多模态-VQA

世界存在着万物，可以通过视觉，听觉，触觉，嗅觉等体现。一般来说，模态是指某物发生或经历的方式。大多数人把模态这个词与代表我们主要沟通和感知渠道的感觉方式联系起来，如视觉和触觉。因此，当一个研究问题或数据集包含多个这样的模式时，它就具有多模态的特征。

视觉问题解答（VQA）是多模态机器学习研究领域的一个典型应用，由于深度学习在低级和中级任务（例如图像分割或物体识别）上成功，激发了研究人员对完成将视觉与语言和高级推理相结合的更复杂任务的信心。

视觉问答是多模态的实际应用之一，而多模态数据的异构性带来了不少问题，如何解决图像或视频和文本数据之间的互补和冗余，以达到汇总数据更佳的表示；如何衡量图像和文本信息之间的，或视频与文本信息之间的相似性，以达到确定多个模态的元素之间的关系；如何将图像、视频与文本信息有效连接，进行进一步的预测。

本文着重于视觉问答中图像与文本之间的表示，对齐，融合问题。并且了解国内外研究现状，其运用到的核心技术及创新点。

VQA 的解决方式通常是用监督学习来训练深度神经网络，该网络将输入的图像和问题经过层层处理输出为候选答案的相对评分。其主要思想是学习单模态，视觉和文本的汇总表示。

首先，图像和问题各自作为单独的模态，通常经过以下处理以获得其矢量表示形式。

使用 ImageNet 预训练的 ResNet-152 CNN 或 ResNet-100 CNN、VGG-19 得到图像表示，使用如 RNN、单层 LSTM、双向 LSTM 或 GRU 来编码问题和答案，而对于视频问答，通常 VGG 和 C3D 组合得到视频的外观和运动特征。

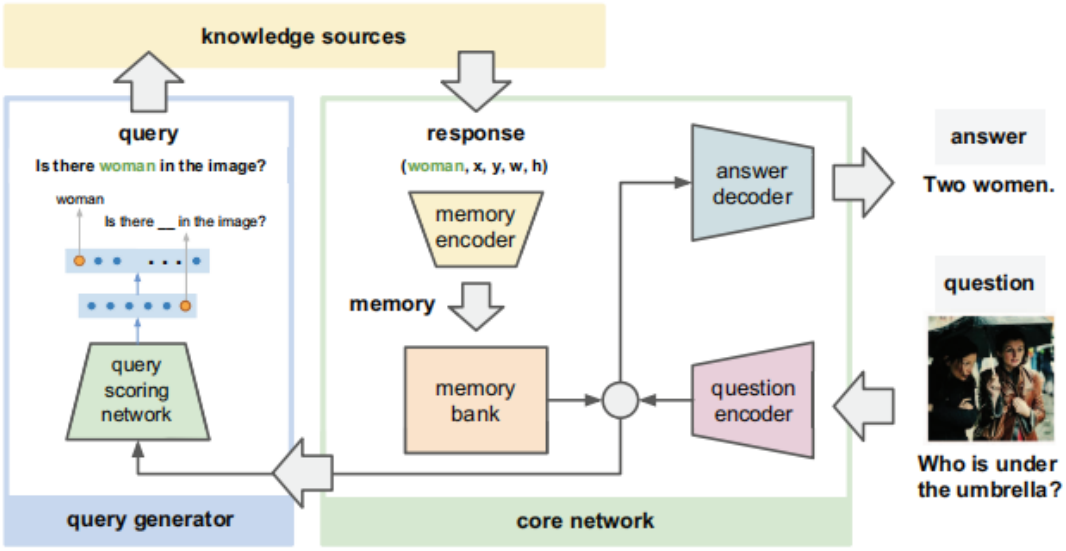
其次，将视觉和文本的汇总表示方式大致分为两种，联合表示和协同表示。

联合表示将视觉信号，文本信号简单组合，用某种线性变换投影到同一维度，也表示为同一空间。

用神经网络来处理双模态联合表示，这种方法十分普遍。

Daniel Gordon[1]等人提出的模型 Hierarchical Interactive Memory Network (HIMN)，Kuo-Hao Zeng[2]等人提出的 4 个扩展模型 E-E2EM ,E-VQA,E-SA,E-SS，Siddha Ganju[39]等人提出的 iBOWIMG-2x 模型，Manoj Acharya[3]等人提出的 RCN 模型，也采取类似的联合表示形成嵌入。

