Visual Dialog

Yingying Zhuang

Apr, 2019

Co-Attention

sequential co-attention generative mode[1]

- **Background：**

- *maximum likelihood estimation (MLE) objective function:*
  only focus on measuring the word-level correctness, the produced responses tend to be generic and repetitive

.

- **Main work:**

- *Combine Reinforcement Learning and Generative Adversarial Networks (GANs)*
  to generate more human-like responses to questions.

- *Propose a sequential co-attention generative model*
  to ensure that attention can be passed effectively across the image, question and dialog history.

- *Propose the discriminator has access to the attention weights the generator used in generating its response.*

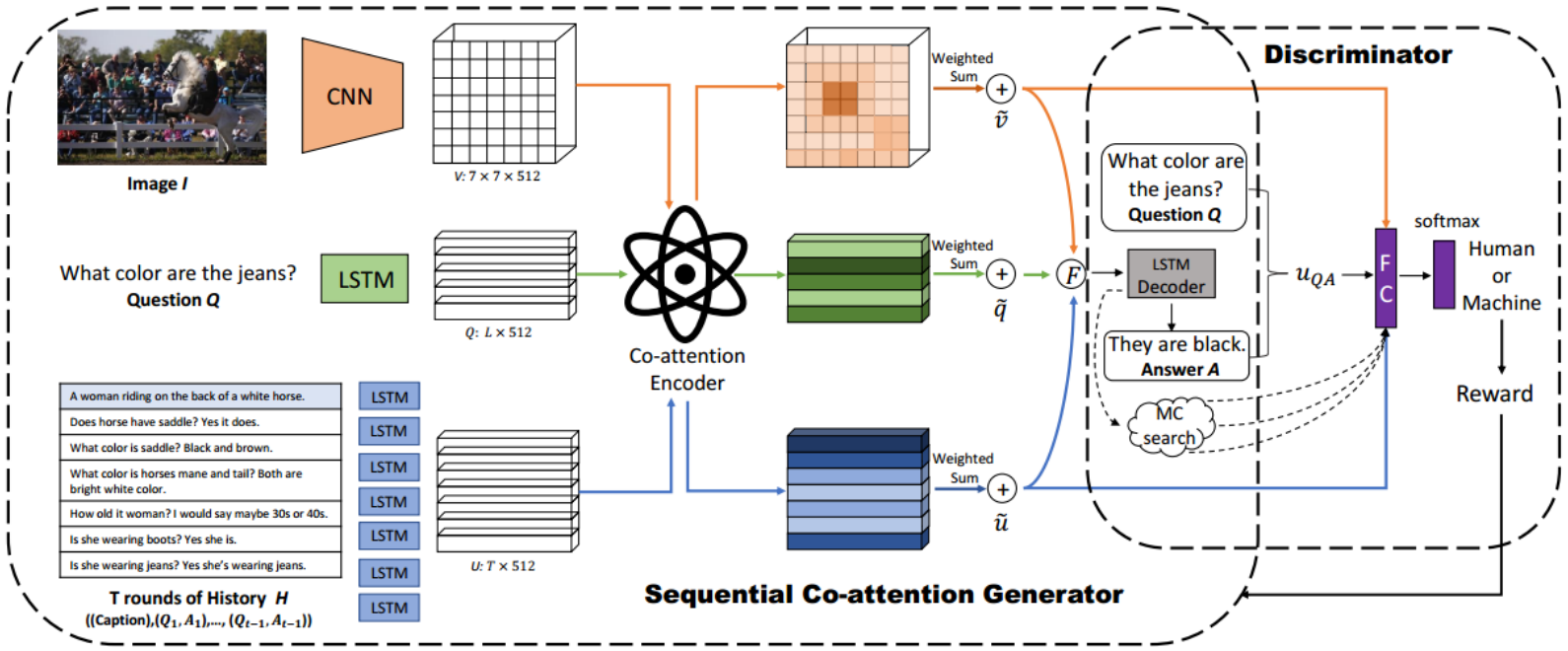
[1] Qi Wu1, Peng Wang2, Chunhua Shen1, Ian Reid1, and Anton van den Hengel1 (2017) Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In ACL

Co-Attention

sequential co-attention generative mode[1]

Model

.



