



中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于局部语义学习的图像描述研究

作者姓名: 张晓丹

指导教师: 焦建彬 教授

中国科学院大学电子电气与通信工程学院

学位类别: 工学博士

学科专业: 计算机应用技术

培养单位: 中国科学院大学电子电气与通信工程学院

2018年6月

Local Semantic Learning for Image Captioning

**A dissertation submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Computer Application Technology**

By

Xiaodan Zhang

Supervisor: Professor Jianbin Jiao

**School of Electronic, Electrical and Communication Engineering
University of Chinese Academy of Sciences**

June 2018

中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日 期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

日 期：

导师签名：

日 期：

摘要

图像描述旨在使用自然语言自动描述图像中的主要内容，是人工智能领域中连接计算机视觉和自然语言处理的一个重要研究问题。与图像分类、目标检测等计算机视觉任务相比，图像描述不仅需要捕获图像中的目标，还需要理解目标的属性、行为及其之间的关系，最后还需要将这些语义信息转化为有意义的自然语言。这些特点使得图像描述工作更加复杂也更加具有挑战性，同时也使得其具有极其重要的研究意义。图像描述有许多人工智能相关的应用，例如，帮助视障人士理解视觉内容、辅助幼儿进行看图说话学习等。生动丰富的图像描述也有助于满足我们的日常需求，如图像检索、聊天机器人等。

近年来在该领域占主导地位的研究方法采用卷积神经网络(Convolutional Neural Network, CNN)提取图像特征，然后利用循环神经网络(Recurrent Neural Network, RNN)来进行语言生成。由于 CNN 强大的视觉表达能力、RNN 优越的序列数据处理能力以及深度学习的端到端机制，基于 CNN-RNN 的方法吸引了许多研究者的兴趣，促进了图像描述的进一步发展。

然而，现有方法一般采用全局视觉特征进行语言模型训练，忽略了图像中的局部语义概念以及视觉和语言之间的模态差异。同时，图像语义内容本身是丰富多样的，而现有方法局限于用单一的上下文信息来表达图像内容，对一幅图像通常只生成一个语句，并且一般仅侧重于表达显著性目标。为解决这些问题，本文从局部语义学习的角度出发，将图像目标区域或目标类别信息引入到描述学习中，以此来增强图像理解并提升描述性能。本文主要工作归纳如下：

1. 提出了一种基于图像目标区域生成局部语义表达的方法。通过目标检测获取目标区域，训练全局描述模型 Long Short Term Memory (LSTM)并对目标区域进行局部描述生成，将基于全局图像的单个描述扩展为针对目标的多个局部语义表达，从而更加充分地表达整幅图像的内容。与全局图像描述相比，局部描述能够生成更加具体和准确的语句，对于局部语义探索和挖掘起到了重要作用。

2. 提出了一种局部语义特征表达方法, 将语义特征集成到提出的语义元素嵌入 LSTM 模型(Element Embedding LSTM, EE-LSTM)中以提升图像描述性能。首先, 利用第一个工作获得的多个局部描述语句生成语义特征, 该特征由图像中的局部语义元素构成, 不仅包含图像的细节信息, 而且还与期望描述语句共享同一个语义空间, 弥补了视觉图像和语义表达之间的模态差距。然后, 将 CNN 特征与语义特征集成到提出的 EE-LSTM 模型中生成最终的语言表达。实验表明, 所提出的方法有效地利用了局部语义信息, 优于传统的描述方法。

3. 提出了一种关键词驱动的图像描述, 探索针对图像目标概念的局部语义表达, 是对图像描述更深层的细粒度分析和学习。给定关键词作为指导, 提出的上下文依赖的双边 LSTM 模型(Context-dependent Bilateral Long Short-Term Memory, CDB-LSTM)可以根据关键词生成具备不同侧重点的描述。CDB-LSTM 模型基于关键词进行两个方向的语义学习, 并通过上下文传输模块实现模型的统一, 保障了个性化描述的表达准确性和语法连贯性。

关键词: 图像描述, 局部语义学习, CDB-LSTM, EE-LSTM

Abstract

Image captioning, which aims to automatically describe the main content of an image using natural language, is an important problem in artificial intelligence that bridges computer vision (CV) and natural language processing (NLP). Compared with CV tasks like image classification and object detection, image description should not only capture the objects contained in an image but also express how these objects relate to each other, as well as how their attributes and the activities are involved. Moreover, the above semantic knowledge has to be expressed in a natural language. These characteristics make image captioning to be a challenging and meaningful task, which is helpful for many important applications. For example, it can help visually impaired people understanding the visual world and assist children to learn to express what he sees. The vivid and informative image description is also helpful to satisfy our daily needs including image searching and chatting robot.

The recent state-of-the-art methods apply a convolutional neural network (CNN) to extract the feature of the entire image, followed by a recurrent neural network (RNN) to generate the description of the image content. This CNN-RNN pipeline attracts much research interest due to the strong representation ability of the CNN, the superior ability for sequence data processing of the RNN, and the end-to-end mechanism of neural networks.

These successful approaches, however, are limited to integrate only global visual features, and the local semantic concepts and the modality difference between visual and language spaces have not been considered. An image contains a lot of information from various aspects. Existing image captioning approaches are also limited to describing images with simple contextual information. They typically generate one sentence to describe each image and only focus on the saliency object. In this dissertation, we address these limitations from the view of local semantic learning. Three novel captioning frameworks are proposed to enhance the image

understanding and enrich the captioning by introducing the object regions or object labels. The main contributions of this dissertation can be summarized as follows:

1. We present a method of generating rich image descriptions based on image regions. From object detection, full image captioning model (Long Short-Term Memory, LSTM) training, and region description, our method extends the single full image description to multiple local semantic representations target to object regions, which are sufficient to represent the whole image and contain more information. Comparing with general image level description, generating more specific and accurate sentences on the different regions is helpful for local semantic exploring.

2. We introduce a local semantic feature to improve the image captioning via the proposed Element Embedding LSTM (EE-LSTM) model. The local descriptions obtained in our first work are employed to generate the semantic features, which not only contain detailed information but also share the same semantic space with the target descriptions, and thus bridge the modality gap between visual images and semantic captions. We further integrate the CNN features with the semantic features into the proposed EE-LSTM model to predict the final language description. Experiments demonstrate that the proposed approach effectively utilizes the local semantic information and outperforms recent approaches.

3. We propose the keyword-driven image captioning, which achieves the local semantic learning focus on the object concepts of the image. The proposed Context-dependent Bilateral Long Short-Term Memory (CDB-LSTM) model is utilized to predict a specific sentence driven by an additional keyword. Based on the keyword, CDB-LSTM learns the semantic representations toward two directions, which are unified through a context transfer module and jointly optimized in an end-to-end training framework, which guarantees the accuracy and consistency of the personalized describes.

Key Words: Image Captioning, Local Semantic Learning, EE-LSTM, CDB-LSTM

目 录

第 1 章 绪论	1
1.1 课题背景和研究意义	1
1.2 国内外研究现状	4
1.2.1 基于模板的方法	5
1.2.2 基于检索的方法	5
1.2.3 基于 CNN-RNN 的方法	6
1.2.4 方法对比与问题分析	7
1.3 本文的研究内容	9
1.3.1 基于目标区域的图像描述	9
1.3.2 基于局部语义特征嵌入的图像描述	10
1.3.3 基于目标关键词驱动的图像描述	10
1.4 本文的组织结构	11
第 2 章 图像描述相关技术背景	13
2.1 特征表达	13
2.1.1 视觉特征表达	13
2.1.2 语句特征表达	15
2.2 语言生成模型	15
2.2.1 RNN	17
2.2.2 LSTM	19
2.3 基于 CNN-RNN 的图像描述模型	20
2.3.1 RNN 语言模型训练	21
2.3.1 RNN 语言模型测试	22
2.4 目标检测	23
2.5 数据集	24
2.6 评估方法	26
2.6.1 人工评价标准	26
2.6.2 自动评价标准	27

2.7 本章小结.....	31
第 3 章 基于局部目标区域的图像描述	33
3.1 引言.....	33
3.2 局部描述.....	35
3.2.1 目标检测.....	35
3.2.2 区域优化.....	36
3.2.3 局部描述.....	36
3.3 实验验证与模型分析.....	37
3.3.1 模型参数设置.....	38
3.3.2 实验结果.....	38
3.3.3 定性分析.....	39
3.4 本章小结.....	41
第 4 章 基于局部语义特征嵌入的图像描述.....	43
4.1 引言.....	43
4.2 语义元素嵌入模型.....	46
4.2.1 元素信息挖掘.....	47
4.2.2 语义特征嵌入.....	48
4.2.3 对比模型 VE-LSTM	51
4.3 实验验证及结果分析.....	51
4.3.1 模型参数设置.....	52
4.3.2 实验结果.....	53
4.3.3 模型对比.....	54
4.3.4 模型分析.....	56
4.3.5 定性评估.....	57
4.4 本章小结.....	59
第 5 章 基于目标关键词驱动的图像描述.....	61
5.1 引言.....	61
5.2 模型定义和实现.....	63
5.2.1 子模型训练.....	64
5.2.2 联合优化.....	67

5.2.3 对比模型 I-LSTM	68
5.3 关键词获取.....	68
5.3.1 训练过程关键词来源.....	68
5.3.2 测试过程关键词获取.....	69
5.4 实验验证与模型分析.....	70
5.4.1 模型参数设置.....	70
5.4.2 实验结果.....	71
5.4.3 模型误差分析.....	75
5.4.4 定性分析.....	76
5.5 实验扩展.....	77
5.5.1 单幅图像多个描述生成.....	77
5.5.2 基于元素关键词的描述.....	79
5.6 本章小结.....	80
第 6 章 总结与展望.....	81
6.1 本文工作总结.....	81
6.2 未来工作展望.....	82
参考文献.....	83
致 谢.....	83
作者简历及攻读学位期间发表的学术论文与研究成果.....	99

图目录

图 1.1“视觉-文本/语言”任务对比.....	2
图 1.2 图像描述样例.....	3
图 1.3 图像描述应用领域.....	4
图 1.4 基于模板的方法 ^[39]	5
图 1.5 基于检索的方法 ^[48]	6
图 1.6 基于 CNN-RNN 的方法 ^[52]	7
图 1.7 本文研究内容及其关系框架图.....	10
图 2.1 RNN 及其展开结构.....	17
图 2.2 循环神经网络序列处理模式：输入向量（红色），隐藏层（绿色），输出 向量（蓝色）。.....	18
图 2.3 LSTM 模型内部结构.....	19
图 2.4 基于 CNN-RNN 的图像描述方法.....	20
图 2.5 图像描述数据集展示.....	25
图 3.1 已有工作中局部描述示例.....	34
图 3.2 本章工作局部描述示例.....	34
图 3.3 基于区域的局部图像描述模型.....	35
图 3.4 局部描述示例.....	40
图 4.1 基于 CNN-RNN 的方法存在的问题及本章工作示意图.....	44
图 4.2 全局和局部描述的示例.....	45
图 4.3 技术路线总框架：语义元素信息挖掘和嵌入模型.....	47
图 4.4 元素预处理的变化对比.....	50
图 4.5 EE-LSTM 模型.....	50
图 4.6 模型训练损失误差对比.....	57
图 4.7 EE-LSTM 模型结果示例描述.....	57
图 4.8 反例样本.....	58

图 5.1 图像描述任务本身现存的问题示例.....	61
图 5.2 本章工作样例展示.....	62
图 5.3 CDB-LSTM 模型框架	64
图 5.4 关键词生成.....	69
图 5.5 目标类别到关键词的匹配映射.....	70
图 5.6 CDB-LSTM 与 I-LSTM 对比	73
图 5.7 基于关键词驱动的图片描述.....	74
图 5.8 人工评测结果对比.....	75
图 5.9 模型训练损失误差对比.....	75
图 5.10 传统图像描述模型通过扩大 beam_size 参数生成的语句.....	77
图 5.11 CDB-LSTM 在 Visual Genome 数据集上的描述结果.....	78
图 5.12 基于不同关键词的描述示例.....	80

表目录

表 2.1 图像描述数据集.....	24
表 3.1 BLEU 评测结果(%).....	38
表 4.1 全局和局部描述结果.....	53
表 4.2 在 MS COCO 数据集上的模型对比(%).....	55
表 5.1 自动评测结果对比.....	72
表 5.2 Visual Genome 数据集评测结果.....	79

第1章 绪论

1.1 课题背景和研究意义

随着数字摄影设备和互联网的飞速发展，越来越多的图像、视频等多媒体信息不断涌入人们的视野并且持续传播扩散。面对如此大规模的数据，如何有效地进行处理、组织和分析，已经成为学术界和工业界面临的重要问题。

近年来大数据技术及深度学习算法的显著进步，使得人工智能(Artificial Intelligence, AI)领域经历了巨大的变革。众多子领域迅速发展使得诸如自动驾驶、指纹识别、人脸识别、聊天机器人、智能安全等应用已经出现在我们的日常生活中。在过去二十年中，计算机视觉(Computer Vision, CV)技术和自然语言处理(Natural Language Processing, NLP)技术在图像/视频理解和文本分析方面取得了巨大的进步。CV领域的图像分类、目标检测、目标跟踪，以及NLP领域的机器翻译、智能问答等都取得了极大的成果。

近年来，结合CV和NLP的研究问题引起了学术界和工业界越来越多的关注。例如，图像描述、视频描述、视觉问答系统等跨学科问题都有了飞速的发展。作为此类“视觉-语言”问题的典型，图像描述旨在自动对图像的主要内容进行语言描述。图像描述结合了CV和NLP两大领域内的研究问题，即由视觉图像到语言生成，涉及图像视觉信息理解和语言的自动生成等关键技术。

计算机视觉(CV)是一个重要的研究领域，涉及如何使计算机获得数字图像或视频的高层次理解。从CV的角度来看，将视觉图像映射到文本的研究具有丰富的历史背景。图像分类^[1-5]旨在为每幅图像提供目标的单一类别，例如“狗”、“猫”、“人”等。场景分类^[6-10]判断图像属于哪个场景，例如“街道”、“村庄”、“卧室”等。除了这些图像到单个单词的分类问题，早期将多个单词和图像关联的工作侧重于自动标注，主要将单词与图像多个区域相关联^[11]。目标检测(Object Detection)^[12-14]则能够发现图像中多个目标的类别以及其所在区域。其他一些工作尝试生成属性(Attribute)^[15]或行为(Activity)^[16-18]以便获得丰富的语义表示。无论这些任务想达到什么目标，它们都有相似的技术方案。首先是视觉

特征表达，通过人工设定或自动学习的特征模型实现图像理解；然后利用视觉特征训练一个可以获得相应视觉任务的判别模型，如图像分类、目标检测、场景分类等。尽管这些任务的定义不同，但视觉表达技术是共通的。例如，用于特征提取的 CNN 模型通常是在图像分类数据集中预训练的，而 CNN 特征可以用于完成各种视觉任务^[19]。

这些“视觉-文本”任务的输出是一个或多个非结构化标签列表，这些标签对于判别分析是很有意义的，但不足以产生人类语言表示，也不能满足更高层的语义分析和表达需求。图 1.1 中列举了几种视觉相关任务，蓝色字体为对应任务的目标或产生的结果。

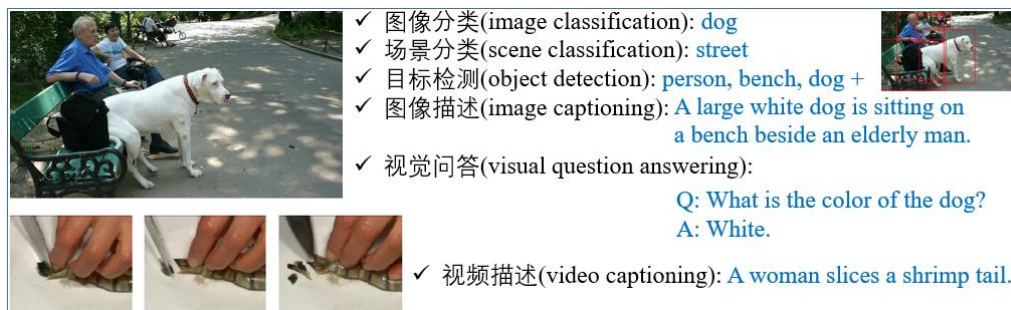


图 1.1 “视觉-文本/语言”任务对比

图像描述能够用语言表达图像中涵盖的语义信息。好的图像描述必须全面而简明，并且必须是形式上连贯的，即由语法正确的语句组成，如图 1.1 中所示，更多样例参见图 1.2。与计算机视觉领域“视觉-文本”任务相比，图像描述工作可进一步归类为“视觉-语言”问题，其更难更具挑战性，是计算机视觉从感知到认知的跨越。这项任务不仅要捕捉包含在图像中的目标，还必须获取它们的属性、行为、关系、场景等相关语义信息，完成深层的图像理解。这些语义知识必须用人类的自然语言（英语、汉语等）来表达成完整的可读的语句，这意味着除了视觉理解之外图像描述还需要语言模型的支撑。例如，图 1.1 中图像描述“A large white dog is sitting on a bench beside an elderly man.”，这种语言形式的信息远多于图像分类、场景分类、目标检测等任务中的个别或多个单词。CV 中的特征表达方法可以直接用在图像描述任务中进行视觉理解，但单独的视觉技术已经不能解决关键的语言描述问题，需要自然语言理解技术的支撑。

自然语言处理(NLP)是计算机科学和人工智能中体现计算机与人类(自然)语言之间交互的领域,主要研究计算机如何处理大量自然语言数据的相关问题。从NLP的角度来看,图像描述是一种自然语言生成(Natural Language Generation, NLG)^[20]问题。NLG是NLP领域的一部分,任务是将计算机输入信息表达映射为自然语言,这些输入信息可以是一个数据库、逻辑表格、专家系统知识库或其他机器可读的数据。在图像描述中,输入是数字图像,因此需要对视觉图像进行特征建模,即利用CV技术进行视觉特征提取,然后通过语言模型将输入的特征按顺序映射为由单词组成的人类可理解的语句,如图1.2所示。



图 1.2 图像描述样例

总体而言,自动图像描述不仅需要全面的准确的图像理解,还需要复杂的自然语言生成。这些特点正是使CV和NLP领域衍生出这样一项极为挑战又有趣的任务的原因。一般来说,处理这个视觉到语言的任务包含两个步骤:第一步是视觉理解模块,从中可以获得视觉特征表示,这一步和CV领域的工作和进展息息相关;第二步是语言生成模块,需要根据视觉特征生成一个语句或段落,这一部分又是NLP的重要内容。而CV和NLP这两个领域的技术的进步都会对图像描述产生有利的推动。

人类的许多日常任务都需要通过视觉与语言进行交互,例如,为视障人士提供前景图像内容理解、为电影或视频添加字幕、使用语言来从大量图像中搜索图像、与机器人进行自然聊天、辅助幼儿教育等,如图1.3所示。图像描述对这些应用及与其相关的涉及图像到语言感知与认知的任务都有直接帮助,是人工智能领域发展中不可或缺的研究方向。



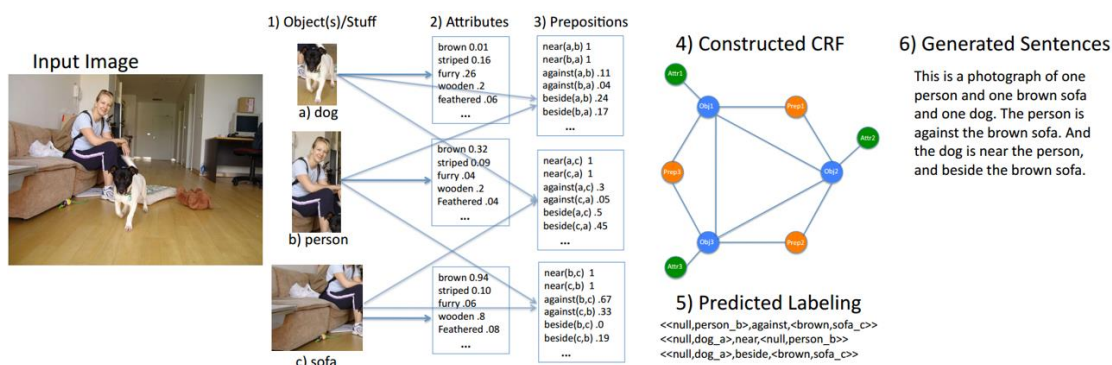
图 1.3 图像描述应用领域

除了图像描述之外，其他“视觉-语言”的问题，如视频描述(Video Captioning)^[21-30]和视觉问答(Visual Question Answering, VQA)^[31-36]也引起了广泛关注。这些任务也是连接 CV 和 NLP 的跨学科领域，并和图像描述具备类似的应用和方法。从视频中生成自然语言描述对基于图像的描述来说又具备了新的挑战，因为它还需要分析时间维度中的目标及其属性和相关关系。视觉问答系统学习一种理解更加局部的视觉内容的模型，并以自然语言形式发现它们与成对问题和答案的关联。但是，大多数现有的描述生成工作都基于静态图像。作为这些“视觉-语言”问题的根本任务，图像描述有助于视频描述和视觉问答。图像描述领域的技术进步可以扩展到“视觉-语言”类问题的解决方法。本文工作重点关注图像描述问题研究，并且针对的图像都是丰富多样的自然图像，而非某一领域的某一场景的特殊图像，描述语句则为英文表达。

1.2 国内外研究现状

现阶段图像描述领域主要有三种研究方法：基于模板的方法、基于检索的方法和基于 CNN-RNN 的方法。

1.2.1 基于模板的方法

图 1.4 基于模板的方法^[39]

基于模板的方法^[37-41]先检测出图像中的目标、动作、场景、属性及其关系，然后根据模板来将这些单词进行组合生成描述语句。模板可以是马尔可夫随机场^[38, 39]或最大熵语言模型^[37]，这些模型可以将多个单词映射为有意义的语句。Kulkarni 等^[39]提出了一个自动生成图像自然语言描述的系统，该系统利用从大量文本数据和图像识别算法中收集的单词（目标、属性等）进行统计分析，利用条件随机场(Conditional Random Field, CRF)进行单词组合，最终进行语句生成，图 1.4 显示了其提出的框架。M. Mitchell 等^[40]利用语法知识进行词语共现统计，并采用视觉方法生成句法树，构成了图像描述。Socher 等^[42]基于递归神经网络对复杂场景图像和语句进行结构预测，从而实现对自然语言描述的语义解析及场景语义分割和标注。

以上这些方法非常直观，也取得了不错的结果，但此类方法的缺点是高度依赖标注（目标、动作、场景、属性等），而目前获取完全正确并且丰富的标注信息是很困难的。另外，基于模板的方法生成的语句不够自然，不能实现类似人类语言的目的。

1.2.2 基于检索的方法

基于检索的方法^[43-48]利用图像检索结果并通过匹配的方式进行图像描述。因为相似的图像更可能拥有相似的语言描述，所以通过图像检索，这类方法可以利用检索到的最相似的图像的描述进行目标图像描述。这类方法利用相似性在视觉空间中将描述问题转移到图像检索问题。如图 1.5 所示，Ordonez 等^[48]

提出了一个包含五个步骤的描述系统：将查询图像输入到描述系统；使用图像描述子从图像库中检索候选匹配图像；提取与图像内容有关的高层信息，例如目标，场景等；根据高层信息重新排列匹配图像；根据检索结果，返回最佳描述，描述也可以在第 2 步之后从与全局最匹配图像相关的描述中生成。

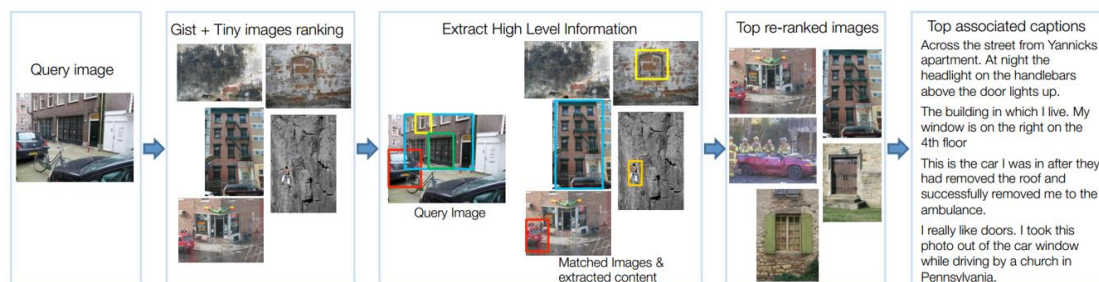
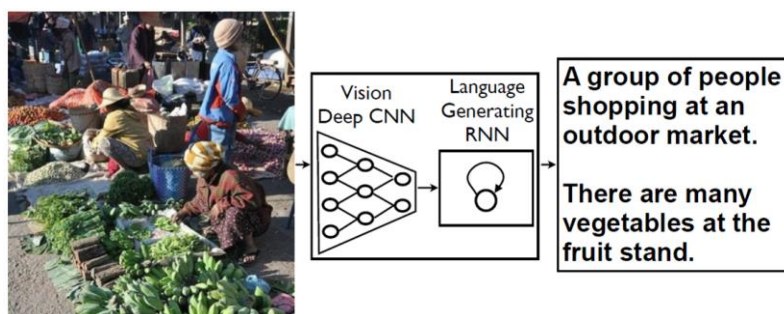


图 1.5 基于检索的方法^[48]

与基于模板的方法相比，基于检索的方法能够获得自然的语句，但缺点是太过依赖检索结果，并且获得的描述结果都是已经存在的固定语句，无法生成新语句，因此生成的描述不能适应灵活的图像内容表达。这也意味着该方法通常需要大量的图像数据集以提供足够的语句数据量，而这又意味着需要增加时间成本。

1.2.3 基于 CNN-RNN 的方法

基于 CNN-RNN 的方法^[49-52]遵循源自神经机器翻译^[53, 54]的编码器-解码器框架，其中源语言的语句可以通过级联的编码和解码过程被翻译为目标语言。图像描述同样可以通过编码和解码来实现图像到语句的转换。图像描述中的编码过程是使用卷积神经网络(Convolutional Neural Network, CNN)将图像映射为特征，解码过程使用循环神经网络(Recurrent Neural Network, RNN)进一步将特征转换为语句。Oriol Vinyals 等^[52]最早提出了这种与以往截然不同的图像描述方法，即先采用 GoogLeNet^[55]对图像进行视觉特征提取，然后用 Long Short Term Memory (LSTM)^[56]建立视觉特征到语言的映射关系，进而对测试图像进行描述生成，如图 1.6 所示。Kapathy 和 Li^[49]采用 VGGNet^[4]对图像进行视觉特征提取，然后用 RNN 进行图像描述。CNN 具备很强的视觉表达能力，RNN 擅于处理序列数据，因此基于 CNN-RNN 的描述方法在近几年取得了极大的性能提升，从而推动了图像描述领域的发展。

图 1.6 基于 CNN-RNN 的方法^[52]