Yuke Zhu[4]等人提出动态的可迭代的 VQA 模型，采用问题和答案的词嵌入平均值进行联合表示其模型的核心网络通过输入问题，预测答案并维持其内部存储库，同时通过查询来反复获得新证据，整个核心网络可以端到端的方式进行联合训练。



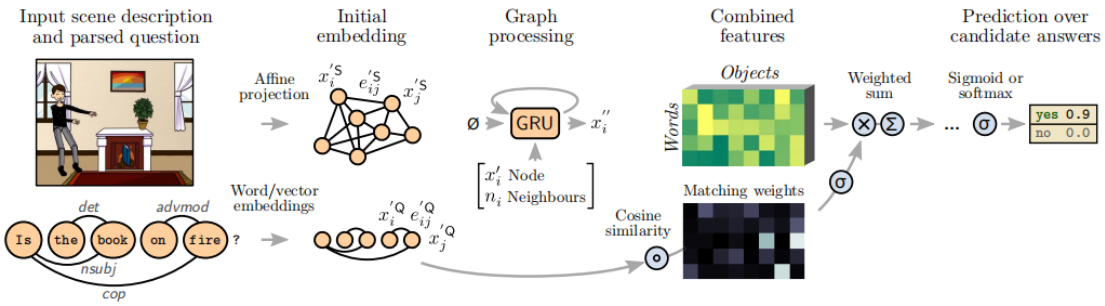
在胶囊网络的基础上，Yiyi Zhou[5]等人提出了解决 VQA 的胶囊注意力模型，视觉特征和文本特征会馈送到 CapsAtt，其输出胶囊用作两种模态的联合表示。

Robik Shresth[6]等人提出 RAMEN(Recurrent Aggregation of Multimodal Embeddings Network)模型解决 VQA，通过将空间定位的视觉特征与问题特征串联起来来进行早期融合。为了学习视觉和文本功能之间的相互关系，把视觉、问题特征的连接通过共享网络传递，从而产生空间局部化的双模态嵌入。

还有另外一种方式来构造双模态联合表示-概率图形模型，使用潜在随机变量来完成。

Monica Haurilet[7]等人提出了可学习场景图的遍历策略的图神经网络，图模型中的 traveler 根据 guide 模块建议的方向遍历 visual graph。将组合图像特征，问题特征，用完全连接层投影到同一空间。

Damien Teney[8]等人使用场景内容和问题的结构化表示来改进视觉问题解答，对图片场景的描述通过图模型得到表示，对解析的问题通过图模型得到表示。通过递归单位（GRU）模型对两个图表示的每个节点关联。在多次迭代中，GRU 更新了每个节点的表示形式，该表示形式集成了图中邻居的上下文。



探索生成模型并学习潜在表示空间。早期的工作主要集中在堆叠式自动编码器上，然后集中在严格的 Botzman 机器上。这些应用的最新成功主要是可变自动编码器（VAE）]和生成对抗网络（GAN）]的结果。利用重新参数化技巧，可以训练 VAE 来学习半监督的潜在空间以生成图像[9]。它们也已扩展到连续状态空间[10，11]和顺序模型[12，13]。

另一方面，GAN 可以学习支持基本线性代数的图像表示[14]，甚至可以通过对贝叶斯程序使用概率推断来实现一次性学习[15]。VAE 和 GAN 都已根据类标签或其他视觉变化解开了

它们的表示[16,17]。

Ranjay Krishna[18]等人将图像表征和答案表征馈入 VAE（variational auto-encoders ），该 VAE 将二者嵌入到一个潜在空间中。该潜在空间通过图像，问题和答案的编码最大化了相互信息。

Qingxing Cao[19]等人提出对抗组成神经模型 ACMN，其对抗模块以输入词嵌入和加权图像空间特征，输出新的注意力图。

协调表示，通常是对视觉信号，文本信号进行单独的变换处理，各自映射到的空间是具有协调性地，这种协调表示可以用最小化余弦相似度，最大化相关等方式。

Chenfei Wu[20]等人用线性变换把视觉特征和文本特征被投影到相同的维度，提出用差分网络融合两种模态信息，显示了其优于全连接网络。在用基于差分网络的融合模型解决 VQA 的过程中，用 multi-glimpse 注意力对齐了视觉特征和文本特征。

Xiangpeng Li[21]等人提出 Positional Self-Attention with Co-attention 模型解决视频问题回答，使用多元素函数来构建一个相似度矩阵，使视频特征和问题特征投影到协调空间，用 V2Q，Q2V 注意力模块隐式对齐与问题向量相关的视频特征，通过共同注意力模型完成融合多个特征，预测答案。

