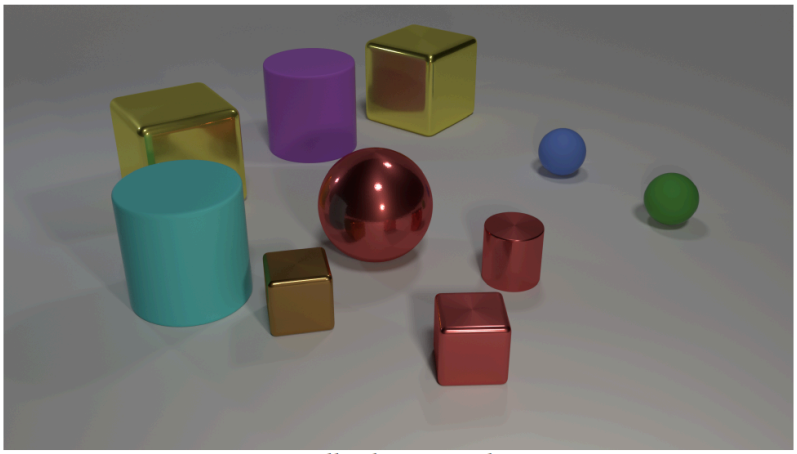
A Survey of Visual Question Answering

摘要

Visual Question Answering（视觉问答，以下简称 VQA）是自 2017 年以来计算机视觉分类下的新兴的话题。它是视觉和语言的桥梁，给定一个图像和自然语言的问题，它需要推理图像的视觉元素和一般知识来推断正确的答案。深度学习模型不仅需要理解图片（或视频）内容，还需要以自然语言的形式回答。本文着重介绍 2019 年最新发表的几篇论文中所使用的处理方法。

介绍

图 1 为 VQA 的一个典型例子。VQA 的数据集分为图片（或视频）、问题以及需要生成的回答。



Q: How many small spheres are there? **A:** 2.
Q: Does the metal sphere have the same color as the metal cylinder? **A:** Yes

图 1 VQA 经典例子

VQA 的步骤大致为：根据图像或视频提取特征（图像使用 Resnet 等模型处理，视频一般需要分帧后按图像的方式处理）、根据问题提取文本特征（LSTM 等模型

处理）。特征提取后进行融合，使用注意力机制等方式，最终获得答案。
接下来的章节将介绍 2019 年最新发表的几篇论文的新型的处理方式。

融入先验知识的模型[1]

随着近来深度神经网络的发展，注意力机制被广泛应用于 VQA 方法中，它更加关注更有意义的图像区域以回答问题。因此，对于问题的图像的视觉内容学习有效的注意力分布是一个主要的挑战。 VQA 中现有的基于注意力的方法可以分为基于自由形式的方法和基于检测的方法。在基于自由形式的方法中，图像被均匀地划分为许多区域，并且在这些区域中分配注意力。虽然它可以自由地将注意力映射到任何大小的区域，但是对有意义的对象分配适当的注意力是无效的。因此，它无法有效地回避有关前景对象的问题。另一方面，基于检测的方法旨在学习在图像的预先指定的检测框上的注意力分布。但是，回答有关更抽象概念的问题是无效的。
图 2 是 ALSA 模型框架。可以看到，它结合了基于自由形式的方法和机遇检测的方法，将两者获得的特征传入鉴别模块和分类起，最终获得回答。

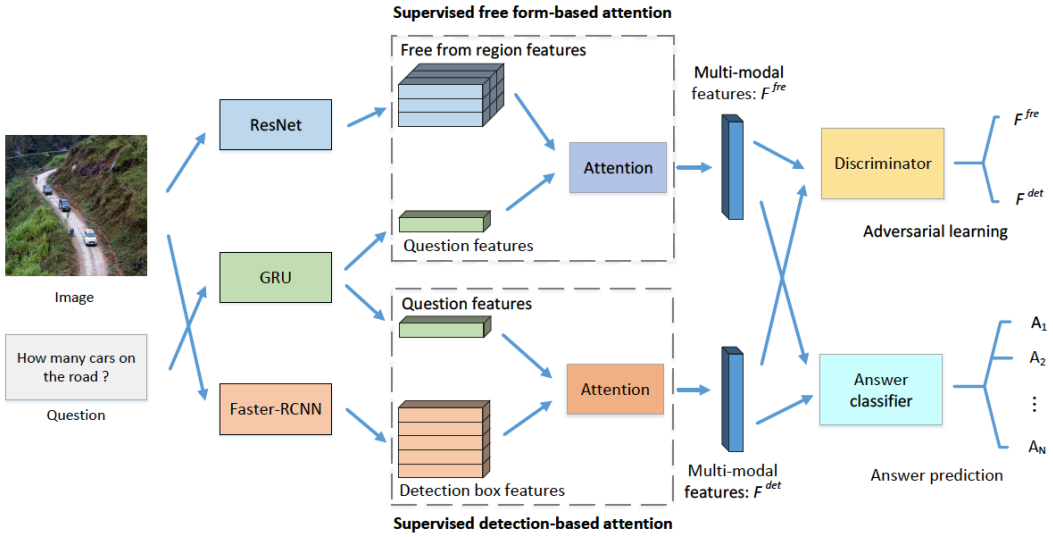


图 2 ALSA 模型框架

具有语义空间映射的推理注意力模型[2]