1.2.4 方法对比与问题分析

1) 方法对比

图像描述问题的难点在于如何挖掘图像中包含的语义概念信息，并且能够用人类可理解的自然语言表达出来。

基于模板的方法通过检测获取图像中的目标、动作、场景、属性及其关系，然后用语言模板进行语句生成。这种方法可以看成是两步策略：图像理解、语言生成。图像理解步骤可以采用计算机视觉领域的算法进行语义知识检测，语言生成步骤则用合适的模板进行语义知识连接。这类方法跟人类认知模式相似，在早期获得了广泛关注，但由于生成的语句过于依赖模板，导致图像描述不够自然。

基于检索的方法通过视觉特征匹配，把数据库中搜索到的相近图像的语义表达作为目标图像的语义描述。这类方法直观有效，缺点是以迁移的方式进行语句匹配，只能生成真实标注中的句子，无法产生更加自由灵活的新句子。在面对和数据库中样本完全不同的图像时，这类方法便会失去描述功能。

从视觉空间到语义空间映射关系的角度来看，基于 CNN-RNN 的方法能够实现基于检索的方法的功能，即采用 CNN 特征进行语义拟合。不同的是，CNN-RNN 的方法经过了复杂的神经网络训练，比直观的检索功能更为强大，能够自由组合并生成真实标注中没有出现过的新语句。比如，真实标注中句子有“a man is sitting on a bench”，“a dog is playing on the grass”，基于 CNN-RNN 的方法能够生成“a man is sitting on a bench next to a dog”，这个句子是模型根据图像内容自动组合的新语句。

从图像理解和语言生成的角度来看，基于 CNN-RNN 的方法和基于模板的

方法也有共通之处。图像理解对应于 CNN 特征获取，取代了之前的检测语义概念；语言生成对应于 RNN 序列模型训练和测试，取代了之前的模板连接。比起基于模板方法的两步处理策略，基于 CNN-RNN 的方法可以将两步统一起来进行联合训练。

总体而言，基于 CNN-RNN 的方法很好地利用了深度学习的优点：CNN 强大的视觉表达能力、RNN 优越的序列数据处理能力、深度学习端到端的统一训练框架，这些优点使得基于 CNN-RNN 的方法生成的语句更加自然、灵活、准确，在图像描述领域很快占据了主导地位。但与基于模板的方法相比，其忽略了可检测的语义概念，如目标、场景、动作、属性、关系等局部语义信息。尽管 CNN 的特征表达性能已经很强大，但视觉和语言空间之间的语义鸿沟仍然是一个巨大的挑战。使用全局 CNN 特征的图像描述方法无法表示和描述图像中所有重要元素，特别是对于具有复杂场景和目标的图像。因此，局部语义信息的挖掘和学习是图像描述研究中不可或缺的工作。

2) 问题分析

本文沿用基于 CNN-RNN 的描述方法，从三个方面分析图像描述领域存在的问题。

a) 局部描述的完整性和丰富性

现有的图像描述大多只关注整幅图像的表达，不能很好地表达图像自身包含的丰富内容，更是无法满足个性化的应用需求。例如，当有人想要获取图像中特定位置的详细信息时，单纯的对整幅图像的描述将不足以满足这种需求。Kapathy 和 Li^[49]提出的模型可以生成针对给定图像区域的短语描述。密集描^[57, 58]为预测的区域候选框生成局部短语表达。但是，这些局部描述都是一些词汇或短句，对于实际应用来说太过简短，也太过分散。

b) 局部信息的表达能力

基于全局视觉特征的现有方法很难将局部细节整合到语言模型中。基于 CNN-RNN 的方法没有考虑全局视觉特征的局限性以及视觉和语言空间之间的形态差距。一些研究者利用视觉注意力^[36, 57, 59-66]进行描述模型增强，即将具备“焦点”的局部视觉信息引入到 CNN-RNN 框架中，取得了一定的成功，对图像描述领域的发展起了极大的推动作用。但是，这些方法忽略了图像中的局部

语义信息。一些其他基于 CNN-RNN 的描述模型的工作^[36, 61, 62, 65, 66]将语义特征嵌入语言模型中。但这些方法仅考虑高层次概念，却忽略了局部区域信息。

c) 图像描述的模糊性

从图像描述任务的角度来看，现有的图像描述方法^[14, 37, 43-46, 48, 49, 52, 66]虽然取得了很大的成果，却局限于使用简单的上下文信息来表达图像内容：只生成一个语句，且仅描述最显著的目标而忽略其他非显著目标或隐含语义信息。因此，生成的语句中很难覆盖图像的所有细节。但是，图像语义表达本身是丰富多样的，尤其对于复杂图像，很难用一句话来完整地描述整幅图像；对于不同的人来说，由于关注点不同，对同一幅图像可能有不同的理解，用一句话来定义一幅图像本身存在语义上的偏差。这些问题导致了对图像局部语义学习与多样性表达进行进一步研究的迫切性。已经有一些工作通过段落描述^[67]或者密集描述^[51, 68]的方式来解决这个问题。但是，这些工作的共同点是都需要重新建立新的数据集，标注更为复杂，已经不同于传统的图像描述工作。另外，图像描述是一个高度个性化的任务，不同用户对于同一幅图像的语义表达可能有不同的侧重点，这也是已有工作无法解决的问题。

1.3 本文的研究内容

本文从局部语义学习的角度解决上述问题。通过引入三种描述方法，从目标区域和目标概念的角度探索和利用局部语义信息，达到增强图像理解和丰富图像内容描述的目的。之所以选择目标为切入点，是因为图像描述是以主要目标为中心的语义表达，针对目标区域进行图像描述探索是有意义的。本文第 3-5 章内容对应提出的三种方法，如图 1.7 所示。

1.3.1 基于局部目标区域的图像描述

提出一种局部描述方法，主要描述通过目标检测方法获得的目标区域。基于目标区域的图像描述工作旨在获取更加丰富完整的视觉语义表达。在这个工作中，局部语义学习通过对目标区域的语义描述引入。首先使用目标检测方法来生成多个候选区域，然后训练 RNN 语言模型以学习图像区域和语句之间的描述关系，最后，所有目标区域被用于生成图像描述。所提出的模型能对图像中的多个区域生成描述语句，如图 1.7 中所示，该样例能够生成 5 个局部描述，

这些语句对整幅图像具有足够的表达能力并且包含更详细的语义信息。与一般图像描述相比，在不同区域生成更加具体和准确的语句可以满足不同用户的更多个性化需求。这部分工作还表明全局图像描述与目标局部描述之间存在互补关系，这对于未来与局部语义表示相关的工作是意义重大的。

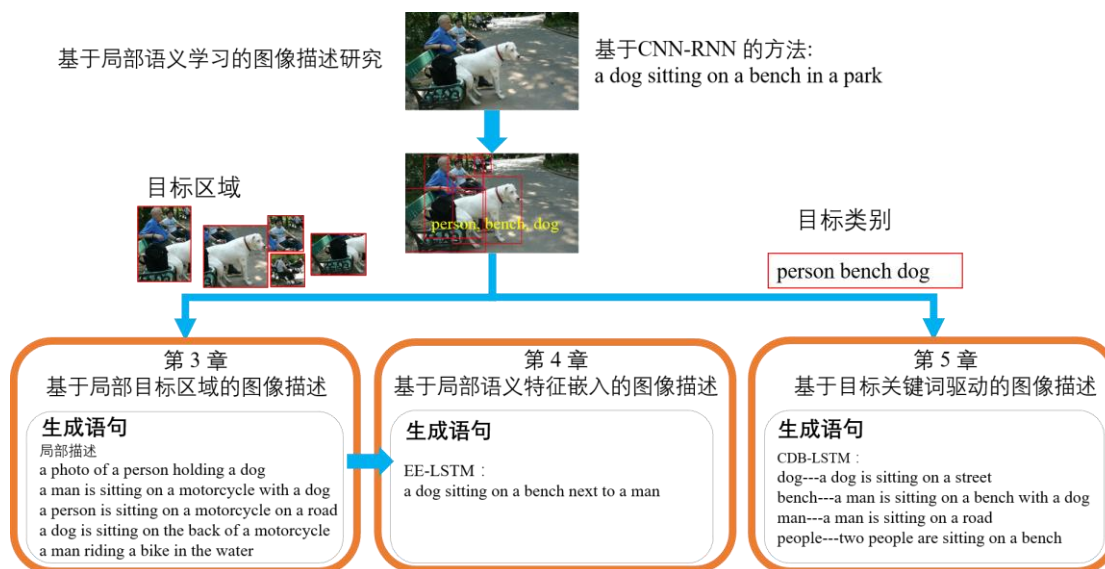


图 1.7 本文研究内容及其关系框架图

1.3.2 基于局部语义特征嵌入的图像描述

提出一种挖掘局部语义特征的方法并通过语义元素嵌入模型(Element Embedding LSTM, EE-LSTM)来提升语言表达能力。这项工作从目标区域的角度来探索局部语义学习。对于语义信息的获取，根据“基于目标区域的图像描述”中局部描述的结果进行单词打散和筛选。语义元素单词被用于生成语义特征，其不仅包含局部目标的详细信息，而且与描述语句共享同一个语义空间，弥补了视觉图像与语义描述之间的模态差距。然后，将 CNN 特征与语义特征集成到 EE-LSTM 模型中以预测最终的语言描述。如图 1.7 所示，与基于 CNN-RNN 的传统方法相比，EE-LSTM 模型的图像描述结果包含了更多细节信息，描述性能得到了提升。

1.3.3 基于目标关键词驱动的图像描述

提出了一种上下文依赖的双边 LSTM (Context-Dependent Bilateral Long Short-Term Memory, CDB-LSTM)模型以预测由目标关键词驱动的语句，从目标

概念/关键词的角度来探索局部语义学习。该模型以给定的关键词作为指导，生成更深层的语义表达，还可以根据不同的关键词为同一图像生成具备不同侧重点的描述，这是前两个工作都不具备的特性。**CDB-LSTM** 包含两个级联的子模型，第一个子模型以逆序（从给定的关键词到语句的开始）生成语句的前半部分，第二个子模型通过考虑第一个子模型的预测结果以正序生成语句的后半部分（从给定单词到语句末尾）。语句的前半部分和后半部分合并在一起构成最终的语句表达。通过上下文传输模块考虑两部分的语义依赖性，两个子模型在端到端框架中统一起来并且联合优化，从而使得模型实现对各种关键词的良好适应性和语义一致性。

1.4 本文的组织结构

本文在研究图像描述方法的基础上，分析了已有方法的局限性，通过引入不同的局部语义学习策略提出了三种图像描述方法。具体结构如下：

第1章，绪论。主要论述图像描述的研究背景和意义，分析国内外的研究现状以及发展趋势，最后说明了本文的主要研究目的和研究内容。

第2章，图像描述相关技术背景。为本文的工作提供了背景知识介绍，包括特征表达、语言生成、目标检测、数据集和评估标准等。

第3章，基于局部目标区域的图像描述。提出了一种生成局部语义描述的方法。首先通过目标检测来对图像生成候选区域，然后根据全局图像和标注语句训练 RNN 语言模型，最后针对目标区域进行视觉特征提取并利用 RNN 语言模型进行局部语义描述生成。基于不同目标区域的描述可以为图像提供更丰富的信息，也使得语义表达更加具体和准确。

第4章，基于局部语义特征嵌入的图像描述。引入局部语义特征以增强图像的表达能力，并通过提出的 **EE-LSTM** 模型进行特征集成，从而实现描述模型的性能提升。首先根据第3章的工作探讨局部描述对全局描述的意义，然后进一步挖掘局部语义信息并生成语义元素特征，最后将元素特征嵌入到所提出的 **EE-LSTM** 模型中以生成表达更完备的语句。实验结果表明，该模型比基准模型具有更好的性能，显示了局部语义特征对于图像理解与表达的有效性。

第5章，基于目标关键词驱动的图像描述。通过引入关键词的方式进行局

部语义学习，以解决图像描述的多样性问题并满足个性化需求。提出了一种新的 CDB-LSTM 语言模型，能够根据关键词从全局视角生成具有不同侧重点的图像内容描述。模型由两个级联的子模型生成，子模型分别基于关键词进行局部语义学习并通过上下文关系在端到端框架中统一起来。该模型能够灵活地处理各种关键词并生成对应的特定描述，自动评测和人工评测都证明了该模型的优越性。

第 6 章简要总结了本论文的主要工作，并展望了未来的研究方向。

第 2 章 图像描述相关技术背景

本章从六个方面详细介绍图像描述相关的背景知识：特征表达、语言生成模型、图像描述模型、目标检测、数据集和评估方法。

2.1 特征表达

图像描述的特征表达涉及视觉特征表达和语句特征表达两部分。

2.1.1 视觉特征表达

图像视觉特征学习是计算机视觉领域的一项基础性工作，是目标检测、图像分类、场景分类等视觉任务的核心问题。原始的 RGB 或灰度图像在计算机中以矩阵像素的形式存储。矩阵可表示为 $W * H * C$ （宽度，高度，通道），其中 RGB 通道为 3，灰色图像为 1。数字矩阵中像素为 0~255 之间的值。矩阵像素值构成了原始视觉信息，通常不能很好地表达高层语义信息。为完成不同的视觉任务，需要进行图像特征提取，接下来从两个方面来介绍可视图像的特征表达：手工设计的特征表达、深度学习特征表达。

1) 手工设计特征表达

底层特征直接反映原始图像的基本特征，例如颜色、纹理和形状，这些特征通过对图像像素进行手工设计的固定方法获得。Szumner 等^[69]使用 MSAR 纹理特征和 Ohta 空间颜色直方图特征来表示图像。Lowe 等^[70]提出了尺度不变特征变换（SIFT）来描述图像的局部特征。SIFT 对图像缩放、平移和旋转具备不变性，已被广泛应用于目标识别、图像匹配和目标追踪等。其他底层特征如 HOG^[12]、LBP^[71]、textons^[72]、颜色直方图^[73]等也广泛用于各种计算机视觉任务。尽管底层特征的应用在视觉特征表达方面获得了很大的成功，但用来表示具备高层语义信息的图像是远远不够的。视觉语义鸿沟^[74-76]加速了其他视觉特征的进展。为了缩小底层特征和高层概念之间的差距，Bag of Words^[77]、Latent Dirichlet Allocation^[78]、Fisher vector^[79]和 Object bank^[80]等中层设计特征引起了广泛关注。

在本论文中，根据面临的具体问题和提出的方法，采用词袋法(Bag of Words,

BoW)生成中层特征，在此对其进行详细介绍。

Bag of Words 最早用于文本处理，其忽略掉文本的语法和语序，用一组无序的单词(Words)来表达一段文字或一个文档。在文本检索中，每个文档可表达为一个词向量。假设有一个 v 个单词的词汇表，那么每个文档都由一个向量表示：

$$\begin{aligned} v_d &= (t_1, \dots, t_i, \dots, t_v)^T \\ t_i &= \frac{n_{id}}{n_d} \log \frac{N}{N_i} \end{aligned} \quad (2.1)$$

其中 t_i 统计词汇表中单词 i 出现在文档 d 中的次数， n_{id} 是文档 d 中单词 i 出现的次数， n_d 是文档 d 中的单词总数， N_i 是包含单词 i 的文档个数， N 是整个数据库中的文档数量。

Josef Sivic 和 Andrew Zisserman^[77]首先提出了视觉词汇的概念，之后 BoW 被广泛应用于计算机视觉中。与应用于文本的 BoW 类比，视觉中将底层的图像特征（如 SIFT、HOG 等底层特征）看作文本中的单词(Words)，进而生成 BoW 中层特征，这个中层特征能够表达更多的底层信息。

因为视觉描述子被量化为视觉词汇，BoW 模型大大减少了计算机视觉系统的计算负担和内存占用。由于其效率和性能，BoW 在图像检索和分类领域非常流行^[81-83]。

2) 深度学习特征表达

通常在大规模图像分类任务中学习到的 CNN 模型被广泛用于视觉特征提取。和基于手工设计的特征相比，CNN 特征在视觉表达上是强大有效的，并可以用于各种视觉任务^[19]。Zeiler 和 Fergus^[19]使用去卷积网络方法将 CNN 的特征层映射投影到像素空间，特征图的可视化结果直观上显示了其良好的表达特性。AlexNet^[84]，VGGNet^[4]，GoogLeNet^[55]和 ResNet^[85]上常用的 CNN 框架，代表着深度学习特征的发展和进步，并且在与视觉表达相关的任务中占主导地位，本文中同样主要采用深度特征来进行视觉表达。

2.1.2 语句特征表达

对于图像 I 及其相应的具有 N 个单词 $Y = \{y_1, y_2, \dots, y_N\}$ 的真实标注语句，其中 y_i 表示一个单词，如“dog”。语句特征表达实际上是按顺序排列的单词的特征表达。这些单词需要以特定的方式表达成为计算机可以处理的特征。单词表达一般使用两种策略：**One-Hot** 特征表达和 **Word Embedding** 特征表达。

1) One-Hot 特征表达

One-Hot（独热编码），又称一位有效编码，其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有独立的寄存器位，并且在任意时候，其中只有一位有效。在自然语言处理中，**One-Hot** 向量是一个 $1 \times n$ 矩阵，用于区分词汇表中的 N 个单词。除了单独用于识别单词的单元中的 1 之外，向量中的其他单元由 0 组成。在实际应用中，每个单词通过唯一的索引表示为一个 **One-Hot** 向量。例如，单词“dog”可以表示为 $\{0, 0, 1, 0, 0, 0, \dots, 0\}$ ，“cat”可以表示为 $\{0, 0, 0, 0, 1, 0, \dots, 0\}$ 。这种方法的缺点是词汇表太大容易导致 **One-Hot** 向量过长，优点是向量很容易设计和修改，编码效率高。

2) Word Embedding 特征表达

词嵌入(**Word Embedding**)是自然语言处理中一组语言建模和特征学习技术的集合名称，该方法将词汇表中的单词或短语映射到实数向量。将每个单词映射到具有固定维度的向量需要训练支持单词嵌入的模型。**Word2vec**^[86]，**GloVe**^[87] 是这一领域较为流行的方法。

与 **One-Hot** 相比，**Word Embedding** 不受词汇量大小的影响，特征维度偏低，如 256、512 等，而 **One-Hot** 特征维度往往大于 1 万。此外，**Word Embedding** 方法可以通过计算两个单词之间的距离来反映其相似度。

在本论文中，第 3 章使用 **One-Hot** 对语句中的单词进行特征表达。第 4 章和第 5 章使用了 **One-Hot** 和 **Gensim**^[88] 词向量这两种表示方法进行语句编码。

2.2 语言生成模型

语言建模是根据前面的单词，预测接下来哪个单词出现，这是自然语言处理中的一个重要概念，可以追溯到 1951 年的工作^[89]，**Claude Shannon** 考虑了一

串输入符号被逐一考虑的情况，通过判断猜测的困难度来衡量下一个输出的不确定性。

语言通常是一组单词的序列表示。语言建模利用序列内所有历史词汇，通过逐个单独预测每个词来定义序列（单词） $y=y_1, y_2, \dots, y_N$ 生成的概率：

$$p(y) = \prod_{t=1}^{t=N} p(y_t | y_1, y_2, \dots, y_{t-1}) \quad (2.2)$$

其中 y_t 表示位置 t 处的单词， N 表示 y 中的单词数量， $p(y_t | y_1, y_2, \dots, y_{t-1})$ 表示给定前 $t-1$ 个单词，即 y_1, y_2, \dots, y_{t-1} ，预测到单词 y_t 的条件概率。

在实际应用中，由于数据量 N 可能极为庞大，因此，广泛使用 **n-gram** 语言模型，即用 $n-1$ 之前的单词来近似 $t-1$ 个历史数据 ($n < t$)。然后，根据相对频率计数来估计条件概率，统计 $y_{t-n+1}, \dots, y_{t-1}$ 出现的次数，以及计算在此基础上出现 y_t 的次数：

$$p(y_t | y_1, y_2, \dots, y_{t-1}) = \frac{C(y_{t-n+1}, \dots, y_{t-1}, y_t)}{C(y_{t-n+1}, \dots, y_{t-1})} \quad (2.3)$$

n-gram 语言模型具有简单和能够快速实现的优点，但是同时存在一些严重的问题，例如受数据稀疏性的影响、对新单词的泛化能力差、对模型存储空间的巨大需求、以及只能考虑 4-6 个上下文单词而无法处理单词的长期依赖性等。

神经语言建模(NLM)^[90-96]提供了一种更好的方式来处理数据稀疏和上下文依赖的问题。在 NLM 中，每个单词用不同的 K 维分布向量表示。语义上相似的词在向量空间中占据相似的位置，显著地缓解了数据稀疏问题。NLM 将上下文单词的向量 y_1, y_2, \dots, y_{t-1} 表示作为输入，并将它们映射为向量表示形式 h_{t-1} ：

$$h_{t-1} = f(y_1, y_2, \dots, y_{t-1}) \quad (2.4)$$

其中 f 表示映射函数，它通常是一个前馈神经网络模型，如循环神经网络或卷积神经网络。给定历史上下文单词，预测词 y_t 的条件概率分布如下：

$$\begin{aligned}
 p(y_t | y_1, y_2, \dots, y_{t-1}) &= p(y_t | h_{t-1}) \\
 s_t &= W[y_t, :] \cdot h_{t-1} \\
 p(y_t | h_{t-1}) &= \exp(s_t)
 \end{aligned}
 \tag{2.5}$$

其中权重矩阵 $W \in \mathbb{R}^{V \times K}$, V 是词汇表大小, K 是词向量表示的维度。 $W[y_t, :]$ 表示矩阵的第 y_t 行, \exp 表示 softmax 函数。 softmax 函数将标量向量 s_t 映射为概率分布向量, 如公式 2-6 所示:

$$\text{softmax}(s_t) = \frac{\exp(s_t)}{\sum_{s \in V} \exp(s)}
 \tag{2.6}$$

由于上下文的维度不受上下文长度变化的影响, 理论上, NLM 能够容纳无限数量的上下文单词而无需存储全部不同的 n-gram 信息。 本文采用该语言生成模型, 并简要介绍属于这类模型的循环神经网络 RNN 及其变体 LSTM。

2.2.1 RNN

循环神经网络 RNN^[97]是专门设计用于处理序列数据的神经网络架构。历史上多用于处理时间序列数据^[98, 99], 并已被成功应用于语言处理中^[93, 100-102]。RNN 模型为仅有输入层 x 、隐藏层 h 和输出层 y 的三层循环结构, 展开后是一个序列处理模式, 如图 2.1 所示:

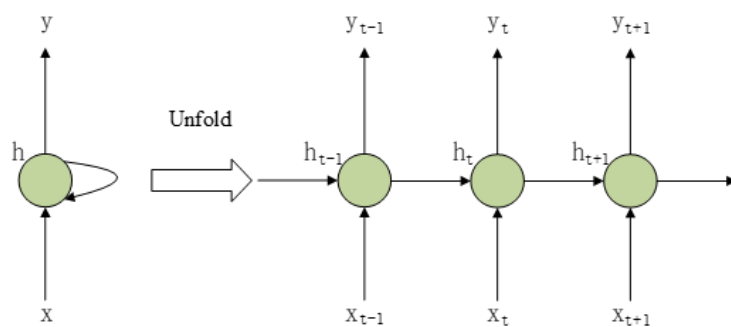


图 2.1 RNN 及其展开结构

给定序列输入 $\{x_1, x_2, \dots, x_t\}$, RNN 能够依次输入 x_t , 通过隐藏层 h_t 输出 y_t , 最终生成输出序列 $\{y_1, y_2, \dots, y_N\}$ 。

RNN 通过隐藏层 h_t 进行序列记忆和处理, 隐藏层 h_t 可以被认为是嵌入了时刻 t 之前输入的所有信息, 即 $\{x_1, x_2, \dots, x_t\}$ 。在数学形式上, h_t 是通过函数 f 获得的, f 为前一时刻 $t-1$ 的隐藏层 h_{t-1} 与当前时刻的输入 x_t 之间的函数:

$$h_t = f(h_{t-1}, x_t) \tag{2.7}$$

函数 f 可以采用不同的形式:

$$f(h_{t-1}, x_t) = \sigma(W_{hh}h_{t-1} + W_{xh}x_t) \tag{2.8}$$

其中 W_{hh}, W_{xh} 同样为权重矩阵, σ 的选择一般为非线性函数, 如 sigmoid、tanh 或 ReLU 等。

RNN 的输出层依然为隐藏层的 softmax 函数::

$$p(y_t) = \exp(W_{hp} \cdot h_t) \tag{2.9}$$

其中 W_{hp} 为输出层的权重矩阵, 将 h_t 映射为和单词空间同维度的向量, 经 softmax 计算后, 条件概率最大的值即为当前预测结果。

循环神经网络可以灵活处理各种序列学习问题, 如图 2.2 所示, 通过在输入端、输出端的设计, 循环神经网络可处理各种序列模式: 一对一、一对多、多对一、多对多问题。

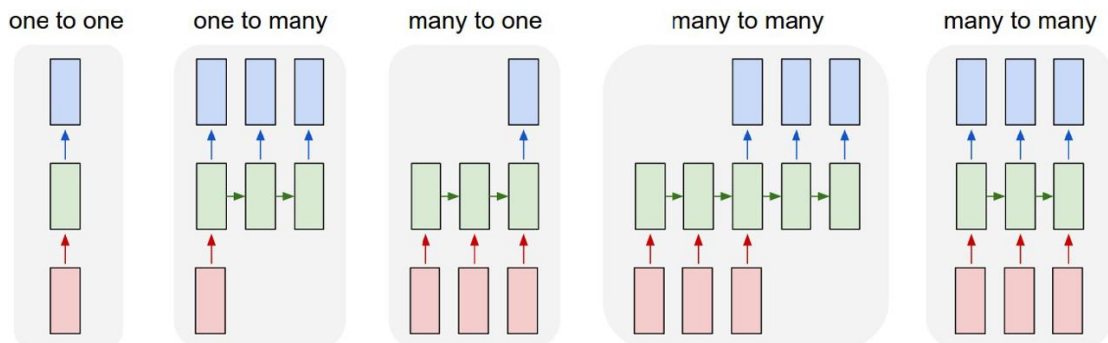


图 2.2 循环神经网络序列处理模式: 输入向量 (红色), 隐藏层 (绿色), 输出向量 (蓝色)。

本文工作主要用图 2.2 中第四种和第五种多对多模式进行语言模型设计。

2.2.2 LSTM

LSTM 可以视为 RNN 的一个变体，其输入层、隐藏层和输出层的结构一致，不同的是隐藏层记忆单元比普通的 RNN 更为复杂，在语言生成上也更为强大。传统的 RNN 训练过程中受梯度消失和梯度膨胀问题的限制^[103]。梯度膨胀指随着时间序列的推移，来自训练目标函数的误差在反向传播时梯度逐渐累积变得非常大的情况；梯度消失指训练误差随着时间序列反向传播梯度趋近于零的情况。这两个问题使得 RNN 模型无法记忆长序列数据的长期上下文信息，也使得模型训练无法正常完成。在 RNN 基础上的长期短期记忆模型(LSTM)是缓解这两个问题最有效的方法之一。LSTM 及其变体在近 20 年来被不断应用和探索^[24, 28, 52, 64, 104-113]，已被广泛应用于图像描述、机器翻译、语音识别、视频描述、视觉问答、图像识别等各个领域。