视觉回答中的对齐是找到实例子元素之间的对应关系。例如，将图像的某个区域或某个物体与问题的某个词、答案相对应，也可以是问题或答案与某个时间定位的视频段对齐等。

对齐大致可以分为显示对齐和隐式对齐，解决 VQA，多数都是用注意力机制来进行隐式对齐。

Zichao Yang[22]等人提出的模型将使用堆叠注意力模型对齐问题序列和图像区域，Ishan Misra[2]等人提出的 LBA 模型中的问题回答模块，Kushal Kafle[23]等人提出 MOM（Multi-Output Model）模型的分类子网模块采用了类似的模型。

Lianli Gao[24]等人提出结构化双流注意力模型，链接并融合了来自视频片段和文本的信息。该注意力层由 N 个双流（即文本和视频特征）注意力组成，并且所有注意力模型共享参数对齐视频段特征和文本特征。用结构化的双流注意力模型的隐层隐式对齐了视频段特征和文本特征。

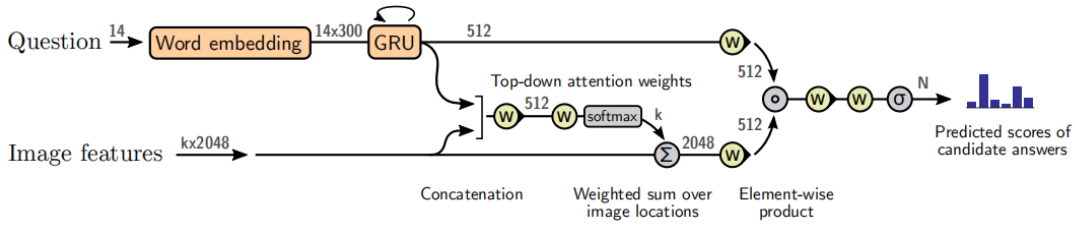
Yiyi Zhou[5]等人在胶囊网络的基础上，提出了解决 VQA 的胶囊注意力模型，直接计算模态之间的耦合系数来显示对齐视觉特征和文本特征。在 CapsAtt 后使用两个全连接层完成答案的预测，实验表明，比传统的多步注意力模型（例如，Yang 等人，2016 年的堆叠式注意力网络，SAN）获得更好的性能。

Robik Shresth[6]等人提出 RAMEN 模型使用双向门控递归单元（bi-GRU）跨场景聚合双模态嵌入，以捕获双模态嵌入之间的交互，对齐了输入图像和问题序列。尽管最新的处理图像 VQA 模型使用注意力或双线性池化机制，但 RAMEN 可以在没有这些机制的情况下达到很好的效果。

Dalu Guo[25]等人提出的协同网络（synergistic network）中使用共同注意模块解决在当前问题中提到的图像中定位对象，潜在的对齐了图像和问题。

Peter Anderson[26]等提出了一种自下而上和自上而下的组合注意力机制，使注意力可以在对象和其他显着图像区域的水平上进行计算。在该方法中，自下而上的机制提出了图像区域，每个区域都具有关联的特征向量，而自上而下的机制确定了特征权重，通过细粒度的分析甚至推理的多个步骤来实现对图像的更深入理解。在解决 VQA，使用“软”自上而下的注意力机制，以问题表示为上下文来加权每个空间图像特征，实现了问题和图像的的联合嵌入。

自上而下的注意力机制也隐式对齐了问题和图像。



Hyeonwoo Noh[27]等人使用该方法的自下而上的注意力机制获得基于图像和边界框的视觉特征，通过任务条件视觉分类器学习答案特征，视觉特征与根据问题说明编码的任务特征的联合分布中的参数，迁移学习解决 VQA。

Lu[28]等人提出了分层协同注意模型来交替学习图像注意和问题注意完成图像和问题的融合，预测出答案。

Nam[29]等人提出了双重注意力网络 DAN，以基于先前注意力的记忆来提炼注意力。但是，这些共同注意模型为每个模态（图像或问题）学习单独的注意力分布，而忽略了每个疑问词与每个图像区域之间的密集交互，这成为理解多模态特征的细粒度关系的瓶颈。

Duy-Kien Nguyen[30]使用密集的共同注意模型融合图像区域和问题的连接。Jin-Hwa Kim[31]也用类似的方式解决 VQA。

Chao Ma[32]等人用共同注意力机制来关注相关的图像区域和问题词，潜在的对齐图像和问题特征，将图像和问题特征向量简单连接起来，并将其馈入记忆增强网络 MAN，将记忆增强网络融合图像和问题对的最终嵌入，并将该嵌入馈入分类器以预测答案 。

