视觉问答文献综述

研究现状：

视觉问答明显是一个结合了计算机视觉与自然语言处理两个研究方向的任务，计算机视觉负责“看”——对图像信息进行感知、识别、理解，而自然语言处理负责“阅读”——对单词、句子进行语义分析、推理。在历史上，这两个方向是分开发展的，随着近年来两个方向日益渐佳的发展和文本视觉信息的共同爆炸性增长，也推动着二者领域的共同进步，视觉问答的概念也被的概念由德国马克斯·普朗克计算机科学研究所的 Mateusz Malinowski 和 Mario Fritz 在 2014 年提出，并受到了人们的关注，成为一个热门的研究方向。理论上求解视觉问答问题，是探索人类感知世界的重要一环。视觉问答当中的理论和方法也可以扩展到其他多模态领域，对于多模态人工智能方面有着很高的研究价值。实际上，视觉问答系统的应用潜力也是巨大的，可用于各种人类感知方面的场景探测回答，最直接的应用便是是为盲人提供服务。

问题描述:

视觉问答属于多模态领域中的一个备受关注的研究方向，不仅是自然语言处理还是计算机视觉中都日益重视的一个研究方向。从目前来看，视觉问答最直接的应用领域应该是为盲人提供服务。视觉问答的任务可以描述为：给定一张图像（或一段视频）和一个关于图像的自然语言问题，通过提取多种模态信息后进行表示、融合后，再经过一定的推理，给出正确的自然语言答案的过程，如图 1 所示。视觉问答与传统的问答系统相比，输出不仅是自然语言方面的信息还包括了视觉方面的信息，这不仅涉及多学科问题，如：计算机视觉中的对象识别与定位、场景识别、行为识别、属性识别，自然语言处理中的分词、单词与句子的编码与语义分析以及内容推理、空间关系的推理、常识推理等，还涉及到多模态问题，如：多模态的表示、对齐及融合这三个多模态中的极具挑战性的问题。本文主要按照上述多模态特征处理过程的三个方面来归纳文献内容。

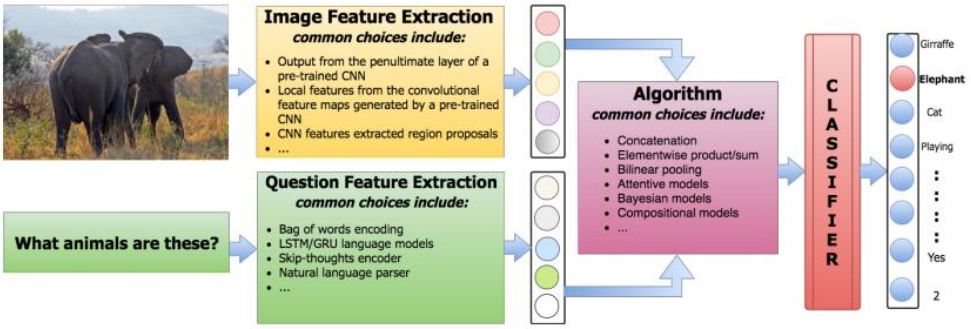

图 1 VQA 基本框架图

核心技术及创新点：

如何用统一的计算机格式来表示不同模态的信息并在融合多模态信息时尽可能大的做到信息互补而减少信息的冗余一直是多模态领域中不小的一个挑战。好的多模态信息表示对算法性能的提升是显而易见的，可以说多模态信息的表示是一个多模态模型的主干部分，占据着十分重要的地位。对于视觉问答中涉及的两种模态——视觉模态和文本模态的各自的单模态表示有：视觉单模态表示的大多数算法使用在 ImageNet 上预先训练的卷积神经网络 (CNN)，常见的例子是 VGGNet，ResNet、GoogLeNet 和 Faster-RCNN；目前已经探索了很多种文本模态表示方法，包括词袋 (BOW)、长短期记忆编码器 (LSTM)，门控循环单位 (GRU 和 skipthought vectors)。目前，目前多模态中的表示可按融合方式分为两种：联合表示和协调表示，如图 2 所示。联合表示是将不同模态的信息投射组合到相同的表示空间，而协调表示是将不同模态的信息单独处理，使其达到一定的相似性约束，也就是我们所说的协调空间。然而在视觉问答中，以协同表示方法来表征多模态信息并不多见，Junwei Liang 等

