，一共包含 19209 个视频。我们参考现有的工作[48], 使用 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。每个样本中视频的平均长度为 2 分钟。

基于视频的视频片段检索：我们在以下数据集上进行评估：

ActivityNet-VRL[49]：它是目前唯一公开发布的数据集。它对动作识别数据集 ActivityNet[50] 共 200 个类别的视频进行了重组，其中任意选取 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，以及剩余 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评估模型的泛化能力。在训练阶段，查询视频和引用视频是随机选取的。在测试阶段，查询视频和引用视频是固定的。

6.3.2 评价指标

基于语句的视频片段检索：我们参考现有工作，使用下列两种通用的评价指标：

R@N, IoU@θ:在测试集中,每个样本预测分数最高的n个的结果重叠度(Intersection-over-Union, IoU）大于θ的百分比。由于自底向上框架的特性，我们仅考虑$N=1$。

mIoU：测试集中所有测试样本的平均的重叠度。

基于视频的视频片段检索：我们使用以下评价指标:

mAP@1：在不同阈值下最高预测结果的平均精度均值（mAP)。

6.4 本章小结

7 基于反事实样本生成的图像视觉问答方法

7.1 引言

7.2 反事实样本生成

7.3 实验设置与性能对比

7.4 本章小结

8 总结和展望

xx

8.1 本文工作总结

xx

本文主要主要的研究内容与贡献如下:

1. xx

8.2 未来研究展望

xx

1. xx

参考文献

[1] 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.

[2] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks[C]. Proc. NeurIPS. 2012:1097–1105.

[3] 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.

[4] 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.

[5] 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.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask r-cnn[C]. Proc. IEEE ICCV. 2017:2961–2969.

[7] 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.

[8] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollár. Panoptic segmentation[C]. Proc. IEEE Conf. CVPR. 2019:9404–9413.

[9] Sepp Hochreiter, Jürgen Schmidhuber. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735–1780.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]. arXiv. 2014.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and tell: A neural image caption generator[C]. Proc. IEEE Conf. CVPR. 2015:3156–3164.

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, Ram Nevatia. Tall: Temporal activity localization via language query[C]. Proc. IEEE ICCV. 2017:5267–5275.

[13] 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.

[14] 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.

[15] Ali Farhadi, Ian Endres, Derek Hoiem, David Forsyth. Describing objects by their attributes[C]. Proc. IEEE Conf. CVPR. 2009:1778–1785.

[16] Bernardino Romera-Paredes, Philip Torr. An embarrassingly simple approach to zero-shot learning[C]. Proc. ICML. 2015:2152–2161.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie. The caltech-ucsd birds-200-2011 dataset[J]. 2011.

[22] Genevieve Patterson, James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes[C]. Proc. IEEE Conf. CVPR. 2012:2751–2758.

[23] Yongqin Xian, Bernt Schiele, Zeynep Akata. Zero-shot learning-the good, the bad and the ugly[C]. Proc. IEEE Conf. CVPR. 2017:4582–4591.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition[C]. Proc. IEEE Conf. CVPR. 2016:770–778.

[25] Alexey Dosovitskiy, Thomas Brox. Generating images with perceptual similarity metrics based on deep networks[C]. Proc. NeurIPS. 2016:658–666.

[26] 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.

[27] 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.

[28] 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.

[29] Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi. Neural motifs: Scene graph parsing with global context[C]. Proc. IEEE Conf. CVPR. 2018:5831–5840.

[30] Alejandro Newell, Jia Deng. Pixels to graphs by associative embedding[C]. Proc. NeurIPS. 2017:2171–2180.

[31] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, Devi Parikh. Graph r-cnn for scene graph generation[C]. Proc. ECCV. 2018:670–685.

[32] 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.

[33] 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.

[34] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition[C]. Proc. ICLR. 2015.

[35] Joseph Redmon, Ali Farhadi. Yolo9000: better, faster, stronger[C]. Proc. IEEE Conf. CVPR. 2017.