Peng Wang[33]等人提出了一个新模型，将输入的问题，事实和图像特征在三个级别（单词，短语，句子）上加权编码问题。通过多个共同注意力机制在 3 个级别进行共加权得到融合表示，可以使用多层感知器（MLP）分类器来预测答案。然后使用排名的事实来生成原因。

Zhou Yu[34]提出了深度模块化协同注意网络，将文本特征和图像特征组合传递到由深度级联的 L 个 MCA 层组成的深度协同关注模块（deep co-attention model），输出得到两个模态的联合表示。

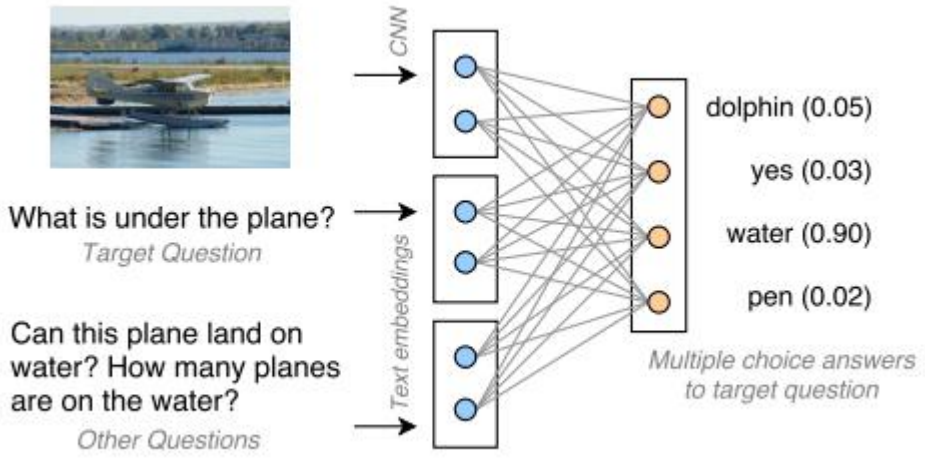
Yunseok Jang[35]等人提出了时空 VQA（ST-VQA）模型，使用两个双层 LSTM 来捕获视频和问题答案句子之间的视觉-文本关联，潜在地对齐了视觉和文本特征。

Dongfei Yu[36]等人提出 multi-level attention network (MLAN)模型，一种用于视觉问题回答的多层注意力网络，使用双向递归神经网络把来自 CNN 的基于区域的中级输出编码为空间嵌入的表示形式，并通过多层感知器进一步明确与答案相关的区域，进行隐式对齐答案与图像区域。最后，通过 softmax 分类器共同优化语义注意，视觉注意力和问题嵌入，以推断出最终答案。

Vasu Sharma[37]使用一种称为 SegAttend Net 的新颖的分段指导基于注意力的网络来解决视觉问题回答的问题。使用由全卷积深度神经网络生成的图像分割图来精炼我们的注意力图，并使用这些精炼的注意力图使模型对齐图像区域和问题。

Bolei Zhou[38]等人提出了 iBOWIMG 的框架。将疑问句和图像中的特征连接起来，完成多模态信息的联合表示，然后馈入 softmax 以预测答案，该模型潜在地将疑问句中的信息性单词和图像中的视觉概念与答案对齐。

Siddha Ganju[39]等人对 iBOWIMG 进行了扩展，提出了 iBOWIMG-2x 模型，此附加特征向量与图像和目标问题特征相连。完成预测目标问题的答案，并且可以认为是图像的更丰富的特征表示。

