计算机视觉基础

中国石油大学（华东）
青岛软件学院、计算机科学与技术学院
宫文娟

计算机视觉基础

- 第一章 概论
- 第二章 基础知识
- 第三章 图像分类
- 第四章 图像语义分割
- 第五章 目标检测
- 第六章 识别
- 第七章 目标跟踪
- 第八章 多目视觉
- 第九章 视觉问答

第九章 视觉问答

- 9.1 联合嵌入方法
- 9.2 组合模型方法
- 9.3 注意力方法
- 9.4 知识增强方法
- 9.5 常用数据集
- 9.6 评估方法

概述

概念

- 视觉问答(VQA)是一项计算机视觉任务，它包含了计算机视觉中的许多子问题，例如：
  - 物体识别（图像中有什么?）
  - 目标检测（图像中有猫吗?）
  - 属性分类（猫是什么颜色?）
  - 场景分类（天气晴朗吗?）
  - 数量统计（图中有多少只猫?）等。
- 还有很多更复杂的问题，比如：
  - 物体之间的空间关系（猫和沙发之间是什么?）
  - 常识推理问题（女孩为什么哭?）等。

✓ 一个鲁邦的VQA系统必须能够解决广泛的经典计算机视觉任务，并且具有能够对图像进行推理的能力。

概 述

概念

➢ 视觉问答VQA有许多潜在的应用：

- 帮助盲人和视力受损的人，使他们能够在网上和现实世界中获得关于图像的信息。
  - ✓ 例如，图像描述生成/图题生成(Image Captioning)系统可以描述图像，然后用户可以使用视觉问答来查询图像，以获得对场景的更多了解。
- 可以作为一种自然的查询可视内容的方式来改进人机交互。视觉问答系统也可以用于图像检索，而不需要使用图像元数据或标记。
  - ✓ 例如，要查找所有在雨天拍摄的照片，我们可以简单地问"下雨了吗?"指向数据集中的所有图像。

9.1 联合嵌入方法

概念

- 在图题生成任务中，首次探索了图像与文本的联合嵌入。由于计算机视觉和自然语言处理两种深度学习方法的成功，人们得以在一个共同的特征空间中进行表示学习。
- 与图题生成的任务相比，在视觉问答需要对两种模式进行进一步的推理，在一个公共空间中的一个表示将降低学习的相互作用，并对问题和图像内容进行推理。

■ 在实际应用中，通常利用：

1. 卷积神经网络对目标识别进行预处理，从而获得图像表征，
2. 文本表示是通过在大型文本语料库上预先训练的词嵌入来获得的。
   - 词嵌入实际上是将词映射到一个能够通过距离反映语义相似性的空间中，然后将问题中每个单词的嵌入信息输入到递归神经网络中，处理变长序列以捕获句法模式。