- **Sequential Co-attention Generator**: accepts as input image, question and dialog history tuples, and uses the co-attention encoder to jointly reason over them.
- **Discriminator** tasked with labelling whether each answer has been generated by a human or the generative model by considering the attention weights.

[1] Qi Wu1, Peng Wang2, Chunhua Shen1, Ian Reid1, and Anton van den Hengel1 (2017) Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In ACL

Co-Attention

- **sequential co-attention generative mode[1]**
  - **Model**

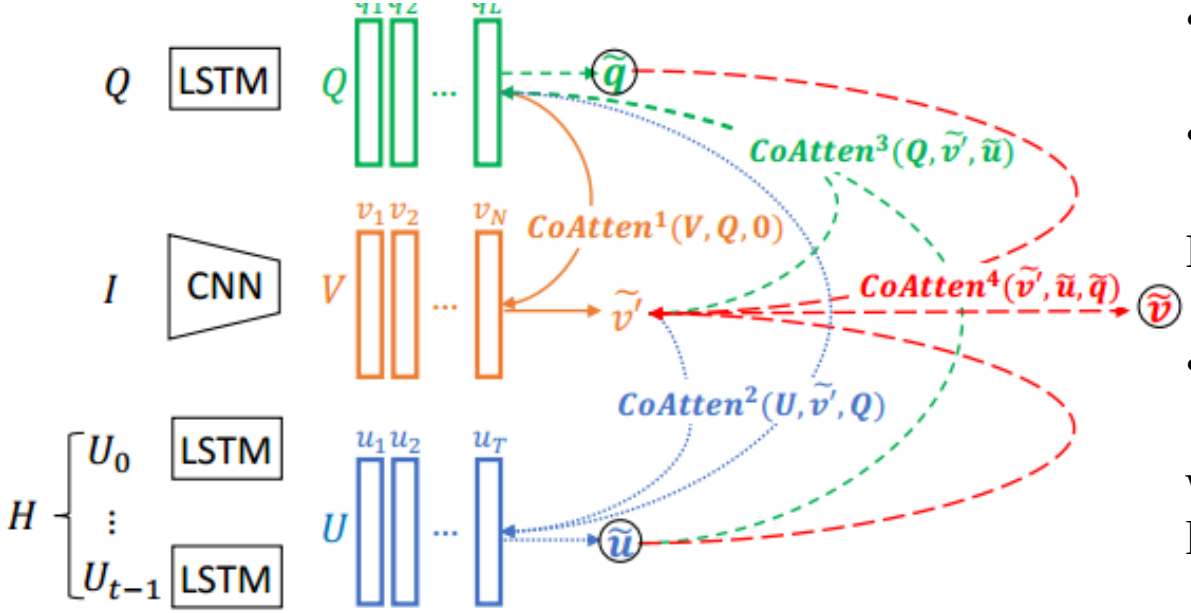


Figure 3: The sequential co-attention encoder. Each input feature is co-attend by the other two features in a sequential fashion, using the Eq.1-3. The number on each function indicates the sequential order, and the final attended features $\tilde{u}$,$\tilde{v}$ and $\tilde{q}$ form the output of the encoder.