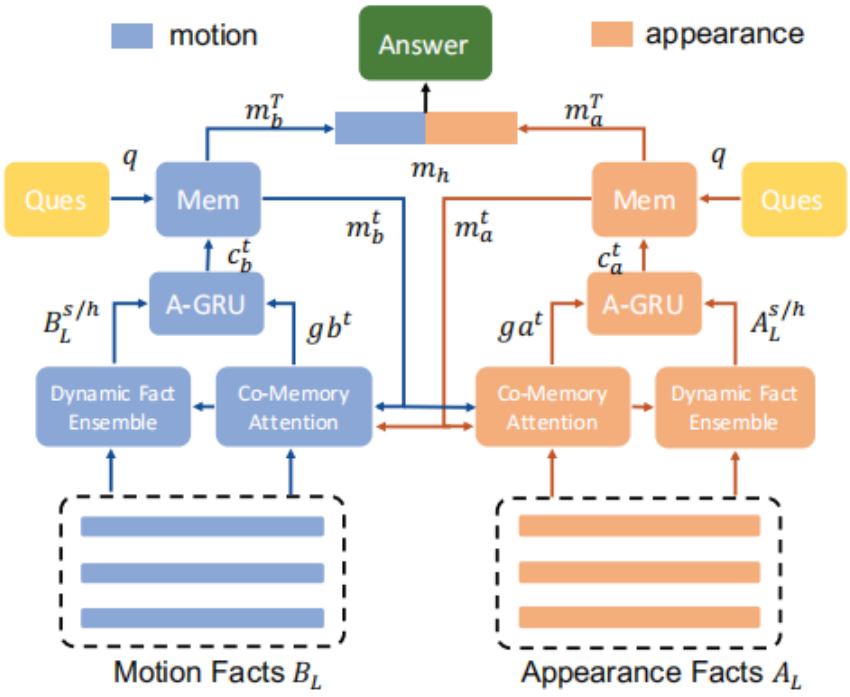


Bo Wang[40]等人提出了 Layered Memory Network（LMN）模型通过其动态字幕记忆模块 Dynamic Subtitle Memory Module 进行句子和视频帧对齐。

动态内存网络（DMN），最早由 Kumar 等人引入。[41]解决基于文本的问题解答，采用了情景记忆和注意力机制，从而允许进行多个推理循环。熊等[42]改进了 DMN 的记忆模块和输入模块，使其可以应用于图像问题回答。

Jiyang Gao[43]提出了模型运动-外观共记忆网络，基于 DMN / DMN + [41，42]的概念，具体来说，设计了一种共同记忆的注意力机制来共同对运动和外观信息进行建模表示视频特征。

然后将运动和外观特征输入到时间卷积和反卷积神经网络中，以构建具有相同时间分辨率但代表不同上下文信息的多级上下文事实。引入一种称为动态事实集成的方法，以在事实编码的每个周期中动态产生时间事实。最终将运动和外观记忆网络的输出串联，用于生成答案。



Hexiang Hu[44]等人提出了一种思想，对某些空间的图像和问题特征进行联合嵌入，得

到答案嵌入去参数化了一个描述答案与图像和问题对的相似程度的概率模型，将学习答案，图像和问题的嵌入通过 PMC，Probabilistic Model of Compatibility 模型将答案的语义相似度与图像和问题对的视觉/语义相似度对齐，以最大程度地提高正确答案的可能性。

解决 VQA 的融合，是把视觉信息和文本信息集成在一起，以进一步预测答案为目标。在最近的研究中，双模态表示和融合的方法有很多的交织。

早期的工作已经用一阶交互对多个模态之间的交互进行了建模。

M. Ren 等人[45]提出 IMG + BOW 模型是第一个使用简单连接去合并全局图像表示和问题嵌入的模型，该问题嵌入是通过对问题中所有学习的单词嵌入进行求和而获得的。

在[46]中，区域的视觉特征和文字嵌入之间的简单连接得到两个模态的嵌入，通过一个注意力框架，每个局部特征得到一个对应于文本特征相似性的分数。这些分数用于区域的权值，融合多模态信息。Bo Wang[40]等人提出了 Layered Memory Network（LMN）模型的静态词记忆模块中用相似的方法得到其多模态表示，Kuo-Hao Zeng[2]等人提出的扩展模型,E-SA，也采取类似的方式完成融合。

Lianli Gao[24]等人提出结构化双流注意力模型中的结构化双流融合模块，对视频段特征和文本特征都使用融合，Monica Haurilet[7]等人将视觉全局表示与嵌入问题连接起来。然后，使用完全连接的层进行融合，Manoj Acharya[3]等人用关系计数网络中的子网络，关系网络，各自处理前景提议与问题序列串联以及前景提议和背景区域的串联，用连接的方式融合特征，都是采取早期融合，最后通过全连接层和 softmax 层完成进一步答案，计数预测。