9.1 联合嵌入方法

神经网络问答算法（Neural Image QA）

- Malinowski等人提出了一种名为神经图像问答的方法[1]，该方法使用了长短期记忆单元连接的递归神经网络(如图9-1)。

1.RNNs背后的动机是处理可变大小的输入(问题)和输出(答案)。

2.对于目标识别任务，图像特征是由经过预处理的卷积神经网络CNN生成的。

3.问题和图像特征都是一起发送给第一个"编码器"。它产生一个固定大小的特征向量，然后将其传递给第二个"解码器"。

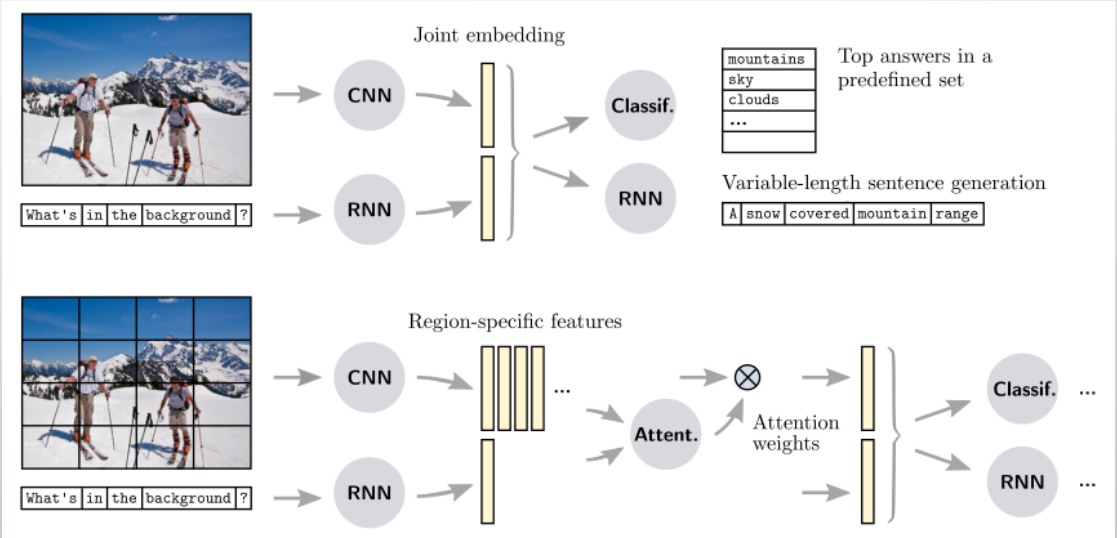


图9-1 神经网络问答模型架构图

[1]. M. Malinowski, M. Rohrbach, and M. Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

9.2 组合模型方法

概念

- 这种方法涉及到连接不同的模块，这些模块是为特定的功能设计的，比如记忆或特定类型的推理。
- 采用模块化结构的优势是可以更好地监督和管理。
  - 它有助于转移学习，因为同一个模块可以在不同的整体架构和任务中使用和训练。
  - 它允许使用“深度监督”，即优化目标时需要依赖于内部模块输出。在注意力模型和与知识库的连接的模型中讨论的其他模型也属于模块化架构的范畴。
- 两个主要贡献在模块方面的特定模型：

1. 神经模块网络(Neural Module Networks,NMN)
2. 动态记忆网络(Dynamic Memory Networks, DMN)

9.2 组合模型方法

9.2.1 神经模块网络(NMNs)

- 神经模块网络由Andreas等人引入并进行了扩展[2,3]。
- 它们是专门为视觉问答设计的，目的是研究"问题"的语言组成结构，而"问题"的复杂程度往往差别很大。例如：
  - "Is this a truck ?"只需要从图像中检索一条信息就可以得到答案，
  - "How many objects are to the left of the toaster ?"需要多个处理步骤，如目标识别和计数。
- 神经模块网络反映了一个网络中问题的复杂性，这个网络是针对问题的每个实例动态设置的。

■ 神经模块网络的一个重要贡献是将这种逻辑推理应用于连续的视觉特征，而不是离散的或逻辑重复的视觉特征。

[2]. Andreas J , Rohrbach M , Darrell T , et al. Learning to compose neural networks for question answering[J]. Computation and language (cs.CL), 2016.
[3]. Andreas J , Rohrbach M , Darrell T, et al. Neural module networks[C].// 2016 IEEE conference on computer vision and pattern recognition(CVPR), June 26 - July 1, Nevada, Las Vegas, 2016, 39-48.

9.2 组合模型方法

9.2.1 神经模块网络(NMNs)