LSTM 的关键思想是将每个时刻与不同类型的控制门相关联，这些门提供了灵活控制信息流的方式：通过遗忘门(forget gate)来控制当前循环想要保存的信息量，通过输入门(input gate)来控制当前循环想要接收多少当前输入信息，通过输出门(output gate)控制当前循环想要输出到下一个时刻的信息量。相对于传统的 RNN 模型，LSTM 通过门控制单元能够记录相隔较长的上下文信息变化，并且能够解决传统 RNN 训练中出现的梯度消失和梯度膨胀问题。

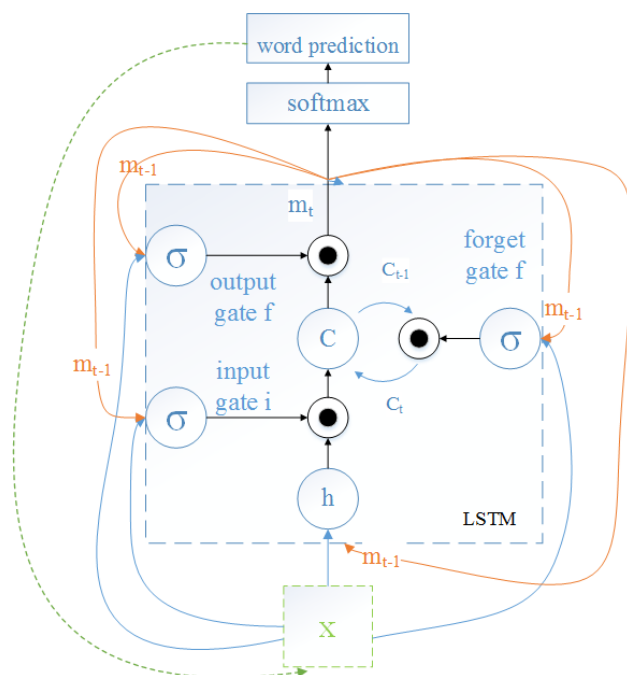


图 2.3 LSTM 模型内部结构

如图 2.3 所示, LSTM 架构的核心由三个门控制: 输入门 i , 遗忘门 f , 输出门 o 。通过三个门的控制, 隐藏层 m 和记忆单元 c 包含了更多的上下文信息。

给定一系列输入 $\{x_1, x_2, \dots, x_N\}$, 每个时刻 t 的隐藏状态 m_t 计算公式为:

$$\begin{aligned}
 i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
 f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
 o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
 m_t &= o_t \odot c_t
 \end{aligned} \tag{2.10}$$

其中变量 i_t, f_t, o_t, c_t, m_t 分别代表输入门、遗忘门、输出门、存储单元和隐藏状态, $W_{ix}, W_{fx}, W_{ox}, W_{cx}, W_{im}, W_{fm}, W_{om}, W_{cm}$ 是训练模型的权重, σ 表示 sigmoid 函数, \odot 表示内积, h 表示双曲正切函数。

由于 LSTM 的长期短期记忆功能比基本 RNN 有显著的优越性, 本文工作中采用的循环神经网络语言模型内核都为 LSTM。

2.3 基于 CNN-RNN 的图像描述模型

近两年在图像描述领域占主导地位的是基于 CNN-RNN 的方法, 该方法先用卷积神经网络 CNN 提取图像视觉特征, 然后用循环神经网络 RNN 来进行语言生成, 如图 2.4 所示。这种方法源自于机器翻译领域, 即将一种语言通过编码和解码的方式转换为另外一种语言。在图像描述领域, 则先将视觉图像编码为特征, 然后通过语言模型将特征解码为自然语言。

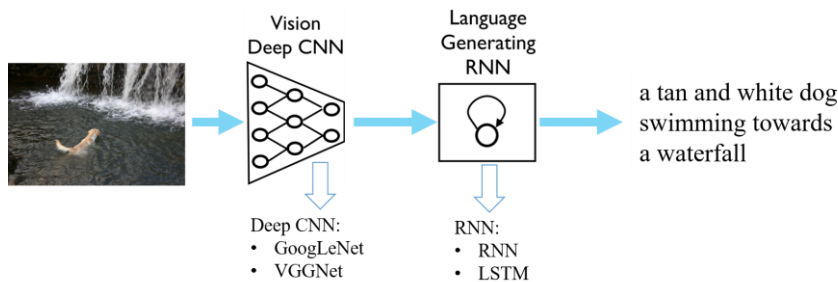


图 2.4 基于 CNN-RNN 的图像描述方法

基于 CNN-RNN 的图像描述方法很好地利用了深度学习的优点：CNN 强大的视觉表达能力、RNN 优越的序列数据处理能力、深度学习端到端的统一训练框架。这些优点使得基于 CNN-RNN 的方法生成的描述语句更加自然、灵活、准确，在图像描述领域很快占据了统治地位。近年来的工作基本都沿用了这一方法，不同的是所采用的 CNN 网络是 GoogLeNet、VGGNet 或其他框架，RNN 网络是基本 RNN、LSTM 或其他变体。

给定一幅图像 I ，图像描述中语言生成任务的目标函数为最大化输出概率 p 的对数似然和：

$$\theta^* = \arg \max \sum \log p(S|I; \theta) \quad (2.11)$$

其中 θ 为模型参数， S 为人工标注的语句描述。 θ^* 的求解通过随机梯度下降(Stochastic Gradient Descent, SGD)的方法训练反馈网络得到。

在具体应用中，图像描述任务首先进行视觉特征表达：

$$v = CNN(I) \quad (2.12)$$

其中 CNN 网络是 GoogLeNet、VGGNet 或其他深度网络。特征表达 v 将作为 RNN 的输入进行语言模型训练或测试。

第 2.2 节已经介绍了 RNN 模型的结构。RNN 擅长处理序列数据，它在机器翻译和自然语言处理领域得到了广泛的应用，因此在图像描述领域被用来进行语言生成。对每幅图像，RNN 逐字处理序列数据，即每次循环仅预测一个单词，然后利用当前生成的单词和隐藏信息进行下一个单词的预测，直到预测到一个代表结束的词为止。

2.3.1 RNN 语言模型训练

给定图像 I 的 CNN 特征 v 及其对应的具有 $N-1$ 个单词的描述语句 $S = \{s_1, s_2, \dots, s_{N-1}\}$ ，RNN 的训练过程为：

$$\begin{aligned}
 x_0 &= W_{ix}v, t = 0 \\
 x_1 &= W_{ex}s_0, t = 1 \\
 x_t &= W_{ex}s_t, t \in \{2, \dots, N\} \\
 h_t &= RNN(h_{t-1}, x_t), t \in \{1, \dots, N\} \\
 y_t, p_t &\propto \exp(W_{hp}h_t), t \in \{1, \dots, N\}
 \end{aligned} \tag{2.13}$$

其中 x_t 为特征编码后的 RNN 模型输入向量， y_t 为模型输出， W 表示相应的线性编码或解码权重：视觉特征编码 W_{ix} ，单词编码 W_{ex} ，输出解码 W_{hp} 。 s_0 代表固定启动词“0”，可以作为一个单词看待。在实验中， s_t 使用 One-hot 编码进行单词特征表达。

由于 $t=0$ 时刻模型输入为 CNN 特征 v ，隐藏层初始化参数值皆为 0，不计算输出，因此定义开始产生输出的时刻为 $t=1$ ，即输入启动单词“0”时，之后的循环依次输入标注语句中的单词 s_t ，通过隐藏层 h_t 进行语义信息记忆和传递，并实现循环过程。

模型损失函数为每次循环预测到的单词的负对数似然和：

$$L = -\sum_{t=0}^{N-1} \log p(s_t | x_t; \theta) \tag{2.14}$$

2.3.1 RNN 语言模型测试

语言模型训练结束后，获得各参数值，模型测试过程是将训练过程中的标注单词 s_t 替换为前一刻的输出 y_t ，直到输出结束词“0”为止：

$$\begin{aligned}
 x_0 &= W_{ix}v, t = 0 \\
 x_1 &= W_{ex}s_0, t = 1 \\
 h_t &= RNN(h_{t-1}, x_t), t \geq 1 \\
 y_t, p_t &\propto \exp(W_{hp}h_t), t \geq 1 \\
 x_{t+1} &= W_{ex}y_t, t \geq 1
 \end{aligned} \tag{2.15}$$

最终的输出结果为 $Y = \{y_1, y_2, \dots, y_{N-1}\}$ ，通过 One-hot 反编码即可得到描述语句。

在本文第 3~5 章的模型训练和测试流程及公式都与本节所介绍的语言模型

形式相关。

2.4 目标检测

目标检测(Object detection)^[12-14, 114-119]是计算机视觉领域最重要的任务之一,在过去的十年中已经得到了广泛的研究。该任务的目标是查找图像中的所有主要目标,输出每个目标的标签和位置。

长期以来,目标检测方法采用滑动窗口和人工设计的描述符的框架^[12, 13],它在 PASCAL VOC^[120]视觉挑战上取得了巨大的成就。Dalal 和 Triggs^[12]采用梯度方向直方图特征(Histogram of Oriented Gradients, HOG)进行特征提取,并利用 SVM 进行特征分类,在行人检测上获得了巨大的成功。Felzenszwalb 等^[13]进一步在 HOG 的基础上扩展可变形的组件模型(Deformable Part Model, DPM),使得目标检测方法上升到了新的高度。卷积神经网络的巨大进展^[84]推动了目标检测方法的进步,最先进的检测方法已经被 R-CNN 系列所取代: R-CNN^[14], Fast R-CNN^[115], Faster R-CNN^[118], Mask R-CNN^[121]。这类方法都是基于提取到的候选窗口(Region Proposal)的分类结果进行目标检测,获得了较好的性能,本章主要采用 R-CNN 系列进行目标检测有关实验,而其他先进的基于回归的方法如 Yolo^[122]、SSD^[123]等性能还不能超越 R-CNN 系列方法,所以不在本文工作的考虑范围。

R-CNN^[14]对在窗口选择上使用 Selective Search^[124],视觉特征表示使用 CNN 模型^[84],对图像分类使用 SVMs^[125]。R-CNN 的出现改变了近几年目标检测领域的研究方向,成为后续研究的里程碑式工作。其优点是性能有了显著提高,缺点是检测速度很慢,平均每幅图像需要 3s 左右。

Fast R-CNN^[115]是对 R-CNN 和 SPPnet^[116]的融合,使得对每幅图片只需要提取一次 CNN 特征,而不是 R-CNN 的平均 2000 次。在测试时,SPPnet 将 R-CNN 加速 10 到 100 倍。由于更快的区域特征提取,训练时间也减少了 3 倍。

Faster R-CNN^[118]将 Fast R-CNN 与区域提取网络(region proposal networks, RPN)合并为一个统一的框架,从而提高了区域提取的质量,并提高了整体目标检测的准确性。

Mask R-CNN^[121]通过添加一个分支来预测目标掩码,并与现有分支进行边

界框识别，从而扩展了 Faster R-CNN。

由于现阶段的图像描述为基于目标的语言描述，目标检测位置代表了图像的显著性区域，基于这些区域的语义研究对于图像描述是有重要意义的。在本文中，R-CNN 和 Faster R-CNN 被用来检测目标并生成局部区域，以进一步支撑局部语义学习。

2.5 数据集

本节简要介绍图像描述领域内常用的数据集，包括收集数据集的常用方法、语句标注特点、数据量等。近几年来发布了大量的数据集^[48, 68, 126-134]用于自动图像描述研究。这些数据集集中的图像与文本描述相关联，并且在某些方面彼此不同，例如大小、描述的格式以及如何收集描述。接下来简要介绍最常用的六种数据集，如表 2.1 所示。

表 2.1 图像描述数据集

数据集	图像个数	标注语句/幅
Pascal1K ^[132]	1000	5
SBU ^[48]	1000000	1
Flickr8k ^[129]	8108	5
Flickr30k ^[133]	31783	5
MS COCO ^[130]	164062	5
Visual Genome ^[68]	108249	50

数据集样本如图 2.5 所示。

Pascal1K 语句数据集^[132]通常被作为评估描述生成系统质量的基准。这个数据集包含 1000 幅从 Pascal 2008 目标识别数据集^[135]中选择的图像，并包含来自不同视觉类的目标，例如人类、动物和车辆等。每张图像都与 Amazon Mechanical Turk (AMT) 服务平台上由人类生成的五种描述相关联。

SBU 图像描述数据集^[48]包含 100 万张带有原始用户生成描述的图像，通过查询特定单词（如目标和动作）、过滤搜集具有特定平均长度描述的图像，在 Flickr 上构建数据集。这个数据集比较庞大，但是每幅图像只有一句标注，并

且不能保障整体标注格式的一致性。



图 2.5 图像描述数据集展示

Flickr8k 数据集^[129]及其扩展版本 Flickr30k 数据集^[133]也都来自 Flickr，分别包含约 8000 和 31000 幅图像。这两个数据集中的图像是通过针对特定目标和动作的用户查询来选择的。这些数据集包含 5 个描述，每幅图像使用类似于 Pascal1K 数据集的策略从 AMT 中收集。公共的划分方法^[49]中，Flickr8k 一般是 6000 幅用来训练，1000 幅用来验证，1000 幅作为测试图像；Flickr30k 一般是 29000 幅作为训练图像，1000 幅验证，1000 幅测试。Flickr30k Entities^[131]将每个目标及其属性一起对应到图像中的局部区域，通过实体关联使得图像语义描述更为详细，增强了原始 Flickr30k 数据集。

MS COCO 数据集^[130]中图像描述数据目前包含 123287 个图像，82783 幅训练图像和 40504 幅验证图像，每幅图像有五种不同的描述。模型验证一般也遵循常用的公共划分方法^[49]，来自训练集的所有 82783 幅图像用于训练，来自验证集的 5000 幅图像用于验证、5000 幅用于测试。每个描述都是使用 AMT 上的人工标注生成的，参与标注者被要求：

- 1) 不要以 “There is” 为开始标注语句;
- 2) 不要描述图像中不重要的细节;
- 3) 不要描述图像中包含的可能发生过或未来将会发生的事情;
- 4) 不要描述图像中的某个人可能会说什么;
- 5) 不要给图像中的人(people)命名;
- 6) 语句应至少包含 8 个单词。

MS COCO 数据集的数据量要远大于前面两种数据集, 另外, MS COCO 数据集还包含了目标位置和 80 个目标类别的标注, 使得该数据集含有更多的可用信息。该数据集促成了 2015 年 MS COCO Captions Challenge (<http://MS-COCO.org/dataset/#captions-challenge2015>), 并且成为视觉和语言研究的基准。

Visual Genome^[68]包含超过 100K 图像的目标、属性、关系和局部描述的密集标注。每幅图像平均包含 42 个局部描述, 每个区域由一个边框进行定位。当描述不同时, 区域允许彼此高度重叠。每个描述都是描述该区域的长度范围从 1 到 16 个字的短语。除了 420 万对区域的描述外, Visual Genome 的标注还含有 170 万视觉问答、210 万目标、180 万属性、180 万关系。这个数据集对于图像描述来说更加详尽, 对于后面的研究工作有重要的意义。

在这些数据集中, 由于对图像描述的定义和标注方法相似, 以及数据量大小的区别, Flickr8k、Flickr30k 和 MS COCO 被更广泛的用于图像描述研究。Visual Genome 相对较为新颖, 也逐渐获得了更多的关注。本文采用 Flickr8k、Flickr30k、MS COCO 和 Visual Genome 进行方法评估。

2.6 评估方法

评估自然语言生成模型的输出是一项艰难的任务^[136]。图像描述的评估极为衡量待评测语句和参考语句之间的相似度, 其中广泛使用两种方法进行评估: 人工评价标准和自动评价标准。

2.6.1 人工评价标准

评估自动生成文本质量的最常用方法是由专家进行主观评价。对于描述生成任务, 已经为图像描述提出了许多不同的评估方案: Ordonez 等^[48]给评估专家提供模型的描述结果, 并要求他们在随机图像和描述的图片之间进行强制选

择。Kuznetsova 等^[45]要求评估者在给定的测试图像中对两个模型的结果进行选择，这种强制选择任务可能会给出明确的模型排名，但不能在不同的实验中进行比较，也不能直接衡量描述的质量。Yang 等^[41]和 Kulkarni 等^[39]对描述进行相关性和可读性的分级评估，Li 等^[137]增加了一个“创造力”得分。Mitchell 等^[40]通过是否描述图像的“主要方面”、是否以适当的顺序引入目标、语义上是否正确、以及是否由人书写来进行描述系统评估。

Hodosh 等^[129]提出将图像描述评估作为一项对给定多个描述进行排名的任务，语句的描述性以四分制评分：分数为 4 意味着语句完美地描述了图像（没有任何错误）、分数为 3 意味着语句几乎表达了图像内容（允许有小错误，例如实体数量），而分数为 2 表示语句仅描述了图像的某些方面，但不能作为图像描述，分数为 1 表示描述与图像无关。将图像描述定为排名任务的一个优势是可以将不同的系统生成结果直接进行比较。四分制评分策略已被广泛用于描述领域^[52, 129, 138, 139]。本文遵循 Hodosh 等^[129]提出的方案，在第 5 章中的实验中要求志愿者评估每个生成的语句，打分范围从 1 到 4。

2.6.2 自动评价标准

由于人工评价的获取方式昂贵且费时，BLEU^[140]，METEOR^[141]，CIDEr^[142]和 ROUGE_L^[143]等自动评价标准被广泛用于描述评估。除了专门为图像描述评估开发的 CIDEr 之外，这些措施最初是为了评估机器翻译引擎或文本摘要系统的输出而开发的。所有这些度量都计算出一个分数，该分数表示系统输出与一个或多个人工标注的参考文本（例如，真实标注翻译或摘要）之间的相似度。

1) BLEU

BLEU(Bilingual Evaluation Understudy)^[140]是一种流行的机器翻译评价指标，也是目前图像描述领域中最常用的度量标准之一，其主要部分是针对参考描述对评测语句进行 n 元组(n -gram)准确率计算。对于图像 I_i ，模型会生成对应的标注语句 c_i ，自动评价标准能够根据参考标注语句（也就是人工标注的语句）的一个集合 $S_i = \{s_{i1}, s_{i2}, \dots, s_{im}\}$ ，对待评价的标准语句 c_i 的质量做出评价。标注语句都是用 n 元组来表示的，一个 n 元组 $w_k \in \Omega$ 是由一个或者多个有顺序单词

组成的序列。 n 元组 w_k 在语句 s_k 中出现的次数被记为 $h_k(s_{ij})$ ， n 元组 w_k 在待评价语句 c_i 中出现的次数被记为 $h_k(c_i)$ 。

首先计算的是全局的 n 元组精度：

$$CP_n(C, S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)} \quad (2.16)$$

其中 k 指的是长度为 n 的可能的 n 元组的集合数。由于 BLEU 倾向于更短的句子，即待评价句子 c_i 较短时，评价得分会较高。为了解决这个问题，经常使用乘以一个简洁性惩罚来防止短句获得高分。令 l_s 为参考句子的总长度， l_c 是待评价句子的总长度，惩罚项为：

$$b(C, S) = \begin{cases} 1, & \text{if } l_c > l_s \\ e^{1-l_s/l_c}, & \text{if } l_c \leq l_s \end{cases} \quad (2.17)$$

最终计算 BLEU 分数：

$$BLEU_N(C, S) = b(C, S) \exp\left(\sum_{n=1}^N w_n \log CP_n(C, S)\right) \quad (2.18)$$

通常使用 4 元组，即 $N=\{1,2,3,4\}$ 。对于所有的 n ， w_n 都是常量，一般为 $1/n$ 。BLEU 在语料库层级上具有很好语句匹配表现，但随着 n 的增加，在句子层级上的匹配越来越差。另外，BLEU 只考虑精确度，不考虑召回率，使得其评价方法并不完善。

2) METEOR

METEOR^[141]方法基于一元组的精度和召回的调和平均（Harmonic mean）来计算 F-mean，且召回的权重比精度大。将待评测语句和参考语句进行单词校准对齐（alignments），则计算公式如下：

$$\begin{aligned}
P_m &= \frac{|m|}{\sum_k h_k(c_i)} \\
R_m &= \frac{|m|}{\sum_k h_k(s_{ij})} \\
F_{mean} &= \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m}
\end{aligned} \tag{2.19}$$

其中 $|m|$ 为校准匹配上单词的个数， P_m, R_m 分别是精度和召回率。公式只对单个单词的一致性进行了衡量，却没有对参考句子和待评价句子中更大的分段进行衡量。为了将其计算在内，使用更长的 n 元组来计算对于校准的惩罚。在参考和待评价句子中没有毗连的映射越多，惩罚就越高。为了计算惩罚，1 元组被分组成最少可能的块 ch (chunks)。块的定义是在待评价语句和参考语句中毗邻的一元组集合。在待评价语句和参考语句之间的毗邻映射越长，块的数量就越少。一个待评价翻译如果和参考翻译相同，那么就只有一个块。惩罚 Pen 的计算如下：

$$Pen = \gamma \left(\frac{ch}{m} \right)^\theta \tag{2.20}$$

最终的评测得分为：

$$METEOR = (1 - Pen) F_{mean} \tag{2.21}$$

和 BLEU 不同，METEOR 同时考虑了基于整个语料库上的准确率和召回率，而最终得出测度，其与人类判断相关性较高，评测得分往往较低。

3) ROUGE

ROUGE 是一个设计用来评价文本摘要算法的自动评价标准集，其中有 3 个评价标准，分别是 ROUGE-N，ROUGE-L 和 ROUGE-S，在图像描述评价中较常用的是 ROUGE_L。ROUGE_L 是基于 longest common subsequence (LCS) 的一种测量方法。所谓 LCS，就是一个同时出现在两个句子中的单词集合，且单词出现的顺序也是相同的。和 n 元组不同的是，在单词之间可能还存在能够创建出 LCS 的单词。将比较的两个句子间的 LCS 的长度记为： $l(c_i, s_{ij})$ 。

ROUGE-L 通过计算 F-mean 来求得:

$$\begin{aligned}
 P_l &= \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \\
 R_l &= \max_j \frac{l(c_i, s_{ij})}{|c_i|} \\
 ROUGE_L(c_i, S_i) &= \frac{(\beta^2 + 1)P_l R_l}{\beta^2 P_l + R_l}
 \end{aligned} \tag{2.22}$$

其中 P_l, R_l 分别是精度和召回率, β 一般等于 1.2, 在这个计算中不需要考虑 n 元组。

4) CIDEr

CIDEr 是图像描述领域特有的评价标准, 它是通过对每个 n 元组进行 Term Frequency Inverse Document Frequency (TF-IDF) 权重计算, 来衡量图像标注的一致性的。CIDEr 为每个 n 元组都计算 TF-IDF 权重:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_j \in \Omega} h_k(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right) \tag{2.23}$$

其中 Ω 是所有 n 元组的词汇表, I 是数据集中所有图像的集合。公式第一个部分计算的是每个 n 元组 w_k 的 TF, 公式第二部分是使用 IDF 来计算 w_k 的稀有程度。从直观上来说, 如果一些 n 元组频繁地出现在描述图像的参考标注中, TF 对于这些 n 元组将给出更高的权重, 而 IDF 则降低那些在所有描述语句中都常常出现的 n 元组的权重。也就是说, IDF 提供了一种测量单词显著性的方法, 这就是将那些容易常常出现, 但是对于视觉内容信息没有多大帮助的单词的重要性打折。

IDF 的计算方法是: 分子为数据集中图像的数量 I , 分母为一些图像的数量, 这些图像是其任意一个描述中出现了 n 元组 w_k 的图像, 然后再对这个分数求对数。对于长度为 n 的 n 元组的 CIDEr- n 分数是使用待评价句子和参考句子之间的平均相似性来计算的, 其中精度和召回率都要占比例:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad (2.24)$$

其中， $g^n(c_i)$ 是一个由 $g_k(c_i)$ 生成的向量，对应的是所有长度为 n 的 n 元组。 $\|g^n(c_i)\|$ 是向量的大小。而 $g^n(s_{ij})$ 的情况类似。更长的 n 元组是用来获取语法性质和更丰富的语义信息的。不同长度的 n 元组的得分计算如下：

$$CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i) \quad (2.25)$$

和 BLEU 一样，权重 $w_n = 1/N$, $N = 4$ 较为常用。

本章采用的自动评测方法为以上四种标准，也是图像描述领域最常用的模型评价方法。

2.7 本章小结

本章从六个方面详细介绍了相关背景知识：特征表达、语言生成模型、图像描述基本模型、目标检测、数据集和评估方法。本文后面的工作都建立在这些背景知识的基础上，具体实施细节将会在各章节结合具体任务详细介绍。

第3章 基于局部目标区域的图像描述

图像内容自动描述是人工智能领域中连接计算机视觉技术和自然语言处理技术的一个基本研究问题。与现有的描述整幅图像内容的方法不同，本章提出了一种根据图像局部区域生成更加丰富的图像描述的方法。首先使用目标检测方法来生成候选区域；然后训练 RNN 语言模型以学习全局图像和标注语句之间的描述关系；最后，利用描述模型对目标区域进行局部描述生成和分析。所提出的方法能够对单幅图像生成针对多个不同区域的语句描述，这些描述具有足够的表达能力并且包含更详细的语义信息。实验评估验证了所提出方法的有效性。

3.1 引言

图像包含丰富的语义内容，使用自然语言自动描述图像的主要内容是一项极具挑战性的任务。图像描述一方面应该准确地提供目标相关语义表达以避免传递错误信息，另一方面，描述应尽可能多地覆盖图像内容以避免错过细节概念。与计算机视觉领域的传统图像分类和目标识别任务相比，图像描述不仅要捕捉包含在图像中的目标，还需要表达这些目标的属性、动作、行为等，最终这些语义知识必须用人类可理解的自然语言（如英语）表达出来。这意味着除了视觉理解外，图像描述还需要合适的语言模型来进行语句生成。

近年来图像描述领域的大部分工作^[39, 40, 42, 52]只关注整幅图像的内容表达，而忽视了图像自身包含的丰富信息，更是无法满足个性化需求。Kapathy 和 Li^[49]提出的多模态 RNN 模型能够为标注图像区域生成短语或单词，如图 3.1 中第一幅图像所示的“large white statue”、“building”、“man in suit”等，是在给定区域边框下生成的。密集描述(Dense Captioning)^[57, 58]进一步为自动预测的区域候选框生成局部短语表达，如图 3.1 中第二幅图像所示的“roof of a building”、“large green trees”、“elephant is standing”等，是对自动生成的区域的表达，与之相比，图 3.1 中第一幅图像的工作是针对手工标定区域的描述。

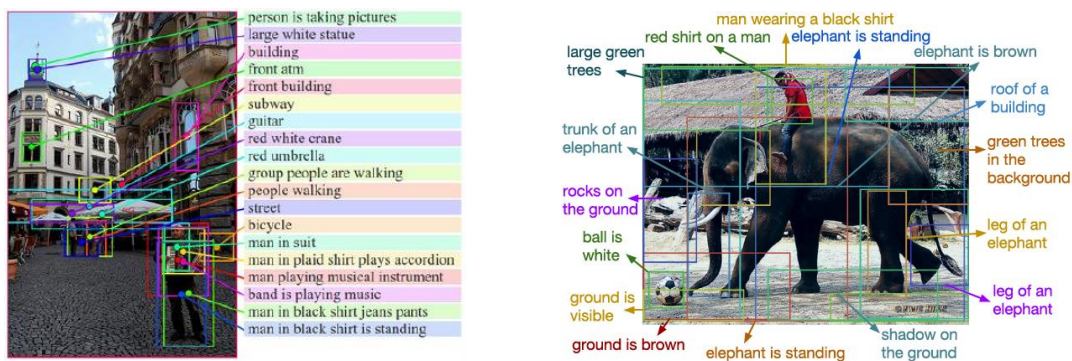


图 3.1 已有工作中局部描述示例

这些局部表达虽然包含了丰富的语义信息,但生成的语句或短句过于简短,全局视角下的局部描述仍然是个研究盲区。本章提出了一种局部描述方法,通过目标检测与全局描述模型对局部目标区域进行语义表达,以填补图像描述领域内局部语义探索的空白,结果示例如图 3.2 所示。

GT: A child holding a flowered umbrella and petting a yak.
 F: (33.33) a group of people riding on the back of a brown horse.