分层的共同注意网络[47]在提取了多个文本和视觉特征之后，将它们串联之后求和合并，完成融合。

二阶模型是对两个嵌入空间之间的交互进行建模的更强大的方法。

在深度学习中，双线性交互在细粒度分类和多模态语言建模中显示出了巨大的成功。

在 VQA 中，在[48]中执行两个向量之间的简单逐元素乘积。

Kuo-Hao Zeng[2]等人提出的扩展模型 E-VQA 也采取类似的方式完成多模态融合。

文献[49]还在更复杂的迭代全局合并方案中使用了基于元素的乘积（Hadamard product）。

在[50]中，他们在注意框架中使用了按元素相乘聚合。为了深入研究双线性相互作用，多模态紧凑双线性池（MCB）[51]在视觉 v 和文本 q 嵌入之间使用了外部乘积 q⊗v。

计数草图投影用于将问题和视觉的元素积投影到较低维度的空间上。

但是，MCB 中的交互参数由计数草图投影固定（在{0; 1; 1}中随机选择），从而限制了其用于建模图像和问题之间复杂交互的表达能力。

在多模态低秩双线性（MLB）池化工作中[8]，图像和问题空间之间的完全双线性相互作用是由张量参数化的。

同样，为了限制自由参数的数量，将张量约束为低秩 r。

MLB 策略在著名的 VQA 数据库上达到了最先进的性能[52]。Zhou Su[53]等人用该方法融合了视觉问题对，视觉问题对和知识条目的一般联合嵌入通过 Visual Knowledge Memory Network VKMN 完成融合，进行预测答案。

Tingting Qiao[54]等人也提出了人类注意网络（HAN）模型用该方法融合了图像特征和问题特征。

Ilija Ilievski[55]等人提出 ReasonNet 用 MLB 得到模块表示和问题表示的交互向量，串联交互向量进行多模态融合。

尽管获得了这些令人印象深刻的结果，但低秩张量结构等效于将可视表示形式和问题表示形式投影到共同的 r 维空间中，并在该空间中计算简单的元素级乘积交互作用。因此，MLB 本质上是为学习强大的文本和图像模态单模态嵌入而设计的，但它依赖于此空间中的

简单融合方案。

Yu 等人指出，MLB 的收敛速度较慢。[56]提出了一种多模态分解双线性（MFB）池，它利用矩阵分解技巧来计算融合特征，以减少参数数量并提高收敛速度。

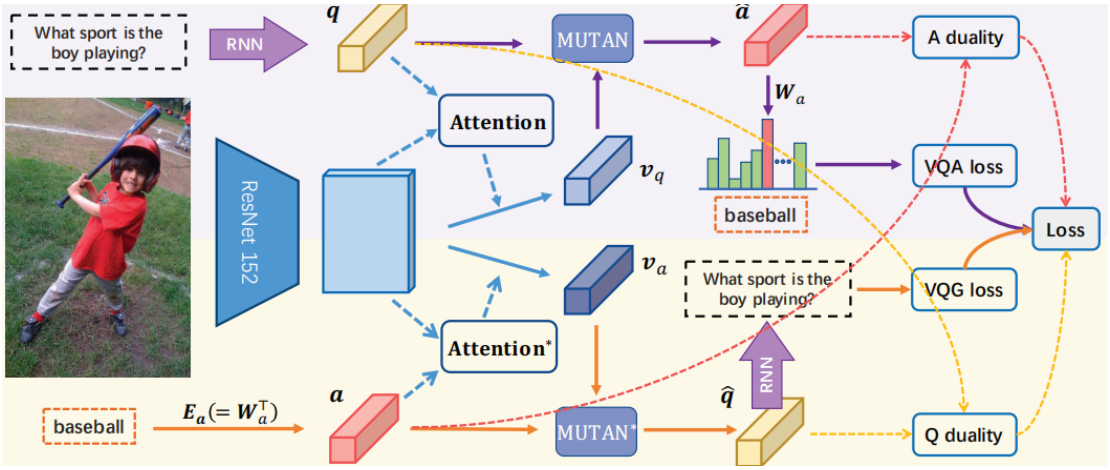
可以代替简单的线性串联的方法，与其他双线性方法（例如 MLB 和 MCB ）相比，MFB 可以提供更丰富的表示。Dalu Guo[25]等人提出的协同网络（synergistic network）中用 MFB 来克服两个特征的分布之间的差异（两个 LSTM 分别编码问题和历史； LSTM 分别用于文本特征和 CNN 用于图片特征），完成两种模态信息的融合。将历史矢量与问题矢量连接起来，然后与图像特征用 MFB 融合。最后，我们再使用 MFB 学习得到文本特征和视觉特征的联合表示。

Zhou Yu[34]提出了深度模块化协同注意网络，将图像特征，文本特征经过 MLP 的注意力减少模块处理后，用线性多模态融合函数 LayerNorm 映射得到融合特征。