- 该方法使用基于自然语言处理社区中著名的工具对查询进行语义解析。解析树被转换成来自预定义集的模块集，然后这些模块集一起用于回答问题。

➢ 对问题的解析导致对在关注空间中运行的模块进行组装：

1. 两个参与模块用于定位红色的形状和圆圈
2. 将注意力转移到圆圈上面
3. 合并计算它们的交集
4. 测量检查最终的注意力并确定它是非空的。

- 对每个问题执行的计算是不同的，在测试时使用与训练时不同的问题实例。

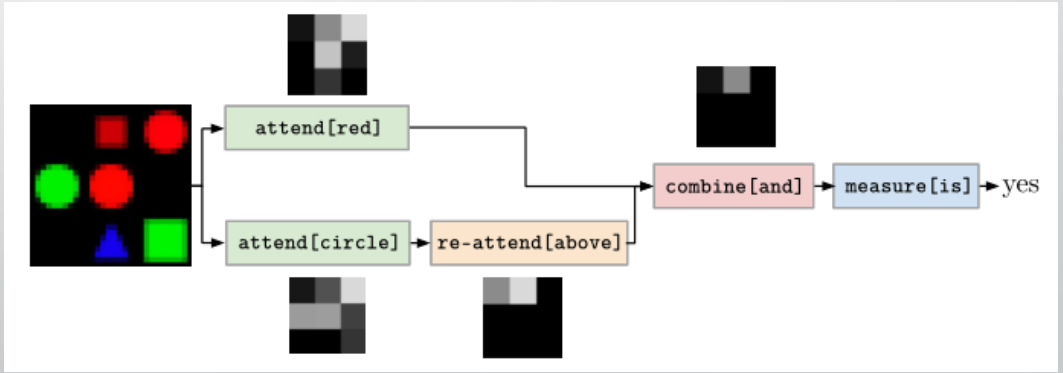


图9-2 神经模块网络方法流程

9.2 组合模型方法

9.2.1 神经模块网络(NMNs)

- 模块的输入和输出可以是三种类型：
  - 图像特征
  - 图像上的注意区域
  - 标签(分类决策)
- 一组可能的模块是预先定义的，每个模块根据其输入和输出的类型，但是通过对特定问题实例的端到端培训，将需要了解它们的确切行为。因此，除了图像、问题和答案三元组外训练不需要额外的监督。

9.2 组合模型方法

9.2.2 动态记忆网络(DMN)

- 动态记忆网络是具有特定模块化结构的神经网络。
- 动态记忆网络属于记忆增强网络，它能够对输入的内部表示进行读写操作。这种机制与注意力机制类似，旨在通过在几次传递中对数据的多个部分之间的交互进行建模，从而解决复杂逻辑推理的任务。

9.2 组合模型方法

9.2.2 动态记忆网络(DMN)

- 动态记忆网络由输入模块、问题模块、情景记忆模块和回答模块4个主要模块组成：

1. 输入模块将输入数据转换成一组称为"事实"的向量。它的实现取决于输入数据的类型；
2. 问题模块使用门控递归单元计算"问题"的向量表示；
3. 事件记忆模块用于检索回答问题所需的"事实"。它包含一个选择相关事实的注意力机制和一个更新机制。
4. 回答模块使用记忆的最终状态和问题相结合，通过对单个单词进行多重分类来预测输出，对长句子的数据集预测时同时使用。