R: (50.00) a group of cows standing in a field.

R: (63.64) a man and a woman standing next to a large elephant

GT: A woman eating vegetables in front of a stove.
 F: (62.50) a woman is holding a plate of food



R: (33.33) a white plate with a piece of cake on top of it

R: (62.50) a woman is eating a piece of pizza

图 3.2 本章工作局部描述示例

图 3.2 中 GT (ground truth)代表真实标注语句, F (Full description)代表全局图像描述结果, R (Region) 代表对应红色目标边框内图像的局部描述, 括号内数字表明 BLEU 1(1-gram)语句评分。可以看出, 本章的局部描述结果比密集描述 (图 3.1)更为完整, 更多地体现了全局视角下的局部细节表达。和全局描述

相比，局部描述可以提供更细节的图像内容信息表示，其中一些描述还包含数据集的真实标注语句中没有出现过的正确表达。本章对局部和全局描述关系的挖掘分析体现了本章方法的优越性，也为进一步的工作打下了良好的基础。

3.2 局部描述

本章提出了一种局部描述方法来表达丰富的图像内容，如图 3.3 所示。

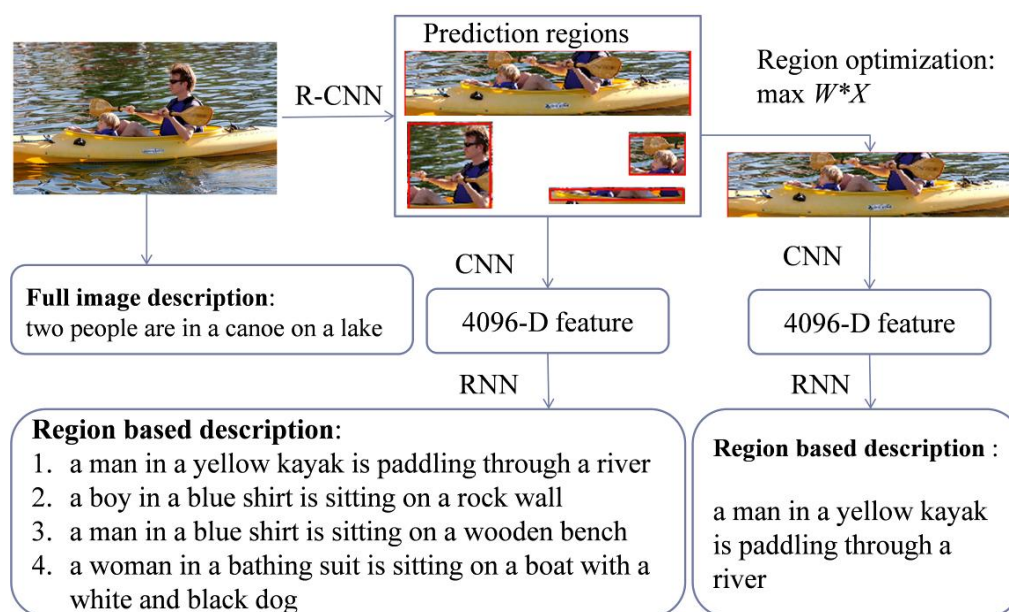


图 3.3 基于区域的局部图像描述模型

首先，采用目标检测方法 R-CNN^[14]来生成目标的图像区域，并用预训练的 CNN 模型 VGGNet^[4]来提取每个区域的视觉特征；然后，在全局图像及其真实标注语句上训练 RNN^[93]语言模型；最后，利用描述模型对目标区域进行局部描述生成。另外，本章还提出了一种区域优化方法来选择一个最佳目标区域进行语句生成，用于和全局图像描述对比。

3.2.1 目标检测

目标检测是计算机视觉领域用于图像理解的经典任务。基于 R-CNN^[14]的方法实现了目标检测历史上里程碑式的突破，极大地提升了检测性能。它使用选择性搜索(Selective Search)^[124]来选择候选目标窗口，然后对 2000 个左右的边框

区域提取 CNN 特征进行目标类别预测，最后采用非极大值抑制方法进行目标合并。本章采用 R-CNN 进行目标区域生成。

在本章中，使用在包含 200 种目标类别的 ILSVRC2014 目标检测挑战数据集上训练获取的 R-CNN 模型进行目标检测并生成图像区域。由于包含目标的区域代表着图像中的突出和重要部分，因此，基于这些区域的图像描述是有意且具有足够信息量的。

3.2.2 区域优化

虽然 R-CNN 是较为优秀的目标检测方法，但它不能保证基于此的局部描述是完全可靠的。一方面，检测结果可能包含一些错误目标；另一方面，即使得到正确的检测目标，如果目标区域过小或者长宽比较大都可能对最终图像描述产生不理想的效果。

为了减少无效检测区域的数量，并探索局部最优描述对全局描述的意义，定义 $X = \{x_1, x_2, x_3, x_4\}$ 为区域选择的四个参考标准：其中 x_1 代表区域面积大小， x_2 代表坐标中心位置， x_3 代表目标类别概率， x_4 代表描述得分。

选择这些指标作为区域选择标准的原因是：在实验中观察到，尺寸较大且更接近图像中心的目标区域在图像描述中表现良好。这一点实际上也是显而易见的，显著的、靠近图像中心的、面积较大的目标区域往往是图像内容的主要构成部分，可以代表整幅图像。另外，人、动物、水果、家具等目标概念在当前自然数据集中也更具代表性，当预测区域目标类别是这些概念时也增加区域的重要性。

定义 $W = \{w_1, w_2, w_3, w_4\}$ 为优化参数，则区域优化的目标函数可表述为：

$$Y = \max(W * X) \quad (3.1)$$

其中 Y 代表局部描述的评测分数。训练获取参数值 W 后，测试时可根据目标区域的四个参考标准进行最优选择。

3.2.3 局部描述

由于领域内公共数据集的真实标注语句都是全局描述，因此局部描述需要

通过全局描述语言模型获取。

1) 全局描述模型训练

首先用整幅图像的 CNN 特征和标注的描述语句进行 RNN 语言模型训练。相关公式和流程与第 2.3 节介绍相同，其中对于 RNN 语言模型，本章采用其变体 LSTM。

2) 局部描述生成

给定目标区域 R ，经过 CNN 特征提取，获得局部特征向量 v ，局部描述生成过程如下：

$$\begin{aligned}
 x_0 &= W_{ix}v, t=0 \\
 x_1 &= W_{ex}s_0, t=1 \\
 h_t &= LSTM(h_{t-1}, x_t), t \geq 1 \\
 y_t, p_t &\propto \exp(W_{hp}h_t), t \geq 1 \\
 x_{t+1} &= W_{ex}y_t, t \geq 1
 \end{aligned} \tag{3.2}$$

和本文 2.3 节介绍相同， y_t 为模型输出，当输出结束词“0”时停止预测。

给定目标检测预测的多个目标区域特征 v ，语言模型将会生成多个语句，其包含丰富的图像内容，如图 3.2 所示，示例图像对四个目标区域产生了 4 个局部描述语句。提出的区域优化方法对每幅图像选出一个最优区域进行语义描述，如图 3.3 所示，区域优化选出了最优区域及其语义描述。结果分析将在实验中详细阐述。

3.3 实验验证与模型分析

本章在两个数据集 Flickr8k^[129]和 MS COCO^[130]上评估所提出的方法。Flickr8k 采用公共划分方法^[49]进行训练(6000 幅)、验证(1000 幅)和测试(1000 幅)。对于 MS COCO 数据，本章从其大规模训练数据中随机选取 10000 幅图像进行训练(8000 幅)、验证(1000 幅)和测试(1000 幅)。两个数据集的数据量近似，主要用于观察本章方法在不同数据集上的表现。

本章采用图像描述领域中最常用的度量标准 BLEU^[140]进行模型评估，采用

4 个评估指标，并用 $B@n$ ($n=\{1,2,3,4\}$)代表 n -gram BLEU，相关详细介绍见第 2 章评价标准部分。

3.3.1 模型参数设置

本章采用经典的基于 CNN-RNN 的描述方法^[49, 52]，并参考其参数设置。CNN 网络使用 16 层卷积神经网络模型 VGGNet^[4]，RNN 模型采用 LSTM，其输入层、隐藏层和输出层的相关参数维度设置为 256。模型的随机梯度下降(SGD)方法使用自适应学习率算法 RMSProp^[144]。学习率开始为 $1e-3$ (Flickr8k)或 $4e-4$ (MS COCO)，并在训练数据每循环 10 次后降低一半。参数 beam_size 根据经验值设置为 5。

语句特征编码时需要创建词汇表，按惯例丢弃训练语句中出现少于五次的单词，因为过少的生僻词汇没有训练的意义。为使用批量数据进行模型训练以提高训练速度及性能，描述的最大长度设置为 16，长度超过此值的将被剪裁。

3.3.2 实验结果

表 3.1 为 Flickr8k 和 MS COCO 数据集上的评测结果。

表 3.1 BLEU 评测结果(%)

数据集	Flickr8k				MS COCO			
	B@1	B@2	B@3	B@4	B@1	B@2	B@3	B@4
FD	51.99	30.71	13.60	5.97	57.27	36.19	17.60	7.88
RD-mean	46.68	25.32	10.23	4.07	50.62	27.49	10.75	4.62
RD-max	52.59	32.20	15.08	6.27	59.30	38.54	18.82	8.98
FD+RD	62.21	42.76	23.93	11.31	64.00	44.57	24.88	12.55
RD-o	48.08	27.12	11.37	4.60	53.90	31.62	13.58	6.01
FD+RD-o	60.36	40.58	21.68	10.68	61.36	41.35	21.48	10.32

本章中目标局部描述和全局描述采用同一个 CNN-RNN 语言模型，因此表 3.1 中每行结果代表同样模型下的不同评测。由于真实标注语句是整幅图像的描述，而且对于局部区域而言没有描述标注，因此所有项目的评价都是基于整幅

图像的标注描述语句。对于目标区域来说，根据整幅图像的标注描述来评估局部语义描述是不公平的，如果训练集有局部的真实标注描述，则局部语义描述将会更加准确。用整幅图像的真实标注语句来评估局部描述有助于探索挖掘图像区域与完整图像之间的语义关系。对于每幅图像，根据区域提取边框可以生成多个局部描述，评测结果就产生多个分数，由于结果对比时需要以每幅图像一个得分的形式，因此评测中有取均值、最大值的操作。

表 3.1 中 FD(Full Description)是基于整幅图像的 CNN 特征生成的全局描述结果。RD(Region Description)代表基于图像区域的 CNN 特征生成的局部描述，RD-mean 表示局部描述的评估分数的平均值，并且 RD-max 是在选择最佳局部描述时的理论最大值。FD + RD 表示全局图像描述和局部描述的组合最优值。局部描述最大值 RD-max 的评测结果高于全局图像描述 FD，这意味着在同样的预测模型下，基于目标检测的局部描述是有意义的，局部描述可以优于全局描述，并且包含更多更详细的语义信息。FD+RD 项目达到最高分数，表明全局和局部描述可以互为补充。

为了显示本章提出的区域优化方法的有效性，表 3.1 中也展示了对优化结果的评测。其中 RD-o 是采用区域优化方法选择一个区域评测的结果。FD+RD-o 表示全局图像描述与区域优化的组合最优。可以看出，区域优化方法结果 RD-o 介于 RD-mean 和 FD 之间，这表明区域优化减少了区域噪声，达到了筛选区域的效果，全局图像可以用有代表性的区域来表达。值得注意的是，对于一些没有重要目标的场景图像，基于区域的描述可能会产生极差的效果。FD+RD-o 实现了最佳性能。这个结果也是显然的，它同样表示了全局描述和局部描述能在一定程度上起到互补作用。

3.3.3 定性分析

探索局部描述的意义和全局描述之间的关系是有重要意义的。因为图像描述是以目标为中心的语义表达，关注目标本身，对目标进行检测并对目标区域进行描述以观察其细节信息有利于挖掘深层次的语义特征。

本章方法能够为每幅图像生成几个局部描述语句，这些语句可能与真实标注完全不同，如图 3.2 和图 3.4 所示。图 3.4 中 GT、F 等符号和图 3.2 相同。

从图 3.2 中可以观察到，这些局部描述是基于目标相互独立的，其中不同区域对应于完全不同的描述。一些局部描述可以完全捕捉整幅图像内容，例如图 3.2 中第二行图像中的“a woman is eating a piece of pizza”，这是因为其对应的目标边框划定了全局图像中最有代表性的范围。一些局部描述中包含更多的细节信息，比如第一行图像中的“a group of cows standing in a field”和第二张图像中的“a white plate with a piece of cake on top of it”，这两个局部描述都是真实标注语句中不存在的正确描述。当然，局部描述可能会得到错误的表达，如第一行图像中的“a man and a woman standing next to a large elephant”对主要目标的判断都出现了错误，这些错误在局部描述中也是不可避免的。

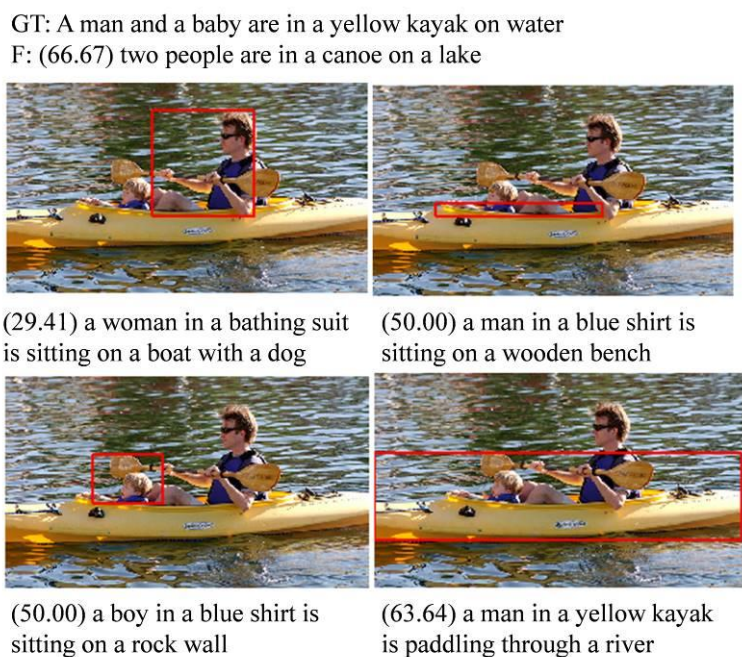


图 3.4 局部描述示例

图 3.4 中展示的局部描述结果没有全局描述结果准确，体现了局部描述的不可靠的一面，但不可否认的是，局部描述语句中含有了全局描述中不存在的语义信息，如“man”、“boy”、“blue”、“yellow”、“river”等，这些细节信息对于以后进一步的语义学习和挖掘是有意义并且有帮助的。

3.4 本章小结

本章提出了一种为图像目标区域生成丰富描述的方法，采用三个深度神经网络模型来生成目标区域(R-CNN)、提取特征(VGG)并生成描述性语句(RNN)，在 Flickr8k 和 MS COCO 数据集上验证了该方法的有效性。基于区域的图像描述可以在很大程度上充分表现整幅图像的内容，所提出的区域优化方法可以为图像选择合适的区域并且实现与全局图像描述相当的描述效果。实验还证明了全局图像描述与提出方法生成的目标局部描述之间是可以互补的，而局部描述语句中还包含真实标注中不存在的有效信息，这些发现对于未来进行深入局部语义探索和学习是非常有益的。

第 4 章 基于局部语义特征嵌入的图像描述

传统的基于 CNN-RNN 的图像描述模型忽略了全局视觉特征的局限性以及视觉和语言空间之间的模态差距。为了解决这些问题，本章提出了一种新的局部语义特征表达方法以增强视觉表达，并通过语义元素嵌入模型(Element Embedding LSTM, EE-LSTM)来进行图像描述语言生成。首先，使用目标检测方法来预测图像中目标所在的区域，并用基准图像描述模型来为这些区域生成局部描述；然后，利用这些描述语句的语义信息和检测到的目标类别生成语义特征，该特征从一定程度上弥补了视觉图像和语义描述之间的模态差距；最后，将 CNN 特征与语义特征集成到所提出的 EE-LSTM 模型中以预测最终的语言描述。在不同数据集上的实验表明，本文所提出的方法优于传统的描述方法，并且具备可扩展性，可以灵活地与基准模型结合以实现更优越的性能。

4.1 引言

自动将图像转换为有意义的语言描述不仅需要全面的图像理解，还需要复杂的自然语言生成。从视觉任务的角度看，图像描述应该能够准确表达图像中涵盖的语义信息，不仅需要捕捉图像中的显著性目标、非显著目标，还必须获取目标的属性、行为、关系等，最后这些信息还要以人类语言的形式呈现出来。如图 4.1 所示，该图像的描述包括显著性目标(dog)、非显著目标(chair)、属性(brown、white)和行为(standing)，最终的期望语句为“two small brown and white dogs standing beside some white chairs”。从实现方法的角度来看，近年来图像描述领域的前沿研究^[49-52, 60]大多都遵循基于 CNN-RNN 的方法来进行图像内容语义生成，如图 4.1 中黑色箭头连接部分所示，首先采用 CNN 进行视觉特征提取，然后用 RNN 语言模型生成描述语句，其中 CNN 和 RNN 可以采用不同的框架结构。Kapathy 和 Li^[49]将 VGGNet^[41]和 RNN 结合起来用于语言模型训练。Vinyals 等^[52]用 GoogLeNet^[55]进行特征提取，并使用 RNN 的变体 LSTM^[56]来进行语言生成。由于 CNN 强大的视觉表达能力和神经网络的端到端机制，基于 CNN-RNN 的方法极大地推动了图像描述领域的发展。

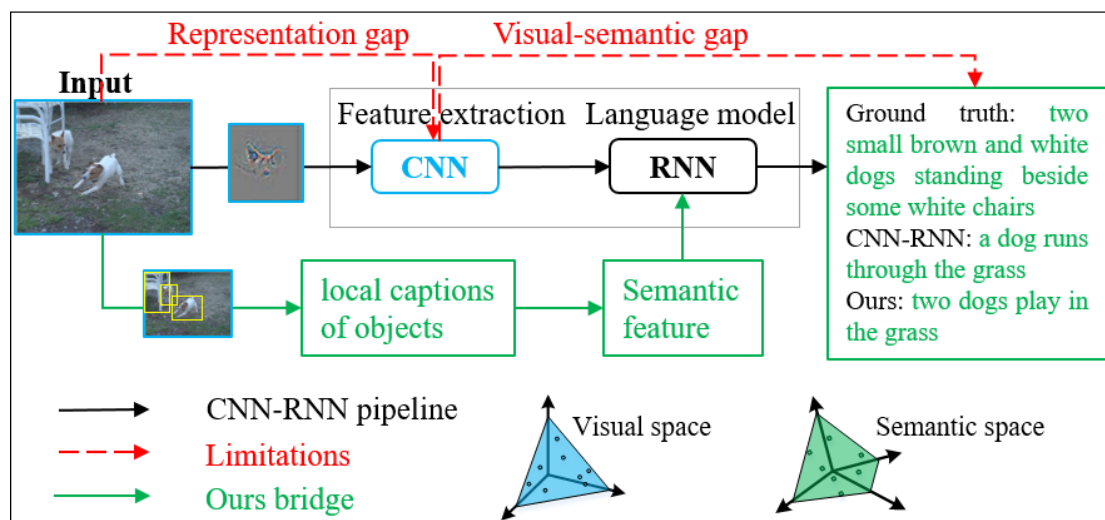


图 4.1 基于 CNN-RNN 的方法存在的问题及本章工作示意图

基于 CNN-RNN 的方法的不足之处是其仅采用全局的 CNN 特征进行语言模型训练，而没有考虑全局视觉特征的限制以及视觉和语言空间之间的模态差异（如图 4.1 中红色虚线箭头所示）。CNN 特征已被证明可以用于各种视觉任务，但预训练的 CNN 模型（通常是在图像分类数据集 ImageNet^[84]上训练）倾向于表达图像中最显著的目标，例如图 4.1 中的“dog”，而忽略了其他细节信息。此外，图像的视觉空间与复杂语句的语义空间之间的模态差异造成的语义鸿沟^[74-76]也加剧了 RNN 语言模型训练的难度。如图 4.1 所示，基于 CNN-RNN 的方法生成语句“a dog runs through the grass”无法描述图像中的所有重要语义信息，如非显著目标单词“chairs”，和非目标单词“two”、“brown”、“white”、“standing”等。

很多研究者致力于采用视觉注意力^[36, 57, 59-66]来指导语言模型的学习。具体而言，Cho 等^[60]和 Xu 等^[64]通过结合视觉注意机制使得模型“看到”关键位置来改善语言模型。Chen 等^[59]扩展视觉注意力模型并引入空间和通道注意力，以更好地理解视觉注意力是什么和在哪里。Lu 等^[63]提出了一种带有视觉哨兵的自适应注意力模型，以决定何时依赖视觉信号以及何时依赖语言模型。视觉注意力将具备“焦点”的局部视觉信息引入到语言模型中，实现了模型性能的提升，对图像描述领域的发展起了极大的推动作用。但是，这些方法忽略了图像中的语义信息。

Jia 等^[62]将从图像检索结果中提取的语义信息作为额外的输入来对 RNN 语

言模型进行指导。Wu 等^[36]将一个属性预测层引入 CNN-RNN 描述框架，用高层语义概念信息作为图像表达以提升描述性能。Yao 等^[65]介绍了带有属性 (Attribute) 的 LSTM 体系结构 (LSTM-A)，将高层属性信息引入到 CNN-RNN 框架。Neo 等^[66]采用语义注意力模型将视觉特征与生成图像描述的循环神经网络中的视觉概念相结合。Gan 等^[61]提出了一种语义组合网络 (Semantic Compositional Networks, SCN) 来有效地将标签文本的单个语义组合成图像描述。这些方法有效地利用了图像的语义信息，但是仅考虑高层次概念而忽略了局部视觉信息。

与已有的工作不同，本章提出一种新的获取局部语义信息的方法，并通过语义元素嵌入模型 (Element Embedding LSTM, EE-LSTM) 来提升语言生成性能。

图像描述是以目标为中心的、共享一个简单而独特的结构的语义表达。如 Le Bret 等^[145]所述，场景中的关键元素用名词短语 (目标) 来描述，介词和动词短语主要用于描述目标属性或目标之间的交互信息。在本章中，所有这些描述涉及的词汇，例如形容词、量词和动词都被称为语义元素。对于语义元素的获取，根据上一章工作中局部描述的结果进行单词打散和筛选，这些单词被用于语义元素特征生成。大多数语义元素是独立于目标的，即它们可以在局部区域内被完全捕获。从不同局部区域生成的描述更多的真值元素，如图 4.2 所示。

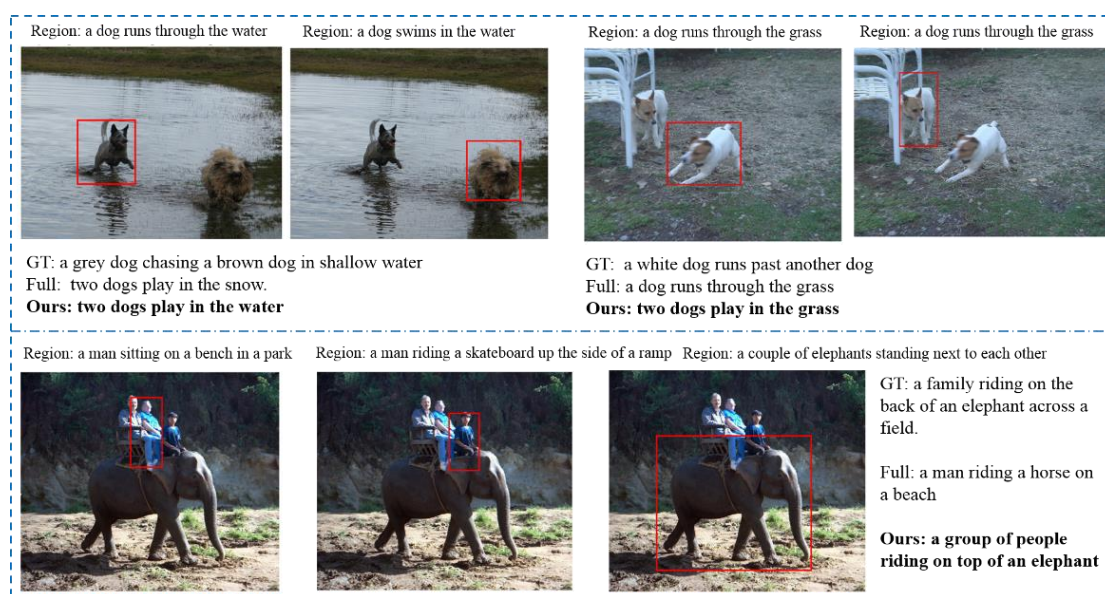


图 4.2 全局和局部描述的示例

其中 GT (ground truth)代表真实标注, Full (Full description)为全局描述, Region 代表对应红色目标边框的局部描述。图 4.2 局部描述中“dog”、“play”、“water”、“elephant”等都为有效语义元素,与图片的真实内容相关。其中第二行图片的局部描述中“elephant”为全局描述中没有的真值元素,其他单词如“of”、“skateboard”、“bench”与图像内容不相关或无意义,为无效语义元素(噪声元素)。