相比之下，MUTAN 能够使用学习的参数对丰富的二阶交互进行建模。

Ben-younes[57]等人提出的 MUTAN，这是一种基于模态之间的双线性相互作用的多模态融合方案。为了控制模型参数的数量，MUTAN 减小了单峰嵌入的大小，同时使用完整的双线性融合方案尽可能精确地模拟了它们的相互作用。

Yikang Li[58]等人提出可逆问答网络（iQAN），其中基于 MUTAN 的注意力模块用于从图像和问题生成感知问题的视觉特征。然后使用另一个 MUTAN 融合模块通过融合图像和问题获得答案特征。最后，使用线性分类器来预测答案。



注意机制也可以被认为是特征融合方法，无论是否明确提及，因为它们被设计为基于它们之间的相互作用来更好地表示图像问题对。

对于其中两个特征被对称对待的共同注意机制尤其如此。Duy-Kien Nguyen[30]提出了 Dense Symmetric Co-Attention 密集对称的共同注意网络，改善视觉和语言表达的融合，它考虑了任何图像区域和任何疑问词之间的每一次相互作用。该层在两个模块之间具有完全对称的体系结构，并且可以堆叠形成一个层次结构，该层次结构使图像问题对之间可以进行多步交互。从输入图像和问题的联合表示开始，层堆栈中的每个密集的共同注意层都会更新表示，然后将其输入到下一层。然后将其最终输出馈送到图层以进行更快速的预测。

Unnat Jain[59]等人提出对称判别基准模型，将问题嵌入，图像嵌入，标题嵌入，历史嵌入和答案嵌入针对每个可能的答案选项进行串联，并采用两层的 MLP 进行晚期融合，相似性网络来预测可能答案的概率分布。

现有的多模态融合方法[49、50、51]仅关注对模态之间交互作用丰富的建模。但是，这些方法与问题无关，因为融合过程并不取决于问题。Junyeong Kim[60]等人提出渐进式注意力机制 PAMN 通过依次计算每个视频情节和问题嵌入，答案嵌入之间的余弦相似度来进行潜

在的对齐，渐进式注意力机制模型 PAMN 的动态模式融合模块在每个渐进式注意力步骤结束时将双重记忆聚合（通过连接双重记忆求余弦相似度得到权重+求和相乘）为融合输出。

Yuke Zhu[4]等人提出动态的可迭代的 VQA 模型，其模型的核心网络通过输入问题，预测答案并维持其内部存储库，同时通过查询来反复获得新证据，整个核心网络可以端到端的方式进行联合训练。

Qingxing Cao[19]等人提出对抗组成神经模型 ACMN 在给定注意力图情况下，由图像中每个网格特征的加权总和生成节点的视觉表示 h。给定问题的依存树的树状结构，ACMN 模块将在每个词节点上顺序使用，以挖掘视觉证据并从下至上整合其子节点的特征，然后在树的根部预测最终答案。

参考文献：

[1] Gordon D, Kembhavi A, Rastegari M, et al. Iqa: Visual question answering in interactive environments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4089-4098.

[2] Zeng K H, Chen T H, Chuang C Y, et al. Leveraging video descriptions to learn video question answering[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[3] Acharya M, Kafle K, Kanan C. TallyQA: Answering complex counting questions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8076-8084.

[4] Zhu Y, Lim J J, Fei-Fei L. Knowledge acquisition for visual question answering via iterative querying[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1154-1163.

[5] Zhou Y, Ji R, Su J, et al. Dynamic Capsule Attention for Visual Question Answering[J]. 2019.

[6] Shrestha R, Kafle K, Kanan C. Answer them all! toward universal visual question answering models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10472-10481.

[7] Haurilet M, Roitberg A, Stiefelhagen R. It's Not About the Journey; It's About the Destination: Following Soft Paths Under Question-Guidance for Visual Reasoning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1930-1939.

[8] Teney D, Anderson P, He X, et al. Tips and tricks for visual question answering: Learnings from the 2017 challenge[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4223-4232.

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. ICLR, 2014. 3, 4

[10] R. G. Krishnan, U. Shalit, and D. Sontag. Deep kalman fifilters. arXiv preprint arXiv:1511.05121, 2015. 3

[11] E. Archer, I. M. Park, L. Buesing, J. Cunningham, and L. Paninski. Black box variational inference for state space models. arXiv preprint arXiv:1511.07367, 2015. 3

[12] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015. 3

[13] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In Advances in neural information processing systems, pages 2980–2988, 2015. 3

[14] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015. 3

[15] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Humanlevel concept learning through probabilistic program induction. Science, 350(6266):1332–1338, 2015. 3

[16] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Advances in Neural Information Processing Systems, pages 3581–3589, 2014. 3

[17] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. arXiv preprint arXiv:1511.05644, 2015. 3

[18] Krishna R, Bernstein M, Fei-Fei L. Information Maximizing Visual Question Generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2008-2018.