人在[16]中将余弦相似度和欧几里得距离结合起来来计算、比较不同模式(视觉模式和文本模式)之间的相似性。而联合表示方法在模型中的应用则比较广泛。

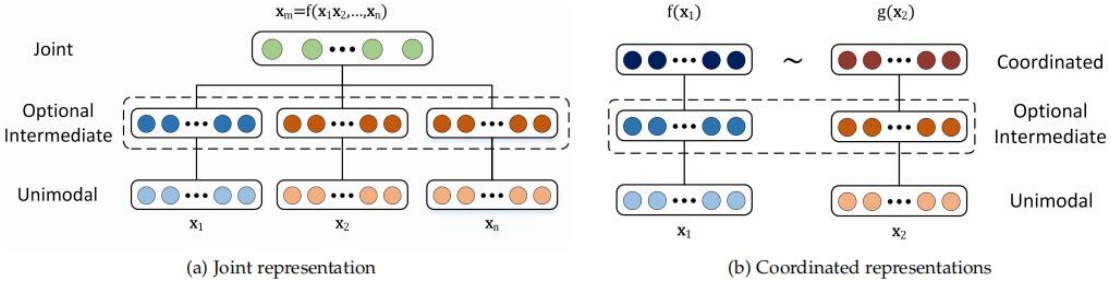

图 2 联合表示与协调表示结构图

Fei Liu 等人[2]提出了一种建模密集相互作用的密集连接注意流框架（DCAF），他们使用 Faster-RCNN 提取初始图像特征。对于问题特征，用 Glove 对每个单词编码，放入一个在每一层的输出进行剩余连接和批归一化处理的两层 GRU 中。在注意力机制中融合视觉和问题特征得到注意力权重用于更新注意视觉特征和注意问题特征，再使用元素积来联合表示多模态特征向量，同样使用 Faster-RCNN 提取初始图像特征，使用 GRU 提取文本特征，再在注意力机制中进行联合表示的还有[5][6]。

Fei Liu 等人[1]提出了一种建模密集相互作用的密集连接注意流框架（DCAF），他们使用 Faster RCNN 提取初始图像特征。对于问题特征，用 Glove 对每个单词编码，放入一个在每一层的输出进行剩余连接和批归一化处理的两层 GRU 中。在注意力机制中融合视觉和问题特征得到注意力权重用于更新注意视觉特征和注意问题特征，再使用元素积来联合表示多模态特征向量。

Badri Patro and Vinay P. Namboodiri 采用基于范例的方法，通过提供不同的注意力来改进 VQA 方法。他们使用 CNN 来处理图像特征，LSTM 处理问题特征，使用加权函数将图像和问题特征联合表示为注意力权重向量[3]。

Pan Lu 等人在[8]中使用 ResNet 提取图片特征，GRU 提取问题特征。Pan Lu 等人在视觉注意过程中使用多模态低秩双线性池化方法（MLB）来将问题和图像两种模态联合表示成上下文向量，经线性操作和 softmax 函数处理得到代表图像区域和输入问题的语义相关性的注意力权重，并以该注意力权重计算表示所有图像区域的加权和表示。最后以元素积方式结合有上下文意识的视觉特征和问题特征来得到最终的视觉表示。将有上下文意识的视觉嵌入和事实嵌入以 MLB 方法联合表示成联合上下文表示，然后计算注意力权重向量，这两者的乘积和用于表示候选事实的最终注意事实表示。将表示候选事实的最终注意事实表示和最终的视觉表示以元素积方式结合经过线性转换、非线性转换激活函数得到联合知识表示。

Zhou Zhao 等在[9]中使用 LSTM 提取问题答案的联合表征，并在其中应用了元素加方式的问题答案对机制。使用二维卷积网络方法提取帧级特征，再在此基础上与有上下文意识的问题表征形成空间注意帧表征。用三维卷积网络方法提取段级特征，之后与有上下文意识的问题表征形成空间注意段表征。以元素积的方式融合空间注意帧表征和空间注意段表征形成基于多流层次关注上下文网络的问题视频表示。

Yangyang Guo 等人在[12]中以 RNN 提取问题特征，利用 CNN 进行预训练并提取图像特征，以注意力机制融合问题特征和图像特征联合表示成最终的视觉特征。

Chen Zhu 等人在[14]中，从 GRU 最后的时间步骤中提取问题特征，从 CNN 的最后一层卷积层中提取图像特征，并使用注意力机制联合表示上下文特征（与问题相联系的视觉特征）。

Ramakrishna Vedantam 提出了一种新的概率神经网络模型，它具有作为潜在的随机变量的符号功能程序[11]。这是一种概率图谱的联合表示方法，用概率图模型来表示单模和多模

数据，是通过使用潜在随机变量来构造表示的常用方法。

多模态对齐

在视觉问答中，通常是通过注意力机制，去找到图像和问题的相关性来赋予与问题相关的图像区域更大的权重来实现多模态特征的对齐。

Remi Cadene 等人[1]提出了一种学习真实图像端到端推理的多模态关系网络 MuRel, 在 MuRel network 中多次使用 MuRel Cell,对齐问题特征和图像区域特征。

Fei Liu 等人[2]提出了一种建模密集相互作用的密集连接注意流框架（DCAF），通过文本自我关注与视觉空间关注对齐文本和视觉信息。

Badri Patro and Vinay P. Namboodiri 采用 Attention 差别的注意力网络（DAN）和 loss 函数对齐目标图片、目标问题、正例图片、反例图片[3]。