大多数现有方法将关节嵌入馈送到分类器中，用于在训练数据集中的大量标记答案上进行答案预测。然而，这些方法无法正确回答问题 - 图像对，这些问题 - 图像对寻求未标记并存在于训练数据集之外的新的回避者。此外，这些方法中使用的注意力模型对于捕获不同单词和图像区域之间的不同重要性是无效的。为了解决这些问题，具有语义空间映射的推理注意力模型（IASSM）被提出。提出了一种零镜头学习方法来预测未标记的答案。为了具有学习新答案的能力，构建了由标记和未标记答案共享的语义空间，其中可以映射问题图像对的联合嵌入并将其聚集在答案样本周围。为了更有效地学习问题 - 图像对的联合嵌入，设计了一种推理注意模型来模拟人类注意力的学习过程。

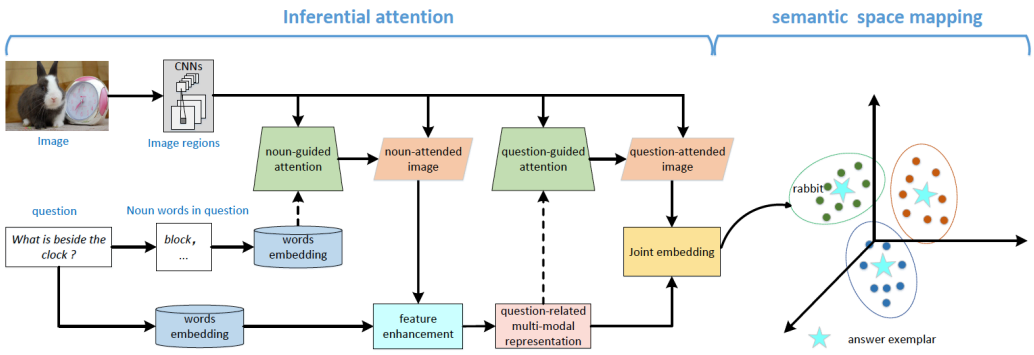


图 3 IASSM 模型框架

如图 3 所示，IASSM 模型包含两个组件。推理注意力网络旨在学习有效的多模态联合嵌入问题和图像。提出语义空间映射以推断问题-图像对的标记或非标记答案。

问题引导模块化路由网络模型[3]

VQA 面临两大挑战：

1）如何更好地融合视觉和文本形式

2）如何使 VQA 模型具有回答更复杂问题的推理能力

此文通过提出新颖的问题引导模块化路由网络（QGMRN）来解决这两个挑战。

QGMRN 可以融合多个语义级别的视觉和文本模式，这使得融合以细粒度的方式发生，它还可以通过在没有额外监督信息或先验知识的情况下在通用模块之间进行路由来学习推理。 拟议的 QGMRN 由三个子网组成：可视网络，文本网络和路由网络。 路由选择网络根据由文本网络生成的问题特征激活的路径选择性地执行可视网络中的每个模块。

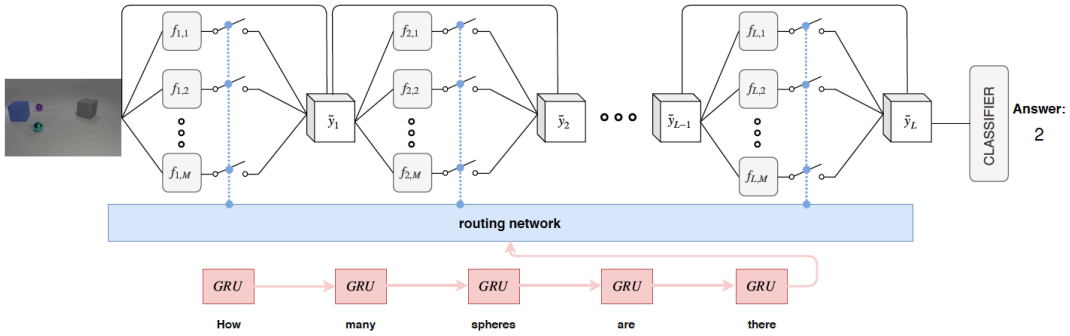


图 4 QGMRN 模型架构

可视网络是由 L*M 模块组成的复合体，每个模块后跟一个受控的 “开关” 通过路由网络。“交换机” 的状态称为路由路径。文本网络提出问题并生成问题特征。路由网络采用从文本网络发送的问题特征来生成路由路径。

总结

以上三者体现了 2019 年 VQA 的发展方向。总体而言，他们都更好的融合了视觉和文本的形式，使新模型拥有回答更为复杂或未知的问题的能力。

参考

[1] ALSA: Adversarial Learning of Supervised Attentions for Visual Question Answering.

[2] New Answers Learning and Inferential Attention for Visual Question Answering.

[3] Question Guided Modular Routing Networks for Visual Question Answering.