本章工作将 CNN 特征与语义元素特征集成到提出的 EE-LSTM 模型中进行语言模型优化。具体来说,通过引入语义元素发现和嵌入来实现本章提出的图像描述方法。首先,基于图像和语句标注训练一个基本的 CNN-RNN 描述模型,如图 4.1 中黑色箭头所示;然后,对整幅图像利用目标检测方法生成有效区域,这些局部区域包含丰富的语义信息(正如在本文第 3 章和 Zhang 等^[146]所研究和论证的结果所示);接下来,对目标区域进行局部描述获得目标的局部语义信息并生成语义特征,如图 4.1 中绿色箭头所示,该特征包含了目标元素的细节信息,又和图像描述共享同一个语义空间,这些特点使得语义特征增强了图像表达能力;最后,将 CNN 特征和语义特征嵌入到 EE-LSTM 模型中进行语言模型训练和预测。EE-LSTM 比传统的 LSTM 多了额外的语义特征输入,即采用包含视觉和语义、全局和局部信息的图像表达进行描述生成,弥补了视觉图像和语义描述之间的差距,实现了描述性能的提升,定量评估证明了其有效性。

4.2 语义元素嵌入模型

语义元素嵌入模型包含两个部分:元素信息挖掘(图 4.3 的上半部分)和元素信息嵌入(图 4.3 的底部)。第一部分是使用 CNN 和 LSTM 训练基准图像描述模型,用于预测整幅图像和所有目标区域的图像描述。这些区域由目标检测方法生成,在本章采用 Faster R-CNN^[118]进行目标检测。在获得多个局部描述和全局描述结果后,将所有描述分割成语义单词,称之为元素(Element)。根据目标检测预测标签对元素集合进行加权,以滤除不相关的信息进而加强目标信息。第二部分将元素信息生成 Bag of Words (BoW)特征,该语义特征和 CNN 特征一起输入到提出的 EE-LSTM 模型中训练并得到最终的描述模型。

- a) 置信度：候选目标的检测分数必须大于 a ;
- b) 交并比：候选目标区域和真实目标区域之间的交并比(Intersection over Union, IoU) 不能大于 b ;
- c) 大小：候选目标边框的宽度和高度分别不小于 n 个像素。

a 、 b 和 n 的值在验证集上通过实验确定。利用 ImageNet 数据集上预训练的 16 层 VGGNet^[4]来提取整幅图像以及有效目标区域的视觉特征。

3) 局部和全局图像描述

基于之前获取的图像描述模型和目标区域，局部和全局描述生成过程为：

$$\begin{aligned}
 x_0 &= W_{ix}v, t=0 \\
 x_1 &= W_{ex}s_0, t=1 \\
 h_t &= LSTM(h_{t-1}, x_t), t \geq 1 \\
 y_t, p_t &\propto \exp(W_{hp}h_t), t \geq 1 \\
 x_{t+1} &= W_{ex}y_t, t \geq 1
 \end{aligned} \tag{4.1}$$

其中 v 为局部或全局的 CNN 特征， y_t 为模型输出，当输出结束词“0”时停止预测。相关参数意义都和第 3 章 3.2 节相同。

对每幅图像，模型能够生成全局描述和多个局部描述，与本文第 3 章的工作相似，区别在于方法细节有所不同。本章局部和全局描述性能有所提升，目的是为后面的进一步工作做铺垫。

4.2.2 语义特征嵌入

基于上一小节的全局和局部描述信息，采取元素特征生成和语义特征嵌入的方法进行图像语义表达和语言模型训练。

1) 元素特征生成

预测的局部描述和全局描述包含丰富的信息。本章的目标不仅仅在于探讨局部描述的意义，而且还在于使用局部语义信息来改进整幅图像的描述。因此，将全局和局部描述信息，即生成的多个语句，进行单词分割打散并获得目标语义元素。这个元素集合不仅包含了目标，还有动作、属性及关系等所有可以组成语句的单元，如“dog”、“white”、“swimming”、“a”、“is”等。

基于元素集合，采用 BoW 对元素集合 elements 进行统计直方图生成，进而

生成元素特征 e 。该特征的维度与语句的单词词汇量的大小相同，即该特征和描述语句共享一个语义空间。元素特征 e 的公式为：

$$e = \text{BoW}(\text{elements}) \quad (4.2)$$

其中 $e \in \mathbb{R}^{n \times 1}$ ， n 为元素集合 elements 的语义空间单词个数。

由于局部描述生成的多个语句并不是完全准确，导致元素集合中的单词并不都与图像内容相关，比如图像中不存在的语义元素，即噪声元素，会对描述模型训练产生严重的干扰。为了减少噪声元素的影响，利用目标检测中获得的目标类别，对已获取的语义元素进行数据预处理，通过加强目标元素权重来进行噪声抑制。

由于元素词汇表和目标类别在单词概念上存在标注差异，如目标类别中代表人的类别“person”，在元素中为“man, woman”等，因此需要将目标类别 C 近似地映射到最相似的元素，例如将“person”映射为“man”，“vehicle”映射为“car”。对于元素和目标类别之间的相似性，可以通过 Gensim^[88]的词向量工具进行计算。假设预测的目标标签包含 k 个类别 $C = \{c_1, c_2, \dots, c_k\}$ ，语义元素集合包含 n 个单词 $W = \{w_1, w_2, \dots, w_n\}$ ，元素特征预处理函数为：

$$e(w) = 1, \text{ if } \text{similarity}(c, w) > th \quad (4.3)$$

其中 $c \in C$ ， $w \in W$ 。该公式定义与目标类别相似的单词 w 将被赋予权重“1”。阈值 th 在实验中根据经验确定。

预处理前后的语义元素权重变化对比如图 4.4 所示，其中横坐标中的单词是通过元素发现过程获得的语义元素。通过数据预处理，与目标检测类别“person, bench, dog”相似的元素中的单词“person”、“man”、“bench”、“dog”都被赋予了更大的权重“1”。其他有效的元素，如“a”、“sitting”、“on”、“road”等都是元素的一部分。像“motorcycle”和“water”这样的噪声元素仍然存在，但权重较轻，对整个元素特征贡献较少。值得注意的是，在整个单词映射空间中，目标类别“person”可以对应一组单词，包括“person, man, woman, child, boy, girl”等。在图 4.4 中所示例子中，目标类别“person”仅映射到对应图像元素

空间的“person, man”，降低了特征映射的模糊性。通过元素预处理，元素特征更好地表达了图像中的语义信息。

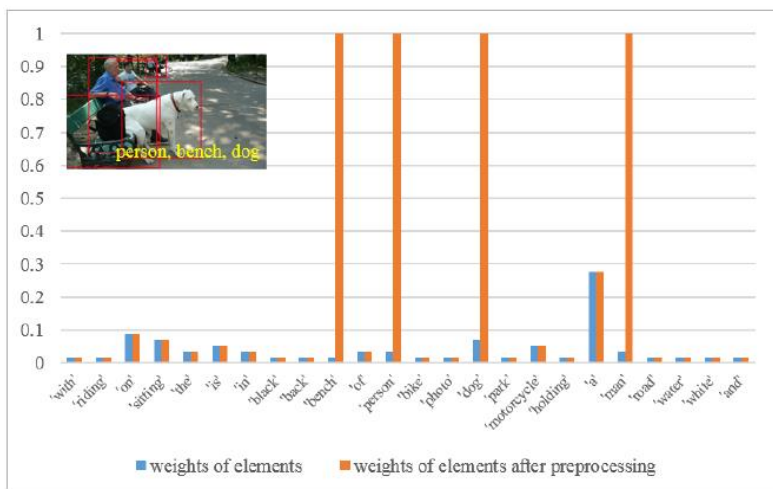


图 4.4 元素预处理的变化对比

2) 元素信息嵌入

为了发挥局部语义特征的作用，本章提出 EE-LSTM 模型进行语义特征嵌入和图像描述生成。EE-LSTM 的输入包含局部和全局、视觉和语义信息，其内部结构如图 4.5 所示：

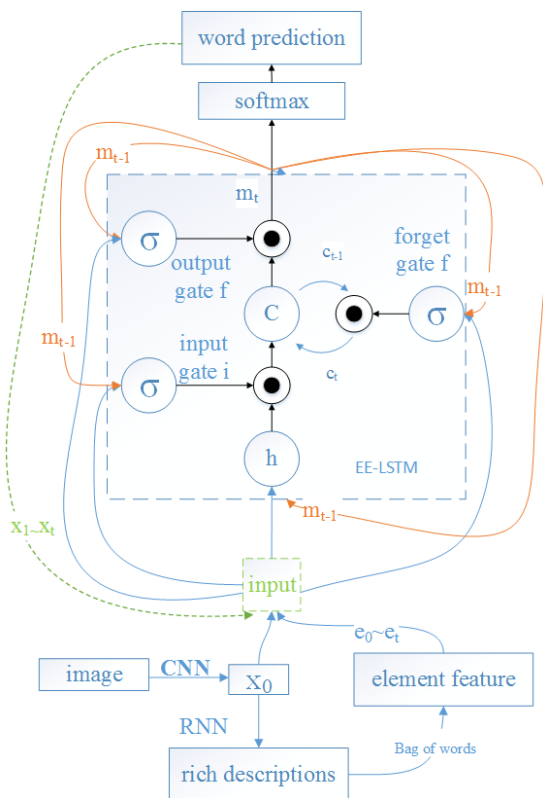


图 4.5 EE-LSTM 模型

其公式表达为:

$$\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ie}e) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fe}e) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oe}e) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1} + W_{ce}e) \\
m_t &= o_t \odot c_t \\
p_{t+1} &= \exp(m_t)
\end{aligned} \tag{4.4}$$

其中 x_0 代表 CNN 特征, 包含全局视觉信息, e 代表元素特征, 包含局部语义信息, 其特征维度和 LSTM 特征空间一致, 并参与每次循环。其他参数都和第 2 章中 LSTM 模型介绍含义相同。模型损失函数和传统 LSTM 一致, 为每次循环的负对数似然和:

$$L(x, e) = -\sum \log p(s_t | x_t, e; \theta) \tag{4.5}$$

在元素特征有表达能力的前提下, 损失函数 $L(x, e)$ 将低于基准语言模型的损失函数 L , 本章第 4.3 节将会进行验证。

4.2.3 对比模型 VE-LSTM

为了进行模型对比, 提出了一种视觉特征嵌入模型 VE-LSTM (Visual Embedding LSTM)。在 VE-LSTM 中, 首先提取所有目标区域的特征表示, 然后通过计算每个维度的最大值的方式来合并区域特征, 最后将新的局部视觉特征用和 EE-LSTM 中语义元素特征同样的方法嵌入到 LSTM 中。理想情况下, 合并后的区域特征将包含局部细节信息。在实验中仅以此与局部语义特征嵌入的方法进行对比, 以显示局部语义元素的有效性。

4.3 实验验证及结果分析

本节首先展示局部描述的有效性, 然后评估本章新提出的方法的描述性能。所提出的方法使用 Torch7^[147] 实现, 并且在具有 i7 3.2GHz CPU, 32GB RAM 和 K40 GPU 的服务器上进行测试。

本章在三个公共数据集上评估本章提出的模型: Flickr8k^[129], Flickr30k^[133]

和 MS COCO^[130]。与本文第 2 章 2.5 节数据集部分的介绍相同，采用的数据量及训练、验证、测试集遵循公共的划分方法^[49]。三种数据集都单独训练测试，没有数据交叉或迁移。

采用自动评价标准进行模型评估，评价标准的详细描述见第 2 章 2.6 节。使用 B@n, M, C, R 来分别代表 n-gram BLEU, METEOR, CIDEr, ROUGE_L。

4.3.1 模型参数设置

基准模型 LSTM 来自于^[148]，该模型在领域图像描述占据重要地位。其输入是整幅图像的 CNN 特征，输出是相应的语句。输入的 CNN 特征是在 ImageNet 数据集上预训练的 16 层 VGGNet^[4]网络中提取的。对于每幅图像，描述模型第一次迭代的输入是 CNN 特征，而其他迭代的输入是标注语句的单词。

模型相关参数根据现有工作^[49, 52]及经验值进行设置。LSTM 模型内部特征维度设置为统一值 512，包括单词编码大小、CNN 特征编码大小、RNN 各层维度（输入层，隐藏层和输出层）、以及元素特征编码大小。学习速率开始为 $4e-4$ (MS COCO)或 $1e-3$ (Flickr8k, Flickr30k)，所有训练数据每迭代 10 次后降低一半，总迭代次数为训练数据集图像个数的 50 倍。神经网络训练时的随机梯度下降(SGD)方法使用自适应学习率算法 RMSProp^[144]。参数 beam_size 为每个单词预测时的概率搜索空间。理论上，更高的 beam_size 会带来更好的性能，但同时会导致测试速度下降，一般将此值设置为 7^[49]、10^[62]或 20^[52]，本章根据经验值将其设置为 5。

在创建词汇表时，丢弃训练语句中出现少于五次的单词。词汇表的最终大小为 Flickr8k 为 2622，Flickr30k 为 5793，COCO 为 9565。另外将描述的最大长度设置为 16，长度超过此值的将被剪裁。

除了引入额外的元素特征作为输入外，EE-LSTM 语言模型训练大多数设置与 LSTM 模型相同。元素特征编码的维度也为 512。目标区域提取和元素特征生成中的相关阈值根据经验值固定，其中 $a = 0.6$ ， $b = 0.5$ ， $n = 50$ ， $th = 0.5$ 。经过阈值控制，每幅图像平均可产生 5 个目标区域。

作为对比模型，VE-LSTM 参数设置都与 EE-LSTM 相同。

4.3.2 实验结果

本节评估局部描述和元素嵌入模型的实验结果。

表 4.1 展示了 Flickr8k, Flickr30k 和 MS COCO 数据集上的结果。前四行代表了基于相同的基准 LSTM 模型进行语句预测的结果, 该模型基于整幅图像的全局 CNN 特征(VGGNet)和真实标定语句进行训练。FD (Full Description)表示基于整幅图像的 CNN 特征生成的全局图像描述, RD (Region Description)表示基于图像区域的 CNN 特征生成的局部描述。

表 4.1 全局和局部描述结果

数据集	Flickr8k				Flickr30k				COCO			
	B@1	B@2	B@3	B@4	B@1	B@2	B@3	B@4	B@1	B@2	B@3	B@4
LSTM (FD)	55.7	37.2	23.7	15.3	57.0	37.7	24.6	16.0	63.9	45.0	30.4	20.5
LSTM (RD-mean)	50.7	30.2	18.7	11.6	50.0	30.7	19.6	11.1	56.9	38.0	23.4	16.5
LSTM (RD-max)	58.7	39.2	26.7	18.3	60.0	40.7	27.6	17.6	63.9	45.0	30.4	18.5
LSTM (FD+RD)	63.8	42.8	33.9	19.5	64.0	43.2	30.9	20.5	69.1	51.1	36.6	25.3
VE-LSTM	51.2	30.5	18.8	11.9	50.6	31.2	20.4	11.7	57.6	38.9	24.3	17.2
EE-LSTM+GT	87.4	74.0	59.2	46.6	90.1	72.0	55.1	41.4	93.5	80.5	66.5	53.8
EE-LSTM	59.8	40.8	27.5	18.4	59.2	39.1	25.7	17.0	67.5	49.8	36.4	26.9

由于本章采用的数据集的真实标注语句是整幅图像的描述, 没有关于区域的描述标注, 因此所有模型的评价都是基于整幅图像的标注描述。对于局部描述来说, 根据整幅图像的标注来评估是有偏差的, 但有助于探索挖掘图像区域与完整图像之间的语义关系。对于每幅图像, 根据区域提取边框可以生成多个局部描述, 而评测时需要选择一个值进行对比, 因此表 4.1 中模型评测有取均值(mean)、最大值(max)的操作。

在表 4.1 中, RD-mean 表示局部描述的多个评估分数的平均值, RD-max 是选择最佳局部描述的理论最大值, FD + RD 表示全局图像描述和局部描述的组合最优值。局部描述最大值 RD-max 的评测结果高于全局图像描述 FD 意味着在同样的预测模型下, 局部描述可以优于全局描述, 甚至包含更详细的语义信息。FD + RD 项目达到最高分数, 表明全局和局部描述可以互补。这部分实

验与第 3 章中的结果相似。实际上, 本章局部描述部分基本与第 3 章中相同, 区别在于本章采用了更加优越的 **Faster R-CNN** 进行目标检测、更多训练数据、更好的模型参数设置。这也体现了本章所提出的方法扩展性较强, 还有很大的提升空间。

全局和局部的描述包含更多可以反映图像内容的有效信息, 可以从该信息中提取语义特征, 从而弥补 CNN 特征的局限性, 并得到更好的图像表达性能。表 4.1 中的 EE-LSTM 使用基准 LSTM 模型生成的元素特征进行模型优化, 最终的性能优于 LSTM (FD), 证明了本章方法的有效性。

局部视觉特征嵌入(VE-LSTM)的结果对比原始 LSTM 模型并无优势, 这意味着将局部视觉特征合并在一起并不能提升语言模型, 因为合并视觉向量时, 可能会造成无序的特征值的混乱。然而, EE-LSTM 采用的是语义特征嵌入的方法, 这种合并对于语言模型来说是有帮助的。

根据本章第 4.2 节的分析, 元素中的很多单词反映了图像内容的一部分, 而由于局部描述不完全准确, 导致语义元素集合中含有很多噪声。局部描述的均值(RD-mean)并不像全局图像描述(FD)那样高; EE-LSTM 模型要高于基准模型 LSTM (FD), 但低于理论最大值 $FD+RD$, 体现了噪声元素的干扰。EE-LSTM+GT 采用真值标定进行语义特征生成和嵌入, 其语义元素都是有效的, 实验性能也极具优势, 进一步表明了所提出的 EE-LSTM 有很大的提升空间, 比如可靠的目标检测器和更好的元素降噪方法都会有助于语言模型句子预测。这些问题促使了本章利用元素预处理方法来增强元素特征, 其中检测得到的目标类别被映射到元素单词空间, 然后与目标类别相似的单词被标记为强元素。元素和目标类别之间的相似性通过 Gensim^[88]的词向量工具进行计算, 相似度高于 0.5 的单词将被赋予更高的权重并构成新的元素特征, 后面的实验结果证明了元素预处理的有效性。

4.3.3 模型对比

由于 EE-LSTM 模型独立于用于生成语义元素的基准模型, 因此它能够灵活地和不同的基准模型进行结合。除了元素预处理之外, 用于特征提取的 CNN 模型也是性能提升的另一个因素。本章采用 VGGNet^[4]、ResNet^[85]和经过微调的 VGGNetFT 来提取特征, 其中 VGGNet 和 ResNet 是在 ImageNet 数据集上进

行预训练，而 VGGNetFT 则在 MS COCO 描述数据集^[148]上进行微调。

表 4.2 在 MS COCO 数据集上的模型对比(%)

模型	B@1	B@2	B@3	B@4	M	C	R
NeuralTalk ^[49]	62.5	45.0	32.1	23.0	19.5	--	--
Google NIC ^[52]	66.6	46.1	32.9	24.6	--	--	--
Soft-Attention ^[64]	70.7	49.2	34.4	24.3	23.9	--	--
Hard-Attention ^[64]	71.8	50.4	35.7	25.0	23.0		--
gLSTM ^[62]	63.8	46.3	33.6	24.8	23.3		--
ATT ^[66]	70.9	53.7	40.2	30.4	24.3	--	--
Adaptive ^[63]	74.0	58.0	43.9	33.2	26.6	108.5	--
SCN-LSTM ^[61]	74.1	57.8	44.4	34.1	26.1	104.1	--
LSTM-A ^[65]	73.0	56.5	42.9	32.5	25.1	98.6	53.8
LSTM (VGGNet)	63.9	45.0	30.4	20.5	20.1	63.1	46.6
LSTM (ResNet)	66.5	48.1	33.5	23.3	21.7	74.2	48.6
LSTM (VGGNetFT)	72.1	55.2	40.4	29.1	24.5	92.4	53.0
EE-LSTM+GT (VGGNet)	93.5	80.5	66.5	53.8	32.4	135.9	65.6
EE-LSTM (VGGNet)	67.5	49.8	36.4	26.9	22.6	80.0	49.9
EE-LSTM (ResNet)	69.8	52.2	38.5	28.7	23.9	89.4	51.9
EE-LSTM (VGGNetFT)	73.6	56.9	43.0	32.5	26.0	101.9	54.3
EE-LSTM-P (VGGNet)	70.4	52.8	39.1	29.2	24.1	90.7	51.8
EE-LSTM-P (ResNet)	71.4	54.1	40.3	30.1	24.7	94.3	52.6
EE-LSTM-P (VGGNetFT)	75.7	59.4	45.3	34.6	26.8	109.6	56.0

在 MS COCO 数据集上的实验结果如表 4.2 所示，其中“-”代表缺失数据。和 LSTM 模型相比，EE-LSTM 性能有了显著提升。例如，采用 VGGNet 特征的 EE-LSTM 比 LSTM 在 B@1 上实现了大约 4% 的增益，采用 ResNet 的增益约为 3%，而采用 VGGNetFT 特征的 LSTM 模型本身性能很高，EE-LSTM 仅增加了约 1.5%。添加了元素预处理的模型 EE-LSTM-P 优于 EE-LSTM，在 B@1 中增益约 2%，证明了语义元素降噪的有效性。基于 VGGNetFT 的 EE-LSTM-P

实现了最佳性能,如表 4.2 中最后一行加粗数据所示。与 VGGNet(LSTM)相比,具有更好的视觉表达模型和更丰富的语义元素信息的 VGGNetFT(EE-LSTM-P)在 B@1 中提高了约 12%。

在表 4.2 中,由于 MS COCO 数据集自身包含了 Flickr8k 和 Flickr30k 所不包含的详细目标类别和区域标注,本章仅在 MS COCO 数据集上进行模型对比。目标类别和区域标注的可用性使本章所提出的模型更加有效。

包括 NeuralTalk^[49]和 Google NIC^[52]在内的早期图像描述方法都是基于基准 CNN-RNN 框架所设计。和基准方法 LSTM 基本一致,它们的评估结果低于本章提出的 EE-LSTM,因为这些基准方法只是从全局视觉特征中学习语言模型,往往无法捕捉图像中的细节信息和语义元素。虽然基于视觉注意力的方法 Soft-Attention^[64]和 Adaptive^[63]考虑了局部信息,但只是把注意力转向了视觉空间。尽管它们比传统的 CNN-RNN 方法性能要好,但和 EE-LSTM 模型仍有一定差距。此外,基于语义信息的方法包括 gLSTM^[62], ATT^[66], SCN-LSTM^[61]和 LSTM-A^[65],虽然考虑了高层语义信息,但它们忽视了局部细节信息。总体而言,本章提出的语义元素特征不仅考虑了目标区域,而且与描述语句共享相同的语义空间。所提出的方法将语义元素特征与视觉特征集成在一起,实现了局部与全局、视觉与语义的多角度特征表达,缩小了视觉图像和语义描述之间的差距,并取得了最优的性能。

4.3.4 模型分析

本小节通过比较模型训练的损失误差值来进行模型分析和验证。

图 4.6 模型训练损失误差对比中对比了三种方法:基准 LSTM 模型(蓝色),EE-LSTM 模型(绿色)以及加入元素预处理的 EE-LSTM-P 模型(红色)。本章提出的元素嵌入模型 EE-LSTM 的训练误差比基准模型更低。在语义元素嵌入特征的指导下,EE-LSTM 在所有图像迭代一次时损失误差下降最大,在后面的迭代中也缓慢下降,这表现出视觉和语义特征组合的优越性。

与没有元素预处理的 EE-LSTM 模型相比,EE-LSTM-P 模型训练误差更低,显示了语义元素预处理的优点。更具体地说,在评估中元素预处理的改进大约为 2%,如表 4.2 所示。未来对元素降噪方法的研究将有助于进一步减少 EE-LSTM 的损失误差。

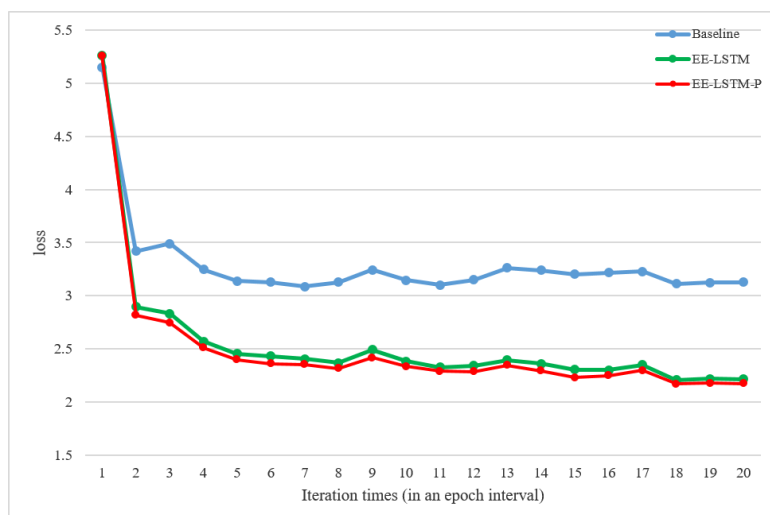


图 4.6 模型训练损失误差对比

4.3.5 定性评估

图 4.7 显示了本章提出的方法生成的一些示例描述。图像中的红色边界框和黄色单词是目标检测结果，其中前两行图例中显示了局部描述结果，最后一行则显示了预处理后的语义特征权重对比图。由于区域各不相同，生成的局部描述也大部分是不相关的，但包含足够多的可以表达图像的语义元素。经过语义元素的预处理，所提出的模型合并了有效的元素信息并利用有效的语义特征产生了更好的描述。