- V =[v1; : : : ; vN] : spatial image features from the convolutional layer, where N is the number of image regions.
- Q =[q1; : : : ; qL],: question features where ql = LSTM(wl; ql-1), which is the hidden state of an LSTM at step l given the input word wl of the question. L is the length of the question
- U = [u0; : : : ; uT] H is composed by a sequence of utterance being extracted separately to make up the dialog history features where T is the number of rounds of the utterance (QA-pairs). u is th last hidden state of an LSTM,

[1] Qi Wu1, Peng Wang2, Chunhua Shen1, Ian Reid1, and Anton van den Hengel1 (2017) Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In ACL

Co-Attention

- **sequential co-attention generative mode[1]**
  - **Model**

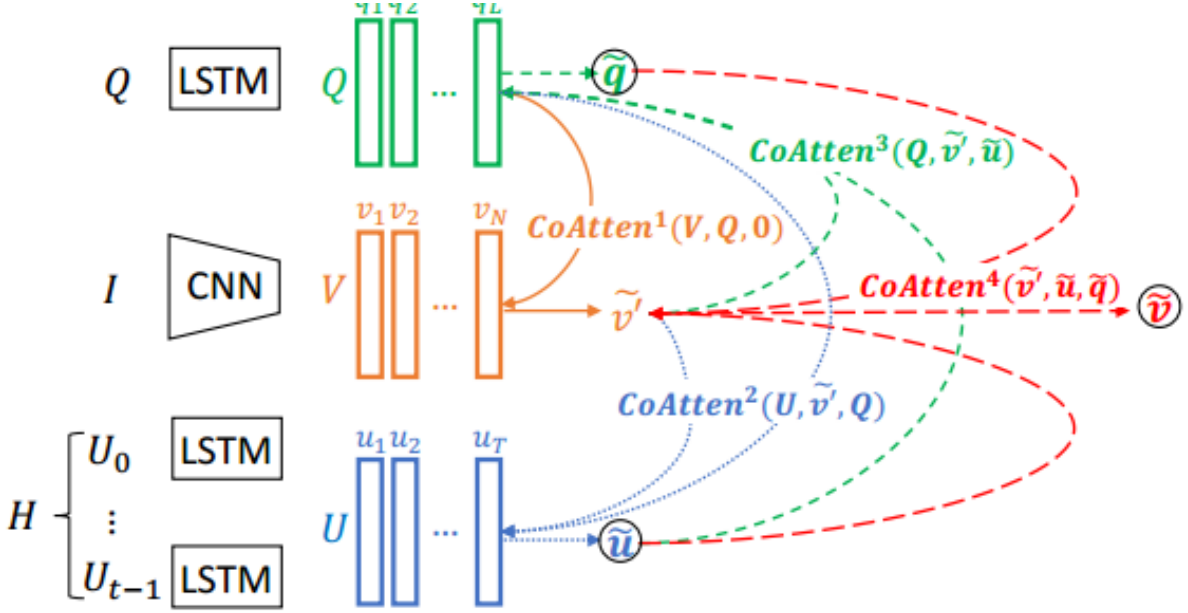


Figure 3: The sequential co-attention encoder. Each input feature is co-attend by the other two features in a sequential fashion, using the Eq.1-3. The number on each function indicates the sequential order, and the final attended features $\tilde{u}$,$\tilde{v}$ and $\tilde{q}$ form the output of the encoder.

$$\tilde{x} = \text{CoAtten}(X, g_1, g_2),$$

$$\begin{aligned} H_i &= \tanh(\mathbf{W}_x x_i + \mathbf{W}_{g_1} g_1 + \mathbf{W}_{g_2} g_2), \\ \alpha_i &= \text{softmax}(\mathbf{W}^T H_i), \quad i = 1, \ldots, M, \\ \tilde{x} &= \sum_{i=1}^{M} \alpha_i x_i, \end{aligned}$$

$$F = \tanh(\mathbf{W}_{eg}[\tilde{v}; \tilde{u}; \tilde{q}])$$

[1] Qi Wu1, Peng Wang2, Chunhua Shen1, Ian Reid1, and Anton van den Hengel1 (2017) Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In ACL

Co-Attention

- **sequential co-attention generative mode[1]**
  - Results

.