Chenyou Fan 等[4]设计了一个视觉记忆网络层，将相关的视觉内容与关键问句对齐来同时处理视觉记忆和问题记忆,将视频的外观和运动特征和问题特征在时间上进行了对齐。

Jingkuan Song 等人[6]在从像素到对象：用于回答视觉问题的立体视觉注意方法中，采取了基于对象区域的空间关注对齐了局部区域和问题特征。

Peng Gao 等人[7]提出的动态融合模式内和模式间注意流模块可以多次堆叠以通过在字和区域之间的信息流，以迭代地建模用于视觉问题应答的潜在对准。

Pan Lu 等人[8]提出了一种自由区域和多模态乘性特征嵌入探测的协同注意力机制方法，联合了如下两个注意力模型来实现问题和图像的特征对齐：学习使用多模态乘性特征的自由形式图像区域注意力嵌入视觉特征和探测盒子注意力嵌入视觉特征。

Zhou Zhao 等在[9]中通过 spatial attention 和 temporal attention 在空间和时间上对齐多模态信息，在[10]中使用帧级注意力模块对齐问题表示和图像帧。

Yangyang Guo 等人在[12]中提出的 VQA 模型利用问题特征通过多层感知器(MLP)网络或 CNN 对每个图像区域进行关注，得到每个图像区域的注意权值，用于问题与图像区域的对齐操作。

Idan Schwartz 等人在[13]中使用了三阶注意力模块：单元注意、成对注意、三元注意用于问题、答案、图像的特征对齐。

多模态融合

目前在视觉问答领域中多模态信息融合通常有以下几种组合方式：简单拼接、逐元素加或乘、向量外积和双线性等。

Remi Cadene 等[1]通过一个有效的双线性融合模块融合了问题和区域特征向量，提供了局部多模态特征。

Badri Patro and Vinay P. Namboodiri[3]通过加权 softmax 函数融合 image 和 question 向量。

Yuetan Lin 等人[5]以区域和问题特征融合并经 softmax 函数得到注意权重，又以元素积方式融合视觉特征和问题表征得到二者的联合表示，[7，8，9，15]同样是以元素积方式融合多模态表示。

当然也有通过提出各种 attention 模块进行多模态信息融合的，例如：[6，14，15]。

总结：

上文将视觉问答方面的文献从多模态表示、对齐、融合三个方面分别归纳总结，这种分类方法有利于更深层次的理解视觉问答的过程以及扩展到对其他多模态领域的理解。

参考文献

[1] Remi Cadene, Hedi Ben-younes, Matthieu Cord, Nicolas Thome. MUREL: Multimodal Relational Reasoning for Visual Question Answering. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[2] Fei Liu, Jing Liu , Zhiwei Fang , Richang Hong , Hanqing Lu. Densely Connected Attention Flow for Visual Question Answering. IJCAI(2019).

[3] Badri Patro and Vinay P. Namboodiri. Differential Attention for Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[4] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang,Heng Huang. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[5] Yuetan Lin, Zhangyang Pang, Donghui Wang, Yueting Zhuang. Feature Enhancement in Attention for Visual Question Answering. IJCAI(2018).

[6] Jingkuan Song, Pengpeng Zeng, Lianli Gao and Heng Tao Shen. From Pixels to Objects: Cubic Visual Attention for Visual Question Answering. IJCAI(2018).

[7] Peng Gao, Zhengkai Jiang, Haoxuan You,Pan Lu, Steven Hoi , Xiaogang Wang , Hongsheng Li. Dynamic Fusion with Intra- and Inter-modality Attention. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[8] Pan Lu，Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, Jianyong Wang. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. KDD(2018).

[9] Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, Shiliang Pu. Multi-Turn Video Question Answering via Multi-Stream Hierarchical Attention Context Network. IJCAI(2018).

[10] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, Yueting Zhuang. Open-Ended Long-form Video Question Answering via Adaptive Hierarchical Reinforced Networks. IJCAI(2018).

[11] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, Devi Parikh. Probabilistic Neural-symbolic Models for Interpretable Visual Question Answering. ICML(2019).

[12] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, Mohan Kankanhalli. Quantifying and Alleviating the Language Prior Problem in Visual - estion Answering. SIGIR(2019).

[13] Idan Schwartz, Alexander G. Schwing, Tamir Hazan. High-Order Attention Models for Visual Question Answering. NIPS (2017).

[14] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, Yi Ma. Structured Attentions for Visual Question Answering. ICCV(2017).

[15] Chenfei Wu, Jinlai Liu, Xiaojie Wang, Xuan Dong. Chain of Reasoning for Visual Question Answering. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).

[16] Junwei Liang,Lu Jiang,Liangliang Cao,Li-Jia Li, Alexander Hauptmann. Focal Visual-Text Attention for Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.