<p>GT: A large white dog is sitting on a bench beside an elderly man.</p> <p>Full: a dog sitting on a bench in a park Regions: a photo of a person holding a dog a man is sitting on a motorcycle with a dog a person is sitting on a bench on a road a dog is sitting on the back of a motorcycle</p> <p>Ours: a dog sitting on a bench next to a man</p>	<p>GT: a man leaning over a lot as another man catches the frisbee</p> <p>Full: a group of people playing soccer on a field Regions: a man is holding a frisbee in his hand a man in a red shirt is playing frisbee a man in a baseball uniform throwing a baseball a man holding a frisbee in his hand</p> <p>Ours: a group of men playing a game of frisbee</p>																																										
<p>GT: A pair of children sit on a giraffe while other children stand nearby.</p> <p>Full: a couple of giraffe standing next to a man Regions: a man is holding a white frisbee in his hand a man is riding a bike in the woods a woman is sitting on a bench with a baby a young boy is feeding a giraffe in a zoo</p> <p>Ours: a group of children are standing on a small giraffe</p>	<p>GT: a cell phone screwdriver a pair of scissors and a black thing on a desk</p> <p>Full: a cup of coffee and a cup of coffee Regions: a cup of coffee sitting on top of a wooden table a cup of coffee and a cell phone on a table a close up of a banana on a table a close up of a cell phone on a table</p> <p>Ours: a pair of scissors sitting next to a cup of coffee</p>																																										
<p>weights of elements</p> <table border="1"> <tr><td>'hand'</td><td>0.0</td></tr> <tr><td>'base'</td><td>0.0</td></tr> <tr><td>'building'</td><td>0.0</td></tr> <tr><td>'floor'</td><td>0.0</td></tr> <tr><td>'wall'</td><td>0.0</td></tr> <tr><td>'sky'</td><td>0.0</td></tr> <tr><td>'back'</td><td>0.0</td></tr> <tr><td>'ground'</td><td>0.0</td></tr> <tr><td>'air'</td><td>0.0</td></tr> <tr><td>'water'</td><td>0.0</td></tr> </table> <p>GT: people on bicycles ride down a busy street Full: a man riding a bike down a street Ours: a group of people riding bikes down a street</p>	'hand'	0.0	'base'	0.0	'building'	0.0	'floor'	0.0	'wall'	0.0	'sky'	0.0	'back'	0.0	'ground'	0.0	'air'	0.0	'water'	0.0	<p>Weights of elements</p> <table border="1"> <tr><td>'hand'</td><td>0.0</td></tr> <tr><td>'base'</td><td>0.0</td></tr> <tr><td>'floor'</td><td>0.0</td></tr> <tr><td>'wall'</td><td>0.0</td></tr> <tr><td>'sky'</td><td>0.0</td></tr> <tr><td>'back'</td><td>0.0</td></tr> <tr><td>'ground'</td><td>0.0</td></tr> <tr><td>'air'</td><td>0.0</td></tr> <tr><td>'water'</td><td>0.0</td></tr> <tr><td>'kite'</td><td>0.9</td></tr> <tr><td>'person'</td><td>0.1</td></tr> </table> <p>GT: A couple of people standing in a field flying a kite Full: a man is flying a kite in a field Ours: a group of people standing in a field flying a kite</p>	'hand'	0.0	'base'	0.0	'floor'	0.0	'wall'	0.0	'sky'	0.0	'back'	0.0	'ground'	0.0	'air'	0.0	'water'	0.0	'kite'	0.9	'person'	0.1
'hand'	0.0																																										
'base'	0.0																																										
'building'	0.0																																										
'floor'	0.0																																										
'wall'	0.0																																										
'sky'	0.0																																										
'back'	0.0																																										
'ground'	0.0																																										
'air'	0.0																																										
'water'	0.0																																										
'hand'	0.0																																										
'base'	0.0																																										
'floor'	0.0																																										
'wall'	0.0																																										
'sky'	0.0																																										
'back'	0.0																																										
'ground'	0.0																																										
'air'	0.0																																										
'water'	0.0																																										
'kite'	0.9																																										
'person'	0.1																																										

图 4.7 EE-LSTM 模型结果示例描述

所提出的方法能够纠正目标之间的关系，即使它们没有在所有局部描述中提及。例如，第一行第一幅图像中基准预测(Full)是“a dog sitting on a bench in a park”，EE-LSTM 的预测结果(Ours)是“a dog sitting on a bench next to a man”，其中“next to a man”并没有出现在局部描述中，EE-LSTM 模型通过语义元素“man”结合图像视觉特征进行了新的模型学习和表达。另外，实验中发现，对于含有多个相似目标的图像，EE-LSTM 对目标量词更容易修正，如底部样本中，用“a group of”取代了 LSTM 的“a”。这些观测意味着所提出的 EE-LSTM 不是简单地从语义元素中选择单词，而是通过影响语言模型对图像视觉特征的理解，进行新的单词或目标交互信息的推断。

图 4.8 显示了本章模型生成的一些反面例子。对于场景简单的图像的描述，比如只有一个目标的情况下（如第一列中的前两个图像），EE-LSTM 不能充分发挥其优势，生成的语句和基准 LSTM 没有太大区别，这是因为局部描述语句单一，含有的信息量也太少。而对于不易分辨或易混淆的图像而言，EE-LSTM 会产生有偏差的描述。如图 4.8 第一列中的第三张图像所示，基准描述结果(Full)为“a giraffe standing in a field with trees in the background”，而 EE-LSTM 模型预测结果是“a close up of a bird on a branch”，两者都是有偏差的描述，这个问题和图像自身的模糊性有关。图 4.8 第二列中的两幅图像的结果中 EE-LSTM 效果更差，原因是复杂的图像内容导致了局部描述的噪声增多，对整体图像描述产生了错误的指导。未来的工作中，可以探讨采用更有效的元素预处理方法来解决这个问题。

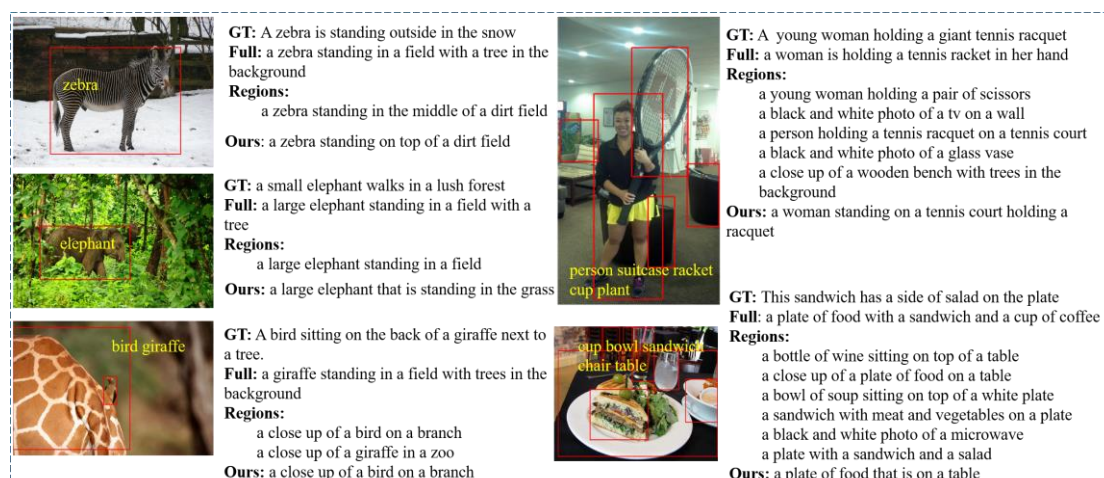


图 4.8 反例样本

4.4 本章小结

本章提出了一种新的图像描述方法，通过引入局部语义特征来强化图像表达和增强语言模型的描述性能。首先，根据本文第3章的工作内容挖掘局部描述与全局描述之间的关系；然后，探索了局部语义信息的意义并生成语义特征，该语义特征包含了图像中更详细的语义信息，并且与描述单词共享同一个语义空间，弥补了视觉图像和语义描述之间的模态差距；最后，通过提出的 EE-LSTM 语言模型进行视觉与语义特征集成，并生成包含更多细节的语句。实验结果表明，本章提出的方法性能优于传统的描述方法，显示了图像理解中局部语义信息的有效性。

第 5 章 基于目标关键词驱动的图片描述

现有的图像描述方法局限于用有限的语义信息来描述图像，通常只能生成一个围绕主要目标的语句来描述整幅图像，图像描述不够充分，也不能满足个性化需求。本章将关键词引入到图像描述模型中，通过关键词的引导生成更深层的描述语句。给定多个关键词，可以在其驱动下为同一图像生成具备不同侧重点的描述，不仅丰富了描述内容，还能够满足不同的用户需求。本章提出了上下文依赖的双边 LSTM (Context-dependent Bilateral Long Short-Term Memory, CDB-LSTM)模型来实现基于目标关键词驱动的图片描述，通过考虑关键词前后单词之间的相关性来实现局部语义学习，达到生成丰富的个性化描述的目的。

5.1 引言

现有的图像描述方法^[14, 37, 43-46, 48, 49, 52, 66]能够为每幅图像生成一个描述语句，却难以解决图像自身的模糊性问题：一幅图像包含了太多信息，很难用一句话来完整地描述整幅图像；对于不同的人来说，关注点不一样，对同一幅图像可能有不同的见解，用一句话来定义一幅图像本身存在语义上的偏差。如图 5.1 所示，右侧语句为五个人对左侧图像的不同描述，第一个人和第三个人只描述了最显著的红色目标，第二个人描述了两个目标，而第四和第五个人则涵盖了所有目标。由此可见图像描述是因人而异的，特别是对于包含多个目标的复杂图片而言，传统的图像描述不能满足语义表达的完整性，因此对图像描述的多样性和个性化探索是有重要意义的。

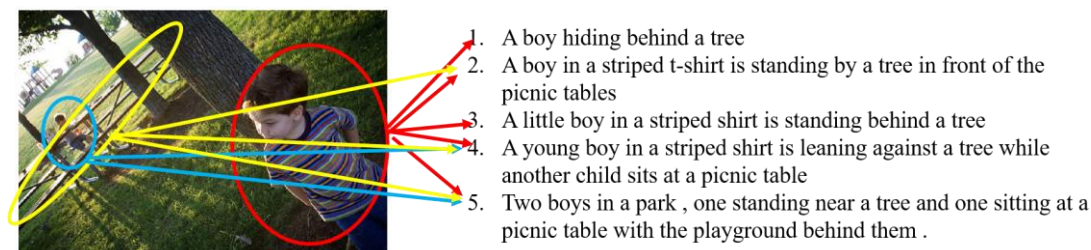


图 5.1 图像描述任务本身现存的问题示例

段落描述^[67]将图像描述的单个语句描述转化为段落描述，从而实现更全面的图像内容表达。但是，这种长段落形式的表达需要重新建立标注更为复杂的数据集，已经不同于传统的图像描述任务。密集描述^[51, 68]通过对图像区域进行细节描述来实现细粒度的图像语义表达。但是，这个工作生成的描述太过简短，通常是一个单词或者几个单词组成的短语，在语义理解方面不够完整。本文第 3 章的工作探索基于目标区域的局部描述，生成了完整的描述语句，对于复杂图像却不能体现全局关系，也过于依赖于局部区域边框的准确度。第 4 章的工作利用局部语义信息进行图像描述增强，却同样存在描述语句单一的问题。本章在现有的包含有限标注的数据集的基础上，将目标类别/概念引入图像描述进行局部语义学习，解决图像描述的模糊性问题。

本章通过引入额外的目标关键词作为先验信息生成个性化的具备不同侧重点的语句，如图 5.2 所示（括号内显示的值是 BLEU 1 评测分数）。由于生成的语句依赖于给定的关键词，给定不同的单词会导致对同一图像的不同描述。图 5.2 中，给定关键词“boy”，描述模型将生成围绕“boy”的语句“a boy in a blue shirt is jumping on a field”，而给定“table”时，生成的语句“a little girl is sitting at a table”是聚焦于图像中“table”区域的描述。关键词的获取至关重要，在实际应用中，关键词可以通过人工给定或自动生成得到，例如图像检索中的语义标签、人机交互中的用户输入或目标检测生成的多个类别概念等。本章采用从真实标定中获取和从目标检测生成类别中获取这两种方法选择关键词。

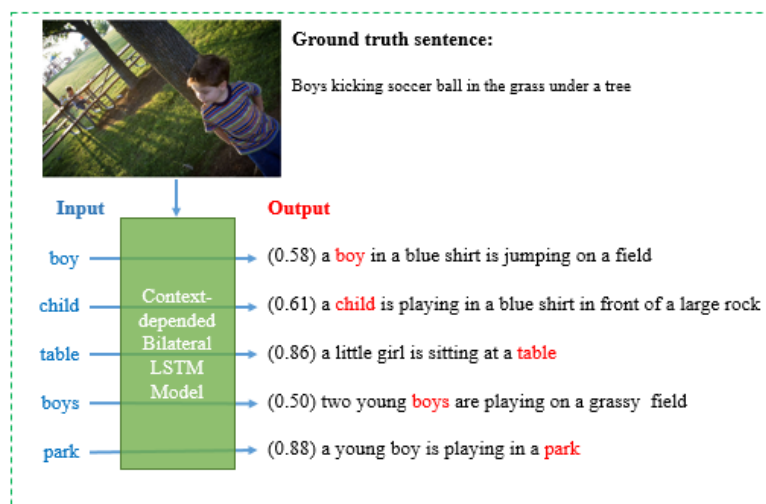


图 5.2 本章工作样例展示

本章提出了一种上下文依赖的双边 LSTM (CDB-LSTM)模型来预测由关键词驱动的图像描述，通过关键词对图像描述进行引导以达到针对一幅图像根据不同关键词生成多个不同描述的预期目标。如图 5.2 所示，输入为图像，通过 CDB-LSTM 模型，输出多个围绕对应关键词的描述语句。CDB-LSTM 包含两个级联的子模型，第一个模型以相反的顺序（从给定的关键词到语句的开始）生成语句的前半部分，而第二个模型根据第一个模型的预测结果生成语句的后半部分（从给定单词到语句末尾）。语句的前半部分和后半部分连接在一起构成最终的一个完整语句。通过考虑上下文传输模块的单词依赖性，这两个子模型在端到端训练框架中进行联合优化。定性评估表明，所提出的模型显示了对各种关键词（真值或检测到的关键词）的良好适应性。定量评估和用户评测也表明，所提出的方法生成的语句比传统的图像描述结果更具多样性和个性化。

5.2 模型定义和实现

对于图像 I ，给定 CNN 特征 v 和输入关键词 w ，目标是生成一个语句 $Y = \{y_1, \dots, y_t | y_i \in D\}$ 来描述图像内容，其中 D (Dictionary) 为存储所有单词的词典， y_i 为语句 Y 中的单词并构成了整个语句，输入关键词 w 为输出语句 Y 的一部分 $w \in Y$ 。为了利用给定的关键词，假设语句从给定单词开始生成，并向前后两个方向分别扩展生成：

- a) 反向预测以逆序生成语句的前半部分 $Y^b = \{y_1, \dots, y_{t^b} | y_i \in D\}$;
- b) 正向预测以正序生成语句的后半部分 $Y^f = \{y_{t^b+2}, \dots, y_t | y_i \in D\}$ 。

其中上标 b (backward) 和 f (forward) 分别代表反向和正向两个方向， t^b 为常数，表示前半部分语句的最后一个单词所在位置（时刻），第 $t^b + 1$ 个单词为关键词 w ，后半部分语句则从 $t^b + 2$ 开始。最终的描述语句为关键词和子句的组合 $\{Y^b, w, Y^f\} = \{y_1, \dots, y_{t^b}, y_{t^b+1}, y_{t^b+2}, \dots, y_t | y_i \in D\}$ 。本章用上下文依赖的双边 LSTM 模型(CDB-LSTM)实现这种交互式描述流程。

为了与预测句子 Y 进行区分，定义真实标注序列为 $S = \{s_1, \dots, s_t | s_t \in D\}$ ，根据关键词 w 将 S 分成三部分 $S = \{S^b, w, S^f\}$ ，其中各符号含义与 Y 相同。图像描述任务的目标函数为最大化对数似然和：

$$\Theta^* = \arg \max \sum \log p(s_t | s_{t-1}, v, w; \Theta) \quad (5.1)$$

其中 Θ 表示模型参数。

5.2.1 子模型训练

本章基于基准 CNN-RNN 图像描述框架（见本文第 2 章 2.3 节）进行模型设计。如图 5.3 所示，CDB-LSTM 包含两个级联子模型：反向 LSTM 模型 (B-LSTM) 和正向 LSTM 模型 (F-LSTM)。子模型分别用来生成从关键词出发向两个方向预测的子句，并通过上下文传输模块连接组成了一个端对端的统一模型。对整个模型来说，CDB-LSTM 的输入是图像特征 v 和关键词 w ，最终的输出语句包含三个部分：后向输出 Y^b 、关键词 w 和前向输出 Y^f 。在训练阶段，定义真实标注序列为 $S = \{s_1, \dots, s_t\}$ ，根据关键词 w 将 S 分成三部分 $S = \{S^b, w, S^f\}$ 。

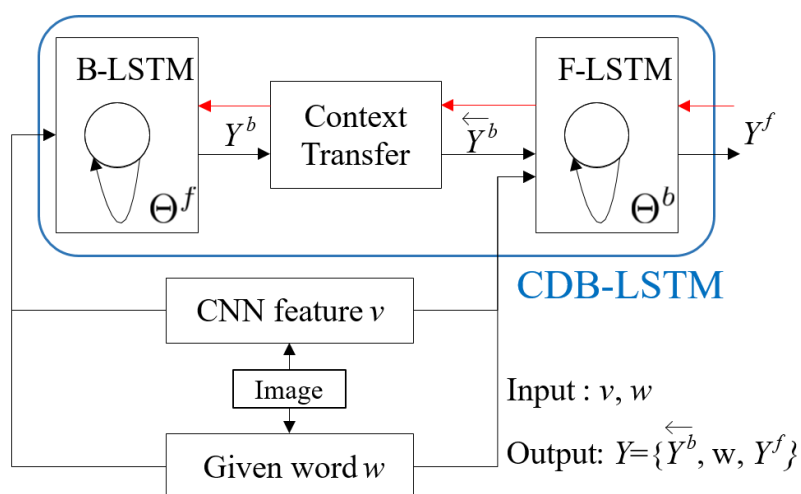


图 5.3 CDB-LSTM 模型框架

在基准 CNN-RNN 描述框架中, 图像 CNN 特征 v 被输入 RNN 描述模型用来进行语句预测。RNN 中 $t-1$ 时刻的隐藏状态 h_{t-1} 和当前输入 x_t 决定了当前时刻输出概率 p_t , 进而生成第 t 个输出单词 y_t 。CDB-LSTM 模型基于此框架进行设计, 并且采用 LSTM 作为 RNN 模型的核心。

1) 反向 LSTM 模型

反向 LSTM 模型(Backward LSTM, B-LSTM)是图 5.3 中 CDB-LSTM 模型的第一个子模型, 按照逆序预测语句的前半部分。模型设计过程类似于传统的 LSTM 模型, 不同之处在于起始单词, 传统模型的起始单词为启动词“0”, 而 B-LSTM 的启动词为关键词 w 。

对于给定 CNN 特征 v 和关键词 w 的图像 I , 假设真实标注语句的前半部分 S^b 包含 M 个单词, 模型训练过程如下:

$$\begin{aligned}
 x_0 &= W_{ix}v, t=0 \\
 x_1 &= W_{wx}w, t=1 \\
 x_t &= W_{ex}s_t, t \in \{2, \dots, M+1\} \\
 h_t &= LSTM(h_{t-1}, x_t), t \in \{1, \dots, M+1\} \\
 y_t, p_t &\propto \exp(W_{hp}h_t), t \in \{1, \dots, M+1\}
 \end{aligned} \tag{5.2}$$

其中 $y_t \in Y^b$ 为模型输出, W_{**} 表示相应的线性编码或解码参数: 视觉特征编码 W_{ix} , 关键词编码 W_{wx} , 单词编码 W_{ex} , 输出解码 W_{hp} 。同本文第 2 章 2.3 节相同, s_t 使用 One-hot 编码进行单词特征表达。

CNN 特征 v 经线性映射为输入 x_0 后在 $t=0$ 时进入 LSTM 网络, 为 LSTM 提供图像内容的全局特征表达; 关键词 w 经线性映射为 x_1 后在 $t=1$ 时输入到 LSTM 网络, 对模型进行强约束, 这也是为了保证后面的预测与这个关键词有关; 之后的循环依次输入真实标注 S^b 中的单词 (在测试阶段, 这部分每一时刻的输入为前一时刻的输出 y_{t-1})。

B-LSTM 的损失函数是每个时刻预测单词的负对数似然和:

$$L(B/w) = -\sum_{t=1}^{M+1} \log p(s_t | s_{1:t-1}, v, w; \Theta^b) \quad (5.3)$$

其中 Θ^b 表示反向子模型的所有学习参数。

B-LSTM 模型主要预测和关键词密切相关的内容。例如, 当给定关键词是一个目标时, 反向预测描述的前半部分, 如同目标的“形容词”描述。B-LSTM 生成的反向输出 \bar{Y}^b 将被输入到之后的正向 LSTM 模型。

2) 正向 LSTM 模型

正向 LSTM 模型(Forward LSTM, F-LSTM)是 CDB-LSTM 模型的第二个子模型, 以正序预测语句的后半部分。

对于图像 I , 给定 CNN 特征 v 、关键词 w 、语句前半部分输出 \bar{Y}^b , 假设真实标注语句的后半部分 S^f 包含 N 个单词, F-LSTM 模型的训练过程如下:

$$\begin{aligned} x_0 &= W_{ix} v, t=0 \\ x_t &= W_{ex} \bar{Y}_t^b, t \in \{1, \dots, M\} \\ x_t &= W_{wx} w, t = M+1 \\ x_t &= W_{ex} s_t, t \in \{M+2, \dots, M+N+2\} \\ h_t &= RNN(h_{t-1}, x_t), t \in \{1, \dots, M+N+2\} \\ y_t, p_t &\propto \exp(W_{hp} h_t), t \in \{M+1, \dots, M+N+2\} \end{aligned} \quad (5.4)$$

其中 $s_t \in S^f$, W_{**} 和 B-LSTM 中对应参数具有相同的含义。F-LSTM 模型在 $t=0$ 时输入为特征 v , 之后的 M 个输入为 B-LSTM 的逆序输出 \bar{Y}^b , 其中 M 为 Y^b 长度。 $t=M+1$ 时输入为关键词 w , 之后的 N 个输入依次为 S^f 中的单词 (在测试阶段, 这部分每一时刻的输入为前一刻的输出 y_{t-1})。F-LSTM 模型的前 M 时刻不产生输出, 仅通过隐藏层记忆输入信息。

F-LSTM 的损失函数值是后面的 N 个预测词的负对数似然和:

$$L(F/B, w) = - \sum_{t=M+1}^{M+N+2} \log p(s_t | s_{M+1:M+N+2}, v, w; \Theta^f) \quad (5.5)$$

其中 Θ^f 表示正向子模型的所有学习参数。

B-LSTM 的输出 Y^b ，F-LSTM 的输出 Y^f 和关键词 w 的组合构成了最终预测语句。由于两个子模型是级联的，因此 CDB-LSTM 模型的循环次数为 $2M+N+3$ 。

5.2.2 联合优化

每幅图像的训练数据由输入图像特征 v ，关键词 w 和真实标注序列 S 组成。根据关键词 w 将真实标注 S 分成三部分 $S = \{S^b, w, S^f\}$ 。本章工作的目标是要学习 CDB-LSTM 中两个子模型的参数：

$$\begin{aligned} \Theta^b &= \{W_{ix}^b, W_{wx}^b, W_{ex}^b, W_{hp}^b, LSTM^b\} \\ \Theta^f &= \{W_{ix}^f, W_{wx}^f, W_{ex}^f, W_{hp}^f, LSTM^f\} \end{aligned} \quad (5.6)$$

模型训练时，F-LSTM 使用 B-LSTM 的输出作为其输入的一部分，而 F-LSTM 的损失误差通过神经网络反向传播反馈给 B-LSTM。两个子模型通过上下文传输模块（图 5.3 中的 Context transfer module）进行连接并联合优化。

在 CDB-LSTM 模型的正向传播过程中，通过上下文传输模块，B-LSTM 模型的输出被迁移到 F-LSTM 模型的输入中：

$$x_t^f = W_{ex}^f \overline{Y_t^b} \quad (5.7)$$

在 CDB-LSTM 模型的反向传播过程中，F-LSTM 的损失误差通过上下文传输模块反馈给 B-LSTM：

$$\frac{\partial L}{\partial y_t^b} = \frac{\partial L}{\partial L_t^b} \frac{\partial L_t^b}{\partial y_t^b} + \frac{\partial L}{\partial L_t^f} \frac{\partial L_t^f}{\partial y_t^b} \quad (5.8)$$

其中第二项 $\frac{\partial L}{\partial L_t^b} \frac{\partial L_t^b}{\partial y_t^b}$ 是通过 F-LSTM 反传到 B-LSTM 的梯度，这部分使得

F-LSTM 对 B-LSTM 产生影响，并将两个方向的子句连接起来。

整个模型的损失函数为两个子模型损失函数之和：

$$\begin{aligned} L(B, w, F) &= L(B/w) + L(F/B, w) \\ &= -\sum_{t^b=1}^{M+1} \log p(y_{t^b}^b | x_{t^b}^b; \Theta^b) - \sum_{t^f=M+1}^{M+N+2} \log p(y_{t^f}^f | x_{t^f}^f; \Theta^f) \end{aligned} \quad (5.9)$$

在给定关键词 w 的引导下，子模型能够更快地收敛，联合损失 $L(B, w, F)$ 使得两个子模型能够进行联合优化（详见实验部分）。

5.2.3 对比模型 I-LSTM

图像描述领域之前的工作中没有引入关键词驱动的方法，为了进行模型对比，本章也提出了不考虑上下文依赖性的双边 LSTM 模型，称为 I-LSTM (Independent Bilateral LSTM)。该模型也由两个子模型组成，每个子模型分别预测前半部分和后半部分。和 CDB-LSTM 相比，I-LSTM 的子模型各自独立互不影响，即不存在联合优化和上下文传输模块。I-LSTM 在实验中用于证明 CDB-LSTM 模型上下文传输模块的有效性，因此其相关模型设置和公式定义都与 CDB-LSTM 相同。

5.3 关键词获取

输入关键词的获取是本章工作的一个重要部分。CDB-LSTM 模型训练中使用的关键词 w 来自于真实标注语句，从训练集所有图像的真实标注语句中选择并构建关键词词汇表。在测试过程中，关键词可以是任意的，但应该与图像相关。

5.3.1 训练过程关键词来源

目标类别对于关键词提取至关重要，本章主要用 MS COCO^[130]数据集进行模型训练和验证，因为 MS COCO 除了图像描述标注，还包含目标类别标注，而其他图像描述数据集大多仅包含图像描述标注，无法为关键词提取提供参考标准。由于图像描述是基于图像中的目标的语义表达，所以基于目标类别选取关键词是合理的，基于关键词的描述也就成了基于不同目标的特定描述，符合