[36] 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.

[37] Micah Hodosh, Peter Young, Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics[J]. J. Arti. Intel. Res., 2013, 47:853–899.

[38] 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.

[39] Andrej Karpathy, Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions[C]. Proc. IEEE Conf. CVPR. 2015:3128–3137.

[40] 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.

[41] 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.

[42] Satanjeev Banerjee, Alon Lavie. Meteor: An automatic metric for mt evaluation with improved

correlation with human judgments[C]. Proc. ACL. 2005:65–72.

[43] Ramakrishna Vedantam, C Lawrence Zitnick, Devi Parikh. Cider: Consensus-based image description evaluation[C]. Proc. IEEE Conf. CVPR. 2015:4566–4575.

[44] Chin-Yew Lin, Eduard Hovy. Manual and automatic evaluation of summaries[C]. Proc. ACL. Association for Computational Linguistics, 2002:45–51.

[45] Matthew D Zeiler. Adadelta: an adaptive learning rate method[C]. arXiv. 2012.

[46] 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.

[47] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, Juan Carlos Niebles. Dense-captioning events in videos[C]. Proc. IEEE ICCV. 2017:706–715.

[48] 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.

[49] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, Jiebo Luo. Video re-localization[C]. Proc. ECCV. 2018:51–66.

[50] 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.

作者简历及在学期间所取得的科研成果

个人简历:

姓名：赵黎明 出生年月：1991 年 7 月

民族：汉族 政治面貌：中共党员

邮箱：zhaoliming@zju.edu.cn 个人主页：www.zhaoliming.net

教育经历:

- 2013.09 –2018.06： 浙江大学 计算机科学与技术学院 直博
- 2009.09 –2013.06： 山东大学 软件学院 本科

发表论文:

1. DeepSaliency: Multi-Task Deep Neural Network Model for Salient Object Detection[J]. IEEE Transactions on Image Processing (TIP), 2016. (学生一作，通讯作者，CCF A 类)
2. Metric Learning Driven Multi-Task Structured Output Optimization for Robust Keypoint Tracking[C]. In AAAI, 2015. (第一作者，CCF A 类)
3. Multi-Task Structure-aware Context Modeling for Robust Keypoint-based Object Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2018. (学生一作，通讯作者，CCF A 类)
4. Deep Convolutional Neural Networks with Merge-and-Run Mappings[C]. In IJCAI, 2018. (第一作者，CCF A 类)
5. Deeply-Learned Part-Aligned Representations for Person Re-Identification[C]. In ICCV, 2017. (第一作者，CCF A 类)

6. Deep Learning Driven Blockwise Moving Object Detection with Binary Scene Modeling[J]. Neurocomputing, 2015. (学生四作，SCI 期刊)

7. Semantics-aware Deep Correspondence Structure Learning for Robust Person Re-identification[C]. In IJCAI, 2016. (学生二作，CCF A 类)

8. Geometry-Aware Scene Text Detection with Instance Transformation Network[C]. In CVPR, 2018. (第二作者，CCF A 类)

参与项目：

1. 基于自适应特征学习和表观建模的目标跟踪算法研究，国家自然科学基金面上项目，2015/01 - 2018/12，61472353

2. 面向公共安全的跨媒体计算理论与方法，国家重点基础研究发展计划（973 计划），2012/01 - 2016/12，2012CB316400

3. 城市智慧安监的相关基础理论和视觉分析技术，国家自然科学基金-浙江两化融合联合基金，2016/01 - 2019/12，U1509206

4. 中国工程科技知识中心关键技术研发，中国工程院工程科技知识中心建设项目，2013/01 - 2017/12，124001-D01703

5. 环境/场景适应的跨媒体综合推理，国家自然科学基金人工智能基础研究应急管理项目，2018/01 - 2020/12，61751209

6. 三元空间群智计算，国家重点基础研究发展计划（973 计划），2014/11 - 2019/11，2015CB352302

致谢

xx

陈隆

2020 年 5 月 于求是园