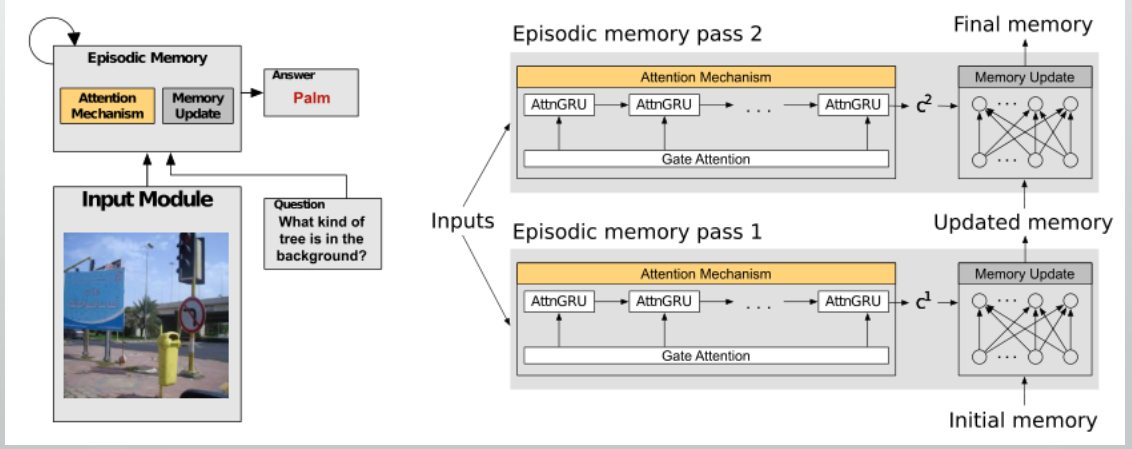


图9-3 用于视觉问答的动态记忆网络

9.3 注意力方法

注意力方法

- 前面的的大多数模型的一个限制是使用图像的全局特征来表示视觉输入，这可能会向预测阶段提供不相关或有噪声的信息。
- 注意机制的目的是通过局部图像特征来解决这个问题，并允许模型对不同区域的特征赋予不同的重要性。
- Xu等人在图像字幕图题生成的背景下提出了对视觉任务注意力的早期应用[4]。模型的注意力部分用于识别图像中的显著区域，通过进一步处理，最终将图题生成集中在这些显著区域。
- 这一概念很容易转化为视觉问答任务，即只关注与问题相关的区域信息。注意力过程在推理过程中执行一个明确的附加步骤，该步骤在执行进一步计算之前确定"从哪里看"。

[4]. Xu K , Ba J , Kiros R , et al. Show, attend and tell: neural image caption generation with visual attention[J]. Computer science, 2015:2048-2057.

9.4 知识增强方法

知识增强方法

- 视觉问答任务包括理解图像的内容，但通常需要事先定义的非视觉信息，这些信息可以是“常识”，也可以是特定主题的知识，甚至更广泛的知识。例如：

➢ 要回答“How many mammals appear in this image ?”这个问题，你必须理解“mammals”这个词，并知道哪些动物属于这一类。

◘ 从这个问题中可以发现联合嵌入方法的两个主要缺点：

1. 它们只能获取训练集中的知识，而无论怎样扩大数据集，也很难完全覆盖现实世界。
2. 在这种方法中训练的神经网络能力有限，我们希望学习的信息量远远超过了这种能力。

9.4 知识增强方法

知识增强方法

- 可以从实际存储的数据或知识中进行推理，近来涌现出大量致力于结构化知识表达的研究。大规模知识库发展迅速，出现了如DBpedia、Freebase、YAGO、OpenIE、NELL、WebChild和ConceptNet等知识库，这些知识库以可读的方式存储常识和事实知识。
- 每一条知识，被称为一个事实，通常表示为一个三元组(arg1,rel,arg2)，其中arg1和arg2表示两个概念，rel表示它们之间的关系。

9.5 常用数据集

常用数据集

- 许多数据集已经被特别提出用于视觉问答任务中的搜索。
  - 它们至少包含一个图像、一个问题及其正确答案构成的三元组。
  - 有时还提供附加说明，如图像说明、支持答案的图像区域或多个候选答案。
- 区别不同数据集的一个主要特征是其图像的类型，图像的类型可以大致分为：