本章工作的出发点，即从局部语义的角度实现图像描述的多样性和个性化。

在语言模型训练阶段，利用 MS COCO 的 80 个目标类别和所有的语句标注进行关键词词汇表构建：

a) 根据所有描述语句标注建立一个语义词典 D ，这个词典中的单词也是语言模型用于语句训练和生成的单词集合；

b) 通过 NLTK 工具箱^[149]中的词性辨别模块对 D 中的所有单词进行词性判断，选出所有名词单词，在 NLTK 中其词性为“NN”或“NNS”；

c) 通过 Gensim^[88]的词向量工具对 80 个目标类别和 D 中名词建立映射表，名词中所有和 80 类目标概念相似的单词构成了最终的关键词词汇表。

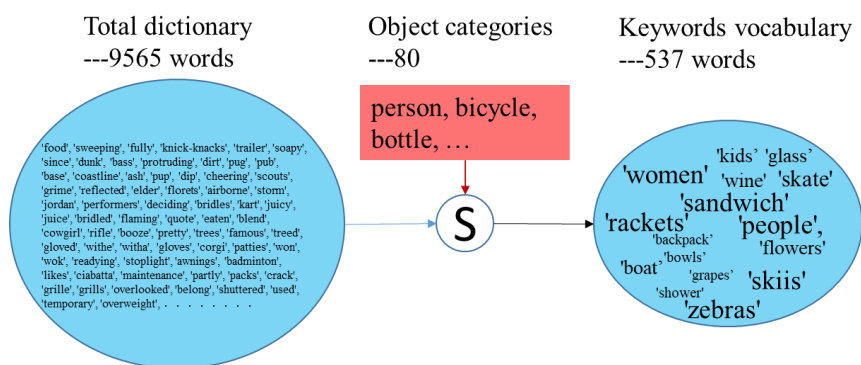


图 5.4 关键词生成

实际操作中，语义词典 D 包含 9565 个单词，最终的关键词词汇表包含 537 个单词，词汇表构建过程如图 5.4 所示。

5.3.2 测试过程关键词获取

对于测试阶段，关键词理论上可以是任意的，但为保障其有效性，关键词应该与图像中的目标相关。本章采用两种不同的策略来选择测试关键词：真实标注关键词和目标类别关键词。真实标注关键词来自于关键词词汇表，其语义空间与语言模型的输入和输出语义空间相同，因此在测试时可直接使用真实标注关键词作为模型输入。目标类别关键词来自于目标检测过程，其语义空间与语言模型语义空间是不同的，因此不能直接用于模型测试。为了解决目标类别与关键词词汇的不匹配问题，本章使用 Gensim 词向量工具^[88]根据单词相似性

将目标类别映射到关键词词汇表，完成一对多的映射，如图 5.5 所示。需要说明的是，这个策略和关键词词汇表构建时从目标类别到词汇表的映射条件是相同的。

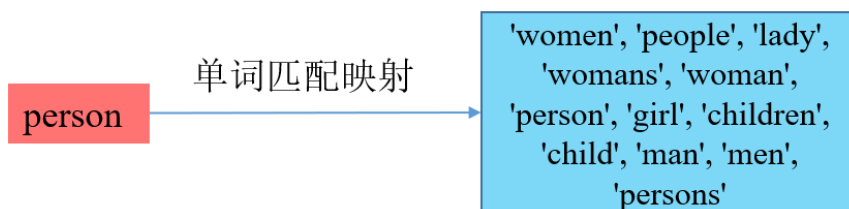


图 5.5 目标类别到关键词的匹配映射

5.4 实验验证与模型分析

本章在 MS COCO 数据集^[130]上对所提出的 CDB-LSTM 进行模型训练和测试。用于训练、验证、测试的数据量和第 2 章 2.5 节所述相同。

本章采用自动评价和人工评测两种方式进行模型评估。自动评测依然使用 B @ n, M, C, R 来分别代表 n-gram BLEU, METEOR, CIDEr, ROUGE_L。人工评测采用征集志愿者进行调查问卷的方式进行评估。

5.4.1 模型参数设置

1) CDB-LSTM

本章提出的 CDB-LSTM 模型训练参数根据 Karpathy 和 Li^[49]的工作中所采用的公共参数以及经验进行设置。内核 LSTM 的输入层、隐藏层和输出层的维度设置为 512，学习率开始为 4e-4，训练数据每循环 10 次后降低一半。模型的随机梯度下降方法使用自适应学习率算法 RMSProp^[144]。参数 beam_size 为每个单词预测时的概率搜索空间，可凭经验将此值设置为 7^[49]、10^[62]或 20^[52]。理论上，更高的 beam_size 会带来更好的性能，但同时会导致测试速度下降。本章将 beam_size 大小设置为 1，因为我们在实验中观察到更大的 beam_size 对整体模型性能提升并不明显。而且，采用较小的 beam_size 节省了很多计算时间。

CNN 网络使用 16 层 VGGNet^[4]，并基于基准 CNN-RNN 框架进行微调以获取更符合描述任务的视觉特征。

在创建单词词典 D 时丢弃训练语句中出现少于五次的单词。将描述的最大长度设置为 20，长度超过此值的将被剪裁。

2) LSTM

在后面的实验结果中 LSTM 代表传统的 CNN-RNN 描述模型，也是本章的基准模型。为了使得实验对比尽可能的公平，LSTM 中和 CDB-LSTM 相同的变量都采用同样的参数设置。

3) I-LSTM

I-LSTM 在实验中用于证明 CDB-LSTM 模型上下文传输模块的有效性。为了使得对比公平，I-LSTM 中所有和 CDB-LSTM 相同的变量都采用同样的参数设置。

5.4.2 实验结果

1) 自动评测

表 5.1 显示了 MS COCO 数据集上不同模型的实验结果。

对于 CDB-LSTM 模型，本章第 5.3 节采取了两种方式来选择关键词：

- a) 从真实标注语句中选取(Ground Truth, G)，即采用训练关键词；
- b) 从目标检测类别(Detected Labels, D)中选取。

对这两种关键词选取方式分别进行测试，用 G 和 D 代表。

本章提出的模型对一幅图像能够生成多个描述，评测结果就产生多个得分。由于结果对比时需要以每幅图像一个得分的形式，所以采用两种方式进行分数选取：

- a) 随机抽取(Random, R)一句进行评测；
- b) 在多个评测结果中选取最大值(Maximum, M)。

当选择真实标注关键词作为输入时，“GR”表示随机选择关键词获得的结果，“GM”表示对多个语句评测选取最大值的结果。当选择目标检测^[118]生成的关键词作为输入时，“DR”表示随机选择检测到的关键词作为输入，“DM”表示对多个检测关键词生成的语句评测的最大值。

表 5.1 自动评测结果对比

模型	B@1	B@2	B@3	B@4	M	C	R
NeuralTalk ^[49]	62.5	45.0	32.1	23.0	19.5	--	--
Google NIC ^[52]	66.6	46.1	32.9	24.6	--	--	--
Soft-Attention ^[64]	70.7	49.2	34.4	24.3	23.9	--	--
Hard-Attention ^[64]	71.8	50.4	35.7	25.0	23.0		--
gLSTM ^[62]	63.8	46.3	33.6	24.8	23.3		--
ATT ^[66]	70.9	53.7	40.2	30.4	24.3	--	--
Adaptive ^[63]	74.0	58.0	43.9	33.2	26.6	108.5	--
SCN-LSTM ^[61]	74.1	57.8	44.4	34.1	26.1	104.1	--
LSTM-A ^[65]	73.0	56.5	42.9	32.5	25.1	98.6	53.8
LSTM	69.8	52.2	38.5	28.7	23.9	53.4	42.9
I-LSTM (GR)	45.3	34.8	24.9	17.3	18.5	64.9	45.0
I-LSTM (GM)	66.1	50.6	35.0	23.4	20.9	77.2	48.0
CDB-LSTM (GR)	73.1	53.2	35.8	23.6	21.6	78.5	49.9
CDB-LSTM (GM)	78.8	58.3	40.4	27.5	23.4	83.6	51.8
CDB-LSTM (DR)	62.9	42.5	27.9	18.4	17.2	47.1	43.2
CDB-LSTM (DM)	76.3	56.1	38.9	26.5	22.5	77.3	51.4

从表 5.1 可以看出,对于本文提出的 CDB-LSTM,采用目标检测结果作为关键词(D)时的描述性能要低于采用真实标注关键词(G),特别是随机选取一个关键词进行评测时,“DR”远远小于“GR”。因为目标检测有一定的误差存在,导致根据检测关键词生成的语句误差也增大,同时对单幅图像来说,生成的多个语句更加不均衡,从而使得随机选择的结果更差。目标检测性能的提升将会直接影响其对应的关键词驱动的描述结果。尽管目标检测获得关键词的误差导致了语言模型描述性能下降,但优点是不需要人工干预。具体应用中,用户给定关键词相当于专家先验,等效于真实标注关键词。在有自动化需求的应用系统中,基于目标检测结果进行关键词选取还是很有意义的。

对于同类型模型来说(I-LSTM 或 CDB-LSTM), 随机取值(R)总是不大于最大值(M), 这也是显而易见的。针对同类型的关键词选取(GR 或 GM), CDB-LSTM 的性能要远远高于 I-LSTM 模型。I-LSTM 中的两个模型彼此是独立的, 导致了不连贯、不完整或不准确的结果。例如, 图 5.6 中第一幅图像给定“trick”或“jacket”时, I-LSTM 都不能预测句子的后半部分, 这是因为在不考虑上下文信息的情况下, 子模型认为给定关键词为结束词, 从而使得整个描述语句不完整。这说明了额外关键词的指导并不一定能保证描述性能一定会提升。过短的描述结果也会导致自动评估时对语句的长度惩罚增加, 使得评测结果更低。理想的最大值 I-LSTM(GM)显著高于随机值 I-LSTM(GR), 这主要是由于长度惩罚造成了多个评测结果的不均衡, 即最高评测值远远高于最低评测。与 I-LSTM 相比, 提出的 CDB-LSTM 与各种输入关键词具有良好的适应性和语义一致性。



图 5.6 CDB-LSTM 与 I-LSTM 对比

CDB-LSTM(GM)达到了最优性能, 这是因为它从由真实标注关键词驱动的多个语句中选择理想的最大值。

结果表明, 上下文依赖性对于基于关键词的图像描述极其重要。直观上看, 额外的输入可以提高图像描述的性能, 但是实验证明, 仅使用简单的关键词嵌入技术(I-LSTM)并不能很好地完成任务。所提出的 CDB-LSTM 能够有效地利用关键词来产生比现有技术方法更准确的、更有针对性的语句。

另外需要说明的是, 本章的模型输出是聚焦于关键词的, 关键词驱动的描述对于同一图像具有不同的侧重点, 因此 CDB-LSTM 生成的描述语句和评测

中采用的真实标注语句是有偏差的。

以图 5.7 中图像为例，由“baseball”和“children”这两个关键词驱动的句子比 LSTM 的结果要好，但评价分数较低，这是因为 CDB-LSTM 的结果与真实标注语句在目标和定义上已经产生了很大不同。真实标注语句本身也是人工标注的，对于复杂的图像内容来说，人工标注本身也是有偏差的。即使在这种情况下，CDB-LSTM(GR)的结果仍然优于 LSTM 模型，这也验证了本章方法的有效性。与传统的描述模型相比，本章提出的模型能够从全局角度生成各种侧重点不同的语句。由于 CDB-LSTM 和 LSTM 之间语义偏差的存在，本章在自动评测之外，额外进行了人工评测以获得更加公平的对比。



图 5.7 基于关键词驱动的图像描述

2) 人工评测

每种自动评测方法都会根据输出语句和真实标注语句之间的相似度计算一个得分。然而，这些自动评测方法在语言理解上饱受质疑，因为它们与人类判断有微弱的负相关或者没有相关性^[39, 150, 151]。此外，作为关键词驱动的任务，人工评测使得用户可以介入交互信息，符合本章工作的出发点。

我们从 MS COCO 测试集中随机选择 100 张图像，并邀请 50 位志愿者进行

人工评测。评分标准沿用 Hodosh 等^[129]和 Oriol Vinyals 等^[52]的做法，志愿者需要对两个实验结果(LSTM 和 CDB-LSTM)进行打分。对于 CDB-LSTM 模型，关键词的选择由志愿者指定。对描述语句的打分采取 4 分制，分数可选值为{1, 2, 3, 4}，越高的评分代表此描述结果更符合志愿者的预期。最后统计其累计分布图，如图 5.8 所示。提出的 CDB-LSTM 的结果要高于传统的 LSTM，这意味着通过用户输入关键词，生成的描述语句更加符合用户的需求或预期，实现了图像描述的多样性和个性化。

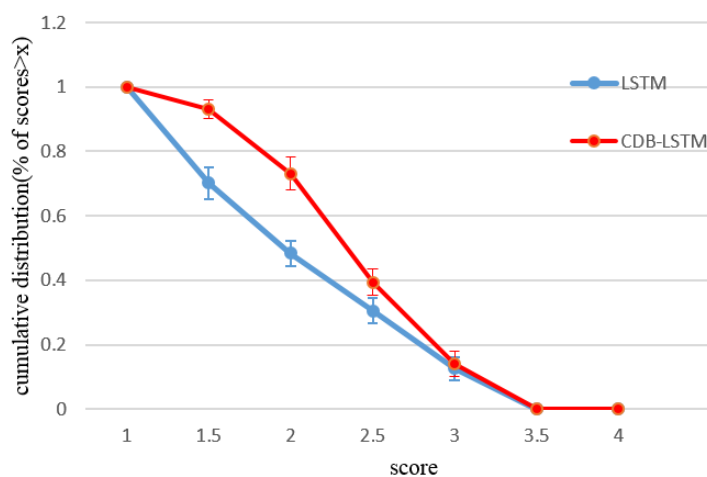


图 5.8 人工评测结果对比

5.4.3 模型误差分析

本节对所提出的模型进行分析，并针对训练损失函数值进行模型对比。如图 5.9 所示，所提出的 CDB-LSTM 模型（红色线条）比传统的 LSTM（灰色）和 I-LSTM（子模型分别为蓝色和绿色）模型具有更低的损失误差。

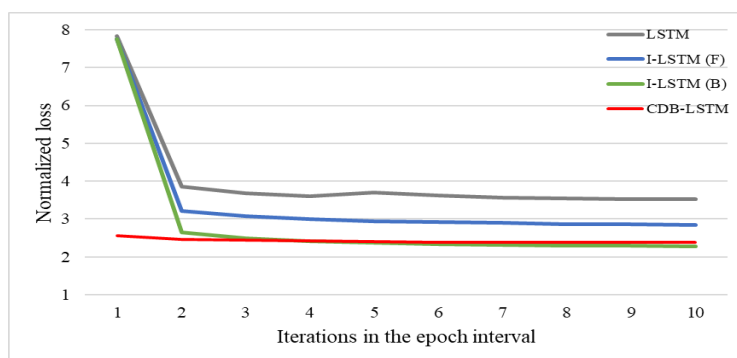


图 5.9 模型训练损失误差对比

I-LSTM 的反向子模型 (B) 比正向子模型 (F) 训练损失误差更小。这是因为一个描述语句中, 关键词的前面部分语句通常是给定目标的属性或量词表达, 与关键词密切相关, 也相对容易进行表达。关键词的后面部分语句通常包含场景、其他目标、目标行为、目标之间关系等信息, 语义更为复杂也更加难以表达。如图 5.7 所示, 第二幅图像真实标注语句为 a group of children playing baseball out side, 关键词 children 前面部分 “a group of” 比较简单, 和 “children” 密切相关, 也比较容易训练和预测。而后半部分 “playing baseball out side” 内容比较丰富, 相对前半部分语句较难训练。因此正向子模型 I-LSTM(F)比反向子模型 I-LSTM(B)的训练难度要大, 这也是 CDB-LSTM 模型选择先反向后正向的级联顺序的原因。

在具体实验中, 提出的 CDB-LSTM 在 I-LSTM 模型的基础上进行训练, 即 CDB-LSTM 的参数初始化采取 I-LSTM 训练后的参数, 这样可以加快模型训练。根据图 5.9 所示, CDB-LSTM 在所有训练数据循环 10 次时, 损失误差达到一个平衡值, 介于 I-LSTM 两个子模型之间, 略高于 I-LSTM(B), 这也说明了 CDB-LSTM 通过引入上下文传输模块, 使得子模型得到了联合优化。

值得说明的是, 图 5.9 中, 虽然 I-LSTM 的训练误差低于 LSTM, 但在测试时, I-LSTM 的预测结果反而比 LSTM 低 (见表 5.1)。这是因为 I-LSTM 中的两个子模型彼此不相关, 导致了不完整、不准确的结果 (见图 5.6), 也使得总体评测时对句子长度的惩罚加大, 导致最终评测性能较低。

5.4.4 定性分析

图 5.7 显示了由本章提出的 CDB-LSTM 生成的一些示例描述。从示例结果中可以得出以下结论:

1) 从整体效果看, CDB-LSTM 的结果增加了描述的丰富性和多样性。CDB-LSTM 可以根据不同的关键词生成不同的包含对应关键词的语句, 实现了在全局视角下针对关键词的特定描述;

2) 从单句效果来看, CDB-LSTM 能够校正传统方法的错误。如图 5.7 第二幅图像的 LSTM 结果中的 “frisbee”, 属于目标识别错误, 而通过给定关键词的方式限定 “baseball”, 直接避免了这种判别错误, 从而提升了整个描述的准确

度;

3) CDB-LSTM 能够处理不显著目标并根据关键词进行局部或全局描述。如第一个示例中的“table”，传统方法会忽略掉这个不太显著的部分，而本章提出的方法可以根据关键词自动定位到 table 区域，并对其进行相关描述，使得图中坐在桌子旁边的小孩也被准确描述出来。第二个示例中的“gloves”也是非显著目标，模型能够据此生成与关键词相关的描述，而这个描述是具有全局视角的。

由于训练集的真实标注描述大多是五个相似的语句，因此对于同一图像，模型生成的最终结果也是相似的。如果真实标注描述更加多样化，提出的模型预测语句也可以具备多样性，这个特点将在实验扩展里详细说明。

5.5 实验扩展

5.5.1 单幅图像多个描述生成

对单幅图像生成多个描述能够弥补单个语句的不足，除了本文提出的方法，还有两种直观的方法可以为单幅图像生成多句个性化描述。

第一种方法是通过扩大输出单词概率取值范围获得多个图像描述，然后从多个描述结果中搜索包含关键词的语句。在实验中，单词输出概率取值范围是由 beam_size 参数控制的，其值越大获得的语句越多。实验中观测到，通过扩大 beam_size 生成的多个语句都倾向于表达显著的目标，而大约一半的测试样本（2454/5000）无法在生成的描述中找到给定的真实关键词。而且，生成的结果彼此太相似，不能满足多样性描述需求。如图 5.10 所示，在模型预测时，设定 beam_size=10，即选择输出概率排行前 10 的语句作为描述结果。

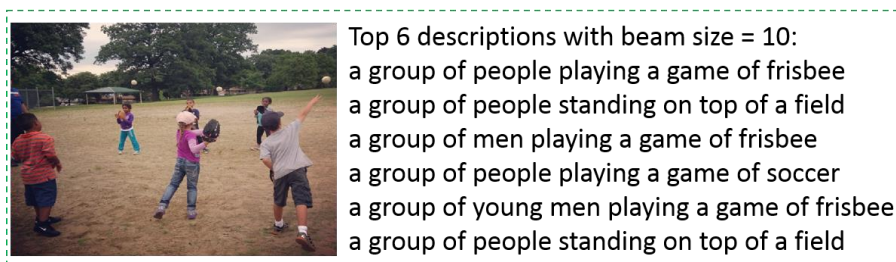


图 5.10 传统图像描述模型通过扩大 beam_size 参数生成的语句

图 5.10 中展示了最终结果中排行前 6 的语句，可以看出，生成的 6 个语句非常相似，且不包含图像中非显著目标如 “gloves” 和 “baseball”。

第二种方法是基于局部区域的图像描述。Johnson 等^[57]和 Krishna 等^[68]提出的模型能够对图像中的多个局部区域进行特征提取和密集语句生成，然而，这些方法倾向于生成短语，而不是完整的语句描述。本文第 3 章的工作也对多个目标区域进行语言描述，生成的句子比较完整，但也受目标检测边框准确率和数据集标注的影响。

本章所提出 CDB-LSTM 着重于在全局视角下描述具有不同侧重点的图像内容。额外关键词的引入使得提出的模型能够满足用户的需求，使得描述内容更加个性化并且具备多样性。由于采用数据集的五个真实标注描述大多是相似的语句，因此对于同一图像的描述更多的是体现了个性化，对多样性的展示还不够充分。如果真实标注语句更加多样化，CDB-LSTM 模型可以预测更丰富多样的描述。因此，在扩展实验中，采用 CDB-LSTM 模型在 Visual Genome 数据集^[68]上进行了密集描述。Visual Genome 中的图像描述标注都是局部描述，每幅图像平均包含 42 个局部描述。这些描述语句足够的丰富多样，每个描述都是表达局部区域的长度范围从 1 到 16 个字的短语。

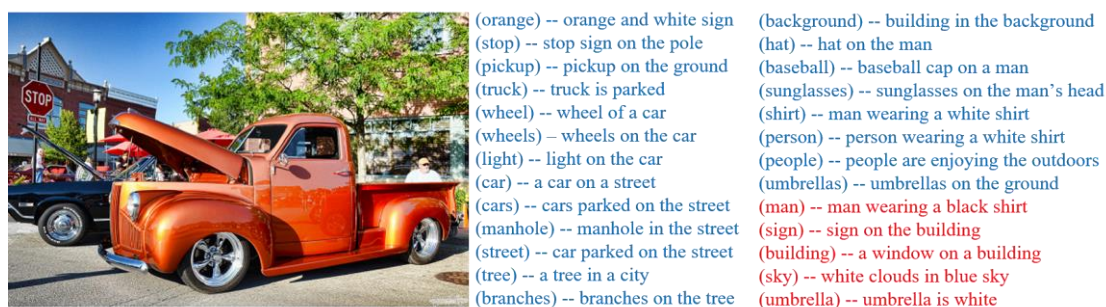


图 5.11 CDB-LSTM 在 Visual Genome 数据集上的描述结果

图 5.11 显示了 Visual Genome 数据集中由 CDB-LSTM 生成的描述示例。其中蓝色的语句意味着正确的预测，而红色的语句代表预测错误的语句。可以看出，这些语句是针对局部区域的长度很短的短语描述，例如 “orange and white sign”， “wheel of a car”。虽然展示的仅为部分示例，但实验中观测到，描述正确的概率较高。值得注意的是，orange 为形容词，在关键词选择时将其认为是

名词。尽管不同词性代表的语义有很大差异，提出的 CDB-LSTM 模型依然可以处理这种语义上为非名词的关键词。在后面的基于元素关键词的图片描述实验中将会进一步探索这个特性。

自动评测结果如表 5.2 所示，这些分数比之前在 MS COCO 数据集上的实验结果要高。这是因为 Visual Genome 数据集下生成的基本都是长度小于 10 的短语。相比于完整的句子，关键词在短语中的占比更大，导致预测结果和真实标注语句之间的相似度更高，评测结果也就更好。

表 5.2 Visual Genome 数据集评测结果

模型	B@1	B@2	B@3	B@4	M	C	R
CDB-LSTM	89.0	70.3	53.1	40.6	30.9	62.4	39.0

5.5.2 基于元素关键词的描述

本章的之前实验中，MS COCO 和 Visual Genome 数据集都包含图像目标标注，因此可以很好地应用提出的 CDB-LSTM 模型。对于没有目标标注的 Flickr8k^[129]和 Flickr30k 数据集^[133]，即缺乏真实标注目标关键词的情况下，我们在扩展实验中进行了基于元素(Element)关键词的描述。

本章定义元素包含目标名词(如 dog)、场景(如 ground, grass)、目标属性(如 brown, tan)、目标行为(如 jumps, swims)等。利用真实标注语句中元素单词的出现频率构建关键词词汇表，其中出现次数排名前 100 的名词被用来构建元素关键词词汇表 V。需要说明的是，英语单词中很多词汇具备一词多性的特点，比如“jump”同时是动词和名词。对所有名词进行排序选择，最终得到的词汇表 V 中的单词不仅仅是名词。很多具备一词多性的关键词在特定句子中会有不同的语义表达，因此得到的元素关键词可以是语句的名词、形容词或动词部分。这种策略很好地剔除了语义不明显的量词、冠词、介词、连词等，同时以较为直观简单的方式增加了关键词的复杂性，从而扩大了模型输入单词的接受范围，也增加了用户需求的选择空间。

图 5.12 中第一行的两幅图像显示了基于元素关键词(element-based)的描述

示例，第二行的两幅图像为本章工作中基于目标关键词(object-based)的描述，在此用来作为对比。从图 5.12 中可以看出，CDB-LSTM 模型对各种类型关键词具有很强的自适应性，能够灵活处理目标(dog, children, baseball 等)或非目标(tan, swims, jumps 等)关键词，实现了从不同角度对图像的细节描述。由于选择元素关键词的方式还有很多噪声，实验出发点和效果不如目标关键词有说服力，在此仅作为扩展实验进行探讨，未来的工作中可以更深入地探索基于不同词性的关键词的局部细节描述问题。



图 5.12 基于不同关键词的描述示例

5.6 本章小结

本章提出了一种基于关键词驱动的图像描述方法，能够根据用户需求进行交互式图像语义表达，从而解决了目前图像描述工作的单一性和局限性问题。通过提出的 CDB-LSTM 模型针对不同的关键词生成对应的个性化描述，实现了全局视角下的局部语义表达。CDB-LSTM 由两个级联的子模型构成，子模型在端到端框架中通过上下文传输模块连接起来并进行联合训练。上下文依赖关系有效地抑制了模型生成不连贯和不准确的语句的可能性，自动评测和人工评测都证明了本文工作的有效性。

第6章 总结与展望

6.1 本文工作总结

本文从局部语义学习的角度来解决图像描述中存在的表达局限性问题。通过三种新的描述方法对局部语义信息进行探索和应用，实现了增强图像理解并丰富图像内容描述的目的。本文主要研究成果总结如下：

(1) 通过图像区域描述进行局部语义表达。与现有的描述整幅图像内容的方法不同，提出了一种基于图像局部区域生成丰富图像描述的方法。首先使用目标检测方法来生成候选区域，然后训练 RNN 语言模型以学习全局图像和标注语句之间描述关系，最后利用 RNN 模型对目标候选区域进行局部描述生成和分析。所提出的方法能够对单幅图像生成多个针对不同区域的局部描述，这些描述具有足够的表达能力并且包含更详细的语义信息，实验验证了所提出方法的有效性。实验中还观测到了全局图像描述与局部描述之间具有互补关系，局部描述语句中甚至包含真实标注中不存在的有效信息。这些发现对于未来更进一步的局部语义学习工作是具有重要意义的。

(2) 利用从图像区域学习到的局部语义信息来生成语义特征并改进描述模型。首先，根据局部描述结果探讨局部语义对全局描述的影响；然后，进一步挖掘局部语义信息并生成语义特征，该特征不仅包含局部目标的详细信息，还与描述语句共享同样的语义空间，从而弥补了视觉图像与语义描述之间的模式差距；最后，将语义特征嵌入到提出的 EE-LSTM 模型中以预测最终的语言描述。EE-LSTM 通过额外的输入通道，将局部和全局、视觉和语义信息同时集成到语言模型中，提升了描述性能。实验结果验证了局部语义特征学习对图像描述模型的改进。

(3) 通过目标关键词驱动的图像描述来实现局部语义学习。与前面在目标区域上进行局部语义挖掘的工作不同，关键词驱动的图像描述探索基于目标概念的局部语义学习。提出了 CDB-LSTM 语言模型以生成由关键词驱动的个性化图像描述，实现了全局视角下的局部语义表达。CDB-LSTM 包含两个级联的子模型，通过考虑子模型之间的上下文依赖性，这两个子模型在端到端训练框

架中进行联合优化，使得模型针对各种关键词能够生成语法连贯又表达准确的语句。

6.2 未来工作展望

图像描述是计算机视觉和自然语言处理的结合，图像理解的进一步发展以及语言模型的进步都会对图像内容的表达产生重要的推动作用。而图像描述研究可以进一步推动视频描述和视觉问答等相关“视觉-语言”扩展问题的发展。基于本文的研究内容，未来可以从如下几个方面进行后续研究：

(1) 多词驱动的图像描述：本文提出的 CDB-LSTM 探讨了基于单个关键词驱动的图像描述，而未来基于多词驱动的描述将会进一步扩展模型对于关键词的接受范围，并使得人机交互方式更为灵活多样。例如，用户可以给定几个关键词，从而通过模型生成包含这些关键词的句子，可用于图像的精准检索。在自动化系统设定下，可以通过弱监督分类模型进行多个关键词的自动预测，然后用以生成更加准确的图像描述。

(2) 目标行为和关系描述：本文从目标区域和目标类别的角度引入局部语义学习，除了目标之外，图像描述还与目标行为和相互关系等信息密切相关，未来探索目标行为和关系在图像中的意义将会推动图像描述的进一步发展。

(3) 构建新的数据集：以构建新的数据集的方式引入更为丰富的图像语义信息对图像描述领域的贡献较为直观。目前，图像描述数据集的真实标注都为全局描述或局部描述，未来构建包含全局和局部描述、目标、场景、关系等标注全面的数据集对图像描述领域会有新的帮助。另外，针对复杂场景的图像描述数据集、具备多种语言描述标注的数据集都是很有意义的扩展研究方向。

(4) 评价机制的全方位分析：图像描述的评价机制是一个挑战性问题，人工评测的大规模应用不太现实，现有的各种自动评测方法和人工评测还是有偏差的，因此如何对图像描述结果进行更合理的评测需要更深入的探索研究，未来可以从图像描述问题的定义和特定视觉认知任务的目标出发，基于图像事实，考虑全局与局部、显著性、描述粒度等全方位多角度的综合评估。

参考文献

- [1] CHAPELLE O, HAFFNER P, VAPNIK V N. Support vector machines for histogram-based image classification[J]. IEEE Transactions on Neural Networks, 1999, 10 (5): 1055–1064.
- [2] HARALICK R M, SHANMUGAM K, OTHERS. Textural features for image classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973 (6): 610–621.
- [3] PERRONNIN F, SÁNCHEZ J, MENSINK T. Improving the fisher kernel for large-scale image classification[C]. Proceedings of European Conference on Computer Vision, 2010: 143–156.
- [4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. Proceedings of International Conference on Learning Representations, 2015.
- [5] YANG J, YU K, GONG Y, et al. Linear spatial pyramid matching using sparse coding for image classification[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2009: 1794–1801.
- [6] BOSCH A, ZISSERMAN A, MUÑOZ X. Scene classification via plsa[C]. Proceedings of European Conference on Computer Vision, 2006: 517–530.
- [7] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37 (9): 1757–1771.
- [8] KOSKELA M, LAAKSONEN J. Convolutional network features for scene recognition[C]. Proceedings of the ACM International Conference on Multimedia, 2014: 1169–1172.
- [9] KWITT R, VASCONCELOS N, RASIWASIA N. Scene recognition on the semantic manifold[C]. Proceedings of European Conference on Computer Vision, 2012: 359–372.
- [10] LI L J, SOCHER R, LI F F. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2009: 2036–2043.
- [11] BARNARD K, DUYGULU P, FORSYTH D, et al. Matching words and pictures[J]. Journal of Machine Learning Research, 2003, 3 (2): 1107–1135.