[19] Cao Q, Liang X, Li B, et al. Visual question reasoning on general dependency tree[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7249-7257.

[20] Wu C, Liu J, Wang X, et al. Differential Networks for Visual Question Answering[J]. 2019.

[21] Li X, Song J, Gao L, et al. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering[C]//The 33rd AAAI Conference on Artificial Intelligence. 2019, 8.

[22] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016. 1, 2

[23] Kafle K, Price B, Cohen S, et al. DVQA: Understanding data visualizations via question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5648-5656.

[24] Gao L, Zeng P, Song J, et al. Structured Two-Stream Attention Network for Video Question Answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 6391-6398.

[25] Guo D, Xu C, Tao D. Image-question-answer synergistic network for visual dialog[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10434-10443.

[26] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[27] Noh H, Kim T, Mun J, et al. Transfer Learning via Unsupervised Task Discovery for Visual Question Answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8385-8394.

[28] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]//Advances In Neural Information Processing Systems. 2016: 289-297.

[29] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 299-307.

[30] Nguyen D K, Okatani T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6087-6096.

[31] Kim J H, Jun J, Zhang B T. Bilinear attention networks[C]//Advances in Neural Information Processing Systems. 2018: 1564-1574.

[32] Ma C, Shen C, Dick A, et al. Visual question answering with memory-augmented networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6975-6984.

[33] Wang P, Wu Q, Shen C, et al. The vqa-machine: Learning how to use existing vision

algorithms to answer new questions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1173-1182.
[34] Yu Z, Yu J, Cui Y, et al. Deep Modular Co-Attention Networks for Visual Question Answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6281-6290.
[35] Jang Y, Song Y, Yu Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2758-2766.
[36] Yu D, Fu J, Mei T, et al. Multi-level attention networks for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4709-4717.
[37] Sharma V, Bishnu A, Patel L. Segmentation guided attention networks for visual question answering[C]//Proceedings of ACL 2017, Student Research Workshop. 2017: 43-48.
[38] Zhou B, Tian Y, Sukhbaatar S, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015.
[39] Ganju S, Russakovsky O, Gupta A. What's in a question: Using visual questions as a form of supervision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 241-250.
[40] Wang B, Xu Y, Han Y, et al. Movie question answering: remembering the textual cues for layered visual contents[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[41] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In ICML, 2016. 1, 2, 3
[42] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In ICML, 2016. 1, 2, 3, 5
[43] Gao J, Ge R, Chen K, et al. Motion-appearance co-memory networks for video question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6576-6585.
[44] Hu H, Chao W L, Sha F. Learning answer embeddings for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5428-5436.
[45] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In NIPS, pages 2953– 2961, 2015.
[46] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In CVPR, 2016.
[47] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In NIPS, pages 289–297, 2016
[48] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In ICCV, 2015.
[49] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In 5th International Conference on Learning Representations, 2017.
[50] R. Li and J. Jia. Visual question answering with question representation update (qru). In NIPS, pages 4655–4663. 2016.
[51] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact

bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847, 2016
[52] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In ICCV, 2015
[53] Su Z, Zhu C, Dong Y, et al. Learning visual knowledge memory networks for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7736-7745.
[54] Qiao T, Dong J, Xu D. Exploring human-like attention supervision in visual question answering[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[55] Ilievski I, Feng J. Multimodal learning and reasoning for visual question answering[C]//Advances in Neural Information Processing Systems. 2017: 551-562.
[56] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In International Conference on Computer Vision (ICCV), 2017. 1, 2, 6, 8
[57] Ben-Younes H, Cadene R, Cord M, et al. Mutan: Multimodal tucker fusion for visual question answering[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2612-2620.
[58] Li Y, Duan N, Zhou B, et al. Visual question generation as dual task of visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6116-6124.
[59] Jain U, Lazebnik S, Schwing A G. Two can play this game: visual dialog with discriminative question generation and answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5754-5763.
[60] Kim J, Ma M, Kim K, et al. Progressive Attention Memory Network for Movie Story Question Answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8337-8346.