1. 自然类
   - 使用最广泛的数据集如DAQUAR、COCO-QA 和VQA-real 都使用自然类图像。
2. 剪贴类
3. 合成类

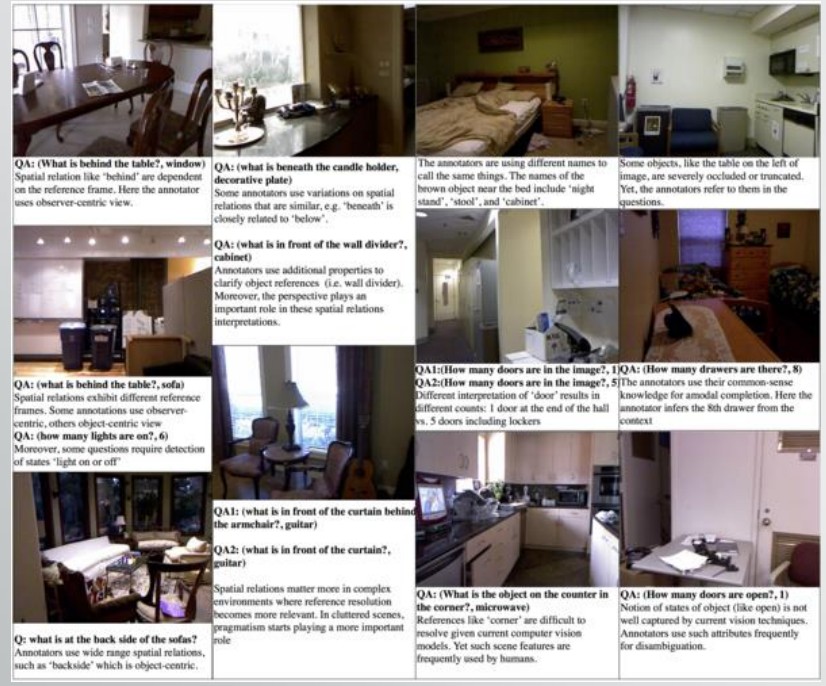
- 数据集之间的第二个关键区别是问答格式，问答格式包括：

1. 开放式没有预定义的答案集，是较为常见的。
2. 多项选择格式每个选项提供了有限的可能答案集。

➢ VQA-real和Visual7W数据集都允许使用开放式或多项选择式进行评估。

- 这两种设置的结果不能进行比较，开放式设置被认为更具挑战性，同时更难进行定量评估。

DAQUAR (DAtaset for QUestion Answering on Real-world images)



- DAQUAR作为基准设计的第一个视觉问答数据集，用于真实世界视觉问题回答数据集：
- DAQUAR的图像分为795张训练图像和654张测试图像。
- 使用来自NYU-Depth v2数据集的图像构建的，该数据集包含了：
  - 1,449张室内场景的RGBD图像，并带有注释的语义分段。
- 收集了两种类型的问题/答案对。
  1. 综合问题/答案是使用8个预定义模板和NYU数据集自动生成的。
  2. 人类的问题/答案来自5个注释者，他们主要针对基本的颜色、数字、物体(894个类别)和它们的集合。
- 总共收集了12,468对问题/答案，其中6,794对用于训练，5,674对用于测试。

■ 大规模的DAQUAR是开发和训练具有深度神经网络的视觉问答VQA早期方法的关键。

□ DAQUAR的主要缺点是答案限制在预定义的16种颜色和894个对象类别。数据集也表现出强烈的偏见，表明人类倾向于关注一些突出的对象，如桌椅。

9.6 评估方法

评估方法

- 视觉问答被设定为一个开放性的任务，即算法生成一个字符串来回答一个问题，或者是一个选择题。
  - 对于多项选择题，通常用简单正确率来评估，如果算法做出了正确的选择，它就能得到正确的答案。
  - 对于开放式的视觉问答VQA，也可以使用简单的准确性。
- 一个算法的预测答案字符串必须与真实答案完全匹配。然而，这种准确性可能过于严格。例如：
  - 如果问题是“What animals are in the photo?”系统输出“dog”而不是正确的标签“dogs”，此时受到的惩罚和输出“zebra”一样严重。
  - 问题也可能有多个正确答案，例如，“What is in the tree?”可能会将“bald eagle”列为正确的真值答案，因此输出“eagle”或“bird”的系统受到的惩罚一样多。

9.6 评估方法

Wu-Palmer Similarity (WUPS)

- WUPS试图通过语义上的差异来衡量一个预测的答案与真实值之间的差异有多大。给定一个真实答案和一个问题的预测答案，WUPS将根据它们之间的相似性在0到1之间分配一个值。它通过查找两个语义之间的最小公共集，并根据需要遍历语义树多远才能找到公共集来打分。
- 采用WUPS进行评估时，语义上相似但不相同的单词受到的惩罚相对较少。

➢ 在前面的例子中，“bald eagle”和“eagle”的相似度是0.96，而“bald eagle”和“bird”的相似度是0.88。

◘ 然而，即使是比较遥远的概念，比如“raven”和“writing desk”，WUPS得分也比较高，有0.4分。

■ 为了解决这个问题，Malinowski建议将WUPS分数降至阈值，即低于阈值的分数将按比例降低一个因子，即设置阈值为0.9，比例系数为0.1。除了简单的准确性之外，这种改进的WUPS度量是用于评估DAQUAR和COCO-QA性能的标准度量。

9.6 评估方法

给一个问题多个独立答案

- 另一种依赖语义相似度度量的方法是为每个问题收集多个独立的真实答案，这是视觉问答数据集DAQUAR-consensus所做的。
- 对于DAQUAR-consensus，平均每个问题收集5个人类注释的真实答案。
- 数据集的创建者提出了两种使用这些答案的方法，他们称之为：

1. 平均共识，最终的分数倾向于注释者提供的答案。
2. 最小共识，答案需要与至少一个注释器一致。