- [12] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2005: 886–893.
- [13] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D A, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32 (9): 1627–1645.
- [14] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2014: 580–587.
- [15] FARHADI A, ENDRES I, HOIEM D, et al. Describing objects by their attributes[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2009: 1778–1785.
- [16] LI L J, FEI-FEI L. What, where and who? classifying events by scene and object recognition[C]. Proceedings of IEEE International Conference on Computer Vision, 2007: 1–8.
- [17] SHARMA G, JURIE F, SCHMID C. Expanded parts model for human attribute and action recognition in still images[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2013: 652–659.
- [18] YAO B, JIANG X, KHOSLA A, et al. Human action recognition by learning bases of action attributes and parts[C]. Proceedings of IEEE International Conference on Computer Vision, 2011: 1331–1338.
- [19] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]. Proceedings of European Conference on Computer Vision, 2014: 818–833.
- [20] REITER E, DALE R. Building natural language generation systems[M]. Cambridge university press, 2000.
- [21] GUADARRAMA S, KRISHNAMOORTHY N, MALKARNENKAR G, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition[C]. Proceedings of IEEE International Conference on Computer Vision, 2014: 2712–2719.
- [22] KHAN M U G, ZHANG L, GOTOH Y. Towards coherent natural language description of

-
- video streams[C]. IEEE International Conference on Computer Vision Workshops, 2011: 664–671.
- [23] KRISHNAMOORTHY N, MALKARNENKAR G, MOONEY R, et al. Generating natural-language video descriptions using text-mined knowledge[C]. Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013: 541–547.
- [24] PASUNURU R, BANSAL M. Multi-task video captioning with video and entailment generation[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1273–1283.
- [25] ROHRBACH A, ROHRBACH M, TANDON N, et al. A dataset for movie description[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 3202–3212.
- [26] ROHRBACH M, QIU W, TITOV I, et al. Translating video content to natural language descriptions[C]. Proceedings of IEEE International Conference on Computer Vision, 2013: 433–440.
- [27] THOMASON J, VENUGOPALAN S, GUADARRAMA S, et al. Integrating language and vision to generate natural language descriptions of videos in the wild[C]. Proceedings of the 25th International Conference on Computational Linguistics, 2014: 1218–1227.
- [28] WU Q, WANG P, SHEN C, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 4622–4630.
- [29] YAO L, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure[C]. Proceedings of IEEE International Conference on Computer Vision, 2015: 4507–4515.
- [30] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]. Proceedings of IEEE International Conference on Computer Vision, 2015: 19–27.
- [31] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[C]. Proceedings of IEEE International Conference on Computer Vision, 2015: 2425–2433.
- [32] BIGHAM J P, JAYANT C, JI H, et al. Vizwiz: nearly real-time answers to visual questions[C]. Proceedings of the 23rd annual ACM symposium on User interface software

- and technology, 2010: 333–342.
- [33] GEMAN D, GEMAN S, HALLONQUIST N, et al. Visual turing test for computer vision systems[J]. Proceedings of the National Academy of Sciences, 2015, 112 (12): 3618–3623.
- [34] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition: volume 1, 2017: 9.
- [35] MALINOWSKI M, FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input[C]. Proceedings of Advances in Neural Information Processing Systems, 2014: 1682–1690.
- [36] WU Q, SHEN C, LIU L, et al. What value do explicit high level concepts have in vision to language problems?[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 203–212.
- [37] FANG H, GUPTA S, IANDOLA F, et al. From captions to visual concepts and back[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 1473–1482.
- [38] FARHADI A, HEJRATI S M M, SADEGHI M A, et al. Every picture tells a story: Generating sentences from images[C]. Proceedings of European Conference on Computer Vision, 2010: 15–29.
- [39] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35 (12): 2891–2903.
- [40] MITCHELL M, HAN X, DODGE J, et al. Midge: Generating image descriptions from computer vision detections[C]. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012: 747–756.
- [41] YANG Y, TEO C L, DAUMÉ III H, et al. Corpus-guided sentence generation of natural images[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011: 444–454.
- [42] SOCHER R, LIN C C, NG A Y, et al. Parsing natural scenes and natural language with recursive neural networks[C]. Proceedings of International Conference on Machine Learning, 2011: 129–136.

-
- [43] DEVLIN J, CHENG H, FANG H, et al. Language models for image captioning: The quirks and what works[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015: 100–105.
- [44] GUPTA A, VERMA Y, JAWAHAR C, et al. Choosing linguistics over vision to describe images[C]. the Association for the Advancement of Artificial Intelligence, 2012: 1.
- [45] KUZNETSOVA P, ORDONEZ V, BERG A C, et al. Collective generation of natural image descriptions[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012: 359–368.
- [46] KUZNETSOVA P, ORDONEZ V, BERG A C, et al. Generalizing image captions for image-text parallel corpus[C]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: volume 2, 2013: 790–796.
- [47] KUZNETSOVA P, ORDONEZ V, BERG T L, et al. Treetalk: Composition and compression of trees for image descriptions[J]. Transactions of the Association of Computational Linguistics, 2014, 2 (10): 351–362.
- [48] ORDONEZ V, KULKARNI G, BERG T L. Im2text: describing images using 1 million captioned photographs[C]. Proceedings of Advances in Neural Information Processing Systems, 2011: 1143–1151.
- [49] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 3128–3137.
- [50] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural language models[J]. arXiv preprint arXiv:1411.2539, 2014.
- [51] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[C]. Proceedings of International Conference on Learning Representations, 2015.
- [52] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 3156–3164.
- [53] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]. Proceedings of International Conference on Learning

- Representations, 2014.
- [54] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation[C]. Proceedings of Conference on Empirical Methods in Natural Language Processing, 2014: 1724–1734.
- [55] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 1–9.
- [56] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9 (8): 1735–1780.
- [57] JOHNSON J, KARPATHY A, FEI-FEI L. Densecap: Fully convolutional localization networks for dense captioning[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 4565–4574.
- [58] YANG L, TANG K, YANG J, et al. Dense captioning with joint inference and visual context[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition: volume 2, 2017.
- [59] CHEN L, ZHANG H, XIAO J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017.
- [60] CHO K, COURVILLE A, BENGIO Y. Describing multimedia content using attention-based encoder-decoder networks[J]. IEEE Transactions on Multimedia, 2015, 17 (11): 1875–1886.
- [61] GAN Z, GAN C, HE X, et al. Semantic compositional networks for visual captioning[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition: volume 2, 2017.
- [62] JIA X, GAVVES E, FERNANDO B, et al. Guiding the long-short term memory model for image caption generation[C]. Proceedings of IEEE International Conference on Computer Vision, 2015: 2407–2415.
- [63] LU J, XIONG C, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017: 3242–3250.
- [64] XU K, BA J, KIROUS R, et al. Show, attend and tell: Neural image caption generation with

-
- visual attention[C]. Proceedings of International Conference on Machine Learning, 2015: 2048–2057.
- [65] YAO T, PAN Y, LI Y, et al. Boosting image captioning with attributes[J]. OpenReview, 2016, 2 (5): 8.
- [66] YOU Q, JIN H, WANG Z, et al. Image captioning with semantic attention[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 4651–4659.
- [67] KRAUSE J, JOHNSON J, KRISHNA R, et al. A hierarchical approach for generating descriptive image paragraphs[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017: 3337–3345.
- [68] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123 (1): 32–73.
- [69] SZUMMER M, PICARD R W. Indoor-outdoor image classification[C]. Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, 1998: 42–51.
- [70] LOWE D G. Object recognition from local scale-invariant features[C]. Proceedings of IEEE International Conference on Computer Vision, 1999: 1150–1157.
- [71] AHONEN T, HADID A, PIETIKAINEN M. Face description with local binary patterns: Application to face recognition[M]. IEEE Computer Society, 2006: 469–481.
- [72] JULESZ B. Textons, the elements of texture perception and their interactions[J]. Nature, 1981, 290 (5802): 91–7.
- [73] SWAIN M J, BALLARD D H. Indexing via color histograms[C]. Proceedings of IEEE International Conference on Computer Vision, 1990: 390–393.
- [74] SMEULDERS A W M, WORRING M, SANTINI S, et al. Content-based image retrieval at the end of the early[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22 (12): 1349–1380.
- [75] TOWN C. Ontological inference for image and video analysis[J]. Machine Vision & Applications, 2006, 17 (2): 94–115.

- [76] YAO B Z, YANG X, LIN L, et al. I2T: image parsing to text description[J]. Proceedings of the IEEE, 2010, 98 (8): 1485–1508.
- [77] SIVIC J, ZISSERMAN A. Video google: a text retrieval approach to object matching in videos[C]. Proceedings of IEEE International Conference on Computer Vision, 2003: 1470–1477.
- [78] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3 (Jan): 993–1022.
- [79] PERRONNIN F, DANCE C. Fisher kernels on visual vocabularies for image categorization[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2007: 1–8.
- [80] LI L J, SU H, LIM Y, et al. Object bank: An object-level image representation for high-level visual recognition[J]. International Journal of Computer Vision, 2014, 107 (1): 20–39.
- [81] KHAN R, BARAT C, MUSELET D. Spatial orientations of visual word pairs to improve bag-of-visual-words model[C]. British Machine Vision Conference, 2012.
- [82] QIN D, CHEN Y, GUILLAUMIN M, et al. Learning to rank bag-of-word histograms for large-scale object retrieval[C]. British Machine Vision Conference, 2014: 1–12.
- [83] TIRILLY P, CLAVEAU V, GROS P. Language modeling for bag-of-visual words image categorization[C]. Proceedings of the international conference on Content-based image and video retrieval, 2008: 249–258.
- [84] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]. Proceedings of Advances in Neural Information Processing Systems, 2012: 1106–1114.
- [85] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
- [86] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
- [87] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]. Conference on Empirical Methods in Natural Language Processing, 2014: 1532–1543.

-
- [88] REHUREK R, SOJKA P. Software framework for topic modelling with large corpora[C]. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.
- [89] SHANNON C E. Prediction and entropy of printed English[J]. Bell Labs Technical Journal, 1951, 30 (1): 50–64.
- [90] GRAVES A. Generating sequences with recurrent neural networks[J]. Computer Science, 2013.
- [91] KIM Y, JERNITE Y, SONTAG D, et al. Character-aware neural language models[C]. the Association for the Advancement of Artificial Intelligence, 2016.
- [92] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]. Proceedings of the 31st International Conference on Machine Learning, 2014: 1188–1196.
- [93] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]. Eleventh Annual Conference of the International Speech Communication Association: volume 2, 2010: 1045–1048.
- [94] MNH A, HINTON G E. A scalable hierarchical distributed language model[C]. Proceedings of Advances in Neural Information Processing Systems, 2009: 1081–1088.
- [95] MNH A, TEH Y W. A fast and simple algorithm for training neural probabilistic language models[C], 2012.
- [96] MORIN F, BENGIO Y. Hierarchical probabilistic neural network language model[C]. Aistats: volume 5, 2005: 246–252.
- [97] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities[C]. Proceedings of the national academy of sciences: volume 79: 2554–2558.
- [98] ELMAN J L. Finding structure in time[J]. Cognitive science, 1990, 14 (2): 179–211.
- [99] FUNAHASHI K I, NAKAMURA Y. Approximation of dynamical systems by continuous time recurrent neural networks[J]. Neural networks, 1993, 6 (6): 801–806.
- [100] MIKOLOV T. Statistical language models based on neural networks[J]. Presentation at Google, Mountain View, 2nd April, 2012.
- [101] MIKOLOV T, ZWEIG G. Context dependent recurrent neural network language model[C]. Spoken Language Technology Workshop, 2013: 234–239.

- [102] MIKOLOV T, KOMBRINK S, BURGET L, et al. Extensions of recurrent neural network language model[C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2011: 5528–5531.
- [103] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE transactions on neural networks, 1994, 5 (2): 157–166.
- [104] CHENG J, DONG L, LAPATA M. Long short-term memory-networks for machine reading[J]. Proceedings of Conference on Empirical Methods in Natural Language Processing, 2016.
- [105] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. Proceedings of Advances in Neural Information Processing Systems, 2014.
- [106] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional lstm and other neural network architectures[J]. Neural Networks, 2005, 18 (5-6): 602–610.
- [107] GRAVES A, SCHMIDHUBER J. Offline handwriting recognition with multidimensional recurrent neural networks[C]. Proceedings of Advances in Neural Information Processing Systems, 2009: 545–552.
- [108] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28 (10): 2222–2232.
- [109] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures[C]. Proceedings of International Conference on Machine Learning, 2015: 2342–2350.
- [110] KALCHBRENNER N, DANIHELKA I, GRAVES A. Grid long short-term memory[J]. Computer Science, 2015.
- [111] OORD A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[C], 2016.
- [112] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C], 2015: 1556–1566.
- [113] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.
- [114] ERHAN D, SZEGEDY C, TOSHEV A, et al. Scalable object detection using deep neural

-
- networks[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2014: 2147–2154.
- [115] GIRSHICK R. Fast r-cnn[C]. Proceedings of IEEE International Conference on Computer Vision, 2015.
- [116] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]. Proceedings of European Conference on Computer Vision, 2014: 346–361.
- [117] OUYANG W, LUO P, ZENG X, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 2403–2412.
- [118] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Proceedings of Advances in Neural Information Processing Systems, 2015: 91–99.
- [119] SZEGEDY C, REED S, ERHAN D, et al. Scalable, high-quality object detection[J]. arXiv preprint arXiv:1412.1441, 2014.
- [120] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88 (2): 303–338.
- [121] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]. Proceedings of IEEE International Conference on Computer Vision, 2017: 2980–2988.
- [122] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 779–788.
- [123] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]. Proceedings of European Conference on Computer Vision, 2016: 21–37.
- [124] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104 (2): 154–171.
- [125] CORTES C, VAPNIK V. Support vector machine[J]. Machine learning, 1995, 20 (3): 273–297.

- [126]BERNARDI R, CAKICI R, ELLIOTT D, et al. Automatic description generation from images: A survey of models, datasets, and evaluation measures[J]. *Journal of Artificial Intelligence Research*, 2017, 55 (1): 409–442.
- [127]ELLIOTT D, KELLER F. Image description using visual dependency representations[C]. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013: 1292–1302.
- [128]FERRARO F, MOSTAFAZADEH N, TINGHAO, et al. A survey of current datasets for vision and language research[M]. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015: 207–213.
- [129]HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. *Journal of Artificial Intelligence Research*, 2013, 47: 853–899.
- [130]LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]. *Proceedings of European Conference on Computer Vision*, 2014: 740–755.
- [131]PLUMMER B, WANG L, CERVANTES C, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]. *Proceedings of IEEE International Conference on Computer Vision*, 2015: 2641–2649.
- [132]RASHTCHIAN C, YOUNG P, HODOSH M, et al. Collecting image annotations using amazon’s mechanical turk[C]. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010: 139–147.
- [133]YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. *Transactions of the Association for Computational Linguistics*, 2014, 2: 67–78.
- [134]ZITNICK C L, PARIKH D. Bringing semantics into focus using visual abstraction[C]. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3009–3016.
- [135]EVERINGHAM M, GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. *International Journal of Computer Vision*, 2010, 88 (2): 303–338.
- [136]REITER E, BELZ A. An investigation into the validity of some metrics for automatically evaluating natural language generation systems[J]. *Computational Linguistics*, 2009, 35 (4):

- 529–558.
- [137] LI S, KULKARNI G, BERG T L, et al. Composing simple image descriptions using web-scale n-grams[C]. Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011: 220–228.
- [138] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39 (4): 652–663.
- [139] ZHANG X, HE S, SONG X, et al. Keyword-driven image captioning via context-dependent bilateral lstm[C]. Proceedings of IEEE International Conference on Multimedia and Expo, 2017: 781–786.
- [140] PAPANENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation[C]. ACL, 2002: 311–318.
- [141] LAVIE A, AGARWAL A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments[C]. Second Workshop on Statistical Machine Translation, 2007: 228–231.
- [142] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation[C]. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015: 4566–4575.
- [143] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]. Text summarization branches out: Proceedings of the ACL-04 workshop: volume 8, 2004.
- [144] TIELEMAN T, HINTON G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSE: Neural networks for machine learning, 2012, 4 (2): 26–31.
- [145] LEBRET R, PINHEIRO P O, COLLOBERT R. Phrase-based image captioning[C]. Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015: 2085–2094.
- [146] ZHANG X, SONG X, LV X, et al. Rich image description based on regions[C]. Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, 2015: 1315–1318.

- [147] COLLOBERT R, KAVUKCUOGLU K, FARABET C. Torch7: A matlab-like environment for machine learning[C]. BigLearn, NIPS Workshop: EPFL-CONF-192376, 2011.
- [148] KARPATY A. Neuraltalk2[EB/OL]. 2016. <https://github.com/karpathy/neuraltalk2>.
- [149] BIRD S, KLEIN E, LOPER E. Natural language processing with python[M]. " O'Reilly Media, Inc.", 2009.
- [150] ELLIOTT D, KELLER F. Comparing automatic evaluation measures for image description[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers: volume 452, 2014: 457.
- [151] GONG Y, WANG L, HODOSH M, et al. Improving image-sentence embeddings using large weakly annotated photo collections[C]. Proceedings of European Conference on Computer Vision, 2014: 529–545.

致 谢

在中国科学院大学攻读博士学位的过程历尽艰辛又丰富多彩，我发自内心感激这段难忘的时光赋予我的欢乐、痛苦和成长。在这四年博士生涯里，作为中国科学院大学模式识别与智能系统开发实验室的一员，先后访问中科院计算所智能信息处理实验室和香港城市大学计算机系，整个过程不仅锻炼了我在科学研究上发现、思考和解决问题的能力，还增加了因多次访问交流带来的思想碰撞和生活历练，使得我自身获得了极大的成长。在毕业论文完成之际，由衷地感谢多年来曾经给予我无数帮助的老师、同学、朋友和家人。

首先要感谢我的导师焦建彬教授给予了我最大的支持，感谢他在我攻读博士学位期间对我的悉心指导和鼓励。焦老师为人谦和、治学严谨，是实验室所有学生的强大后盾，给我们提供了良好的科研环境和科研指导，在生活上对我们也关怀备至，使得我们可以毫无后顾之忧的尽情发挥自己的能力。焦老师开明开放、不拘一格，博士期间我曾两度去校外研究单位进行学习交流，焦老师都给予了有力的支持，使得我的博士生涯多了不一样的精彩。焦老师严谨的治学之风、对学生无私的付出、认真工作和健康生活的言传身教将使我受益终生。

感谢叶齐祥教授、韩振军副教授和秦飞副教授，他们在我的科学学习和理论研究中给予了耐心的指导。叶老师为人热情无私、勤勉敬业，对科研有着狂热的执着，经常有独到的见解。叶老师以身作则，教会了我如何进行科学研究，我博士期间发表的每篇论文都离不开叶老师的详细指导。叶老师也积极推动学生进行对外科研交流，为我们提供了很多宝贵的交流机会。韩老师和秦老师思维缜密、工作认真、经验丰富、平易近人，给我诸多帮助与启迪。他们渊博的专业知识、自强不息的学习精神、扎实的动手能力和对我的谆谆教导，让我受益匪浅。

感谢中科院计算所的蒋树强老师，蒋老师是我在计算所访问交流期间的指导老师，他渊博的专业知识、严谨的治学态度、精益求精的工作作风对我影响至深。在蒋老师课题组的一年时光是我人生的一个重要转折点，蒋老师教会了我如何进行发散思维和深度思考，督促着我一步步提高实验动手能力，对我整个博士生涯都影响深远。在计算所期间的科研工作使得我确定了自己的博士研

究方向，在之后几年内的工作也离不开蒋老师的远程指导，感谢蒋老师一直以来给予我的无私的支持和帮助。

感谢香港城市大学计算机系的杨庆雄老师、Rynson W.H. Lau 老师和何盛烽老师。在香港城市大学联合学习的一年半里，杨庆雄老师做为我在开始一年里的导师，无论是科研还是生活上都给了我极大的帮助，使得我能很快融入香港的科研和生活环境。杨老师科研能力极强，对科研工作要求极高，学生经常跟不上他大脑运转的速度，但他又有足够的耐心进行针对性的指导，使得我们都受益良多。Rynson W.H. Lau 老师是我在香港的最后半年的指导老师，他和蔼可亲、耐心细致，对每个学生都能做到每周至少一次的面谈指导，在频繁的沟通交流中我受到很多启发。在离开香港回到中国科学院大学继续学习的这一年多里，只要有需要，Rynson W.H. Lau 老师都会给予尽可能的帮助。在香港期间，何盛烽老师对我进行细致的科研指导，鼓励我勇于投顶级期刊和会议，精心帮我修改每一篇论文，教会了我很多写作技巧和科研方法，使我得到了直接的成长。我会永远记得三位老师给予的极大的宽容和无私的帮助。

我还要衷心感谢我所经历的三个实验室的所有成员，这几年来我们一起学习一起生活互帮互助，形成了家人一样的氛围。国科大的魏鹏旭、柯炜、万方、张天亮、李兆举等，计算所的宋新航、黎向阳等，香港城市大学的屈靛琮、宋奕兵、焦建波、张佳伟等，有幸与这些可爱可敬的同学一起度过我的博士生涯，并且结下了深厚的友谊，这是我人生中的最珍贵的财富。

感谢我的父母，感谢他们多年来一直给我最无私的帮助与鼓励，让我有勇气面对一切博士期间面临的挫折与困难，让我能够坚定初心，突破自我。

感谢参加开题、中期和毕业答辩的各位指导老师专家，你们丰富的经验和细致的指导，对论文方向和研究进度的指点给我的研究工作带来了巨大的帮助。

最后，再次向在学习、工作和生活中给予过自己关心、支持与鼓励的所有老师、同学、朋友们表示最诚挚的谢意！

张晓丹

2018 年 05 月

作者简介及攻读学位期间发表的学术论文与研究成果

作者简介:

2006年09月——2010年07月,在郑州大学电气工程学院获得学士学位。

2010年09月——2014年07月,在中国科学院大学工程科学学院获得硕士学位。

2014年09月——2018年06月,在中国科学院大学电子电气与通信工程学院攻读博士学位。

已发表(或正式接受)的学术论文:

一作论文:

[1] Xiaodan Zhang, Shengfeng He, Xinhang Song, Qixiang Ye, Jianbin Jiao, Rynson W. H. Lau. Image Captioning via Semantic Element Embedding[J]. Neurocomputing, 2018. (SCI, 已接收)

[2] Xiaodan Zhang, Shengfeng He, Xinhang Song, Pengxu Wei, Shuqiang Jiang, Qixiang Ye, Jianbin Jiao, Rynson W.H. Lau. Keyword-driven Image Captioning via Context-dependent Bilateral LSTM[C]. Proceedings of IEEE International Conference on Multimedia and Expo (ICME, oral), 2017:781--786. (EI)

[3] Xiaodan Zhang, Xinhang Song, Shuqiang Jiang, Qixiang Ye, Jianbin Jiao. Rich image description based on regions[C]. Proceedings of the 23rd ACM international conference on Multimedia (ACM MM), 2015:1315-1318. (EI)

合作论文:

[1] Shengfeng He, Jianbo Jiao, Xiaodan Zhang, Guoqiang Han, Rynson W.H. Lau. Delving into Salient Object Subitizing and Detection[C]. Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017:1059-1067. (EI)