Model	MRR	R@1	R@5	R@10	Mean
LF [5]	0.5807	43.82	74.68	84.07	5.78
HRE [5]	0.5846	44.67	74.50	84.22	5.72
HREA [5]	0.5868	44.82	74.81	84.36	5.66
MN [5]	0.5965	45.55	76.22	85.37	5.46
SAN-QI [40]	0.5764	43.44	74.26	83.72	5.88
HieCoAtt-QI [19]	0.5788	43.51	74.49	83.96	5.84
AMEM [25]	0.6160	47.74	78.04	86.84	4.99
HCIAE-NP-ATT [18]	0.6222	48.48	78.75	87.59	4.81
Ours	**0.6398**	**50.29**	**80.71**	**88.81**	**4.47**

Table 2: Performance of discriminative methods on VisDial v0.9. Higher is better for MRR and recall@k, while lower is better for mean rank.

[1] Qi Wu1, Peng Wang2, Chunhua Shen1, Ian Reid1, and Anton van den Hengel1 (2017) Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In ACL

Co-Attention

➢ Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]

➢ **Background：**

- *Bilinear pooling*

has recently been used to integrate different CNN features for fine-grained image recognition，

but the **high dimensionality** of the output features and the huge number of model parameters may seriously limit the applicability of bilinear pooling

- *Multi-modal Compact Bilinear (MCB)*

pooling model to effectively and simultaneously **reduce the number of parameters** and computation time using the Tensor Sketch algorithm

Nevertheless, the MCB model lies on a **high-dimensional output feature** to guarantee robust performance, which may limit its applicability due to huge memory usage

- *Multi-modal Low-rank Bilinear (MLB)*

generate output features with **lower dimensions** and models with fewer parameters, it is highly competitive with MCB. However, MLB has **a slow convergence rate** and is **sensitive to the learned hyperparameters.**

[1] Zhou Yuy, Jun Yuy∗ , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]

.

Main work:

- *Develop a simple but effective Multi-modal Factorized Bilinear pooling (MFB) approach* to fuse the visual features from images with the textual features from questions.

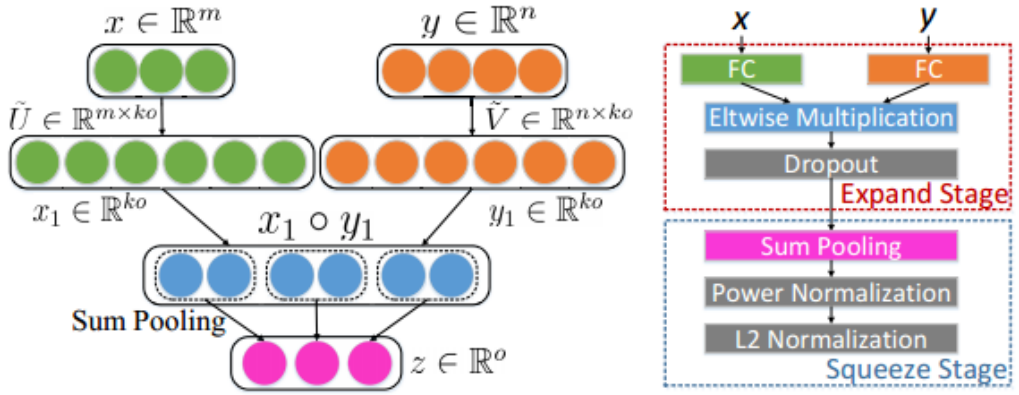
- *Design a co-attention learning architecture based on the MFB* to jointly learn both image and question attention module.

[2] Zhou Yuy, Jun Yuy∗ , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]

- **Model:**



(a) Multi-modal Factorized Bilinear Pooling (b) MFB module

Figure 1. The flowchart of Multi-modal Factorized Bilinear Pooling and completed design of the MFB module.

$$z_i = x^T W_i y$$

$$\begin{aligned} z_i &= x^T U_i V_i^T y = \sum_{d=1}^{k} x^T u_d v_d^T y \\ &= \mathbb{1}^T (U_i^T x \circ V_i^T y) \end{aligned}$$

where k is the factor or the latent dimensionality of the factorized matrices $U_i = [u_1, ..., u_k] \in \mathbb{R}^{m \times k}$ and $V_i = [v_1, ..., v_k] \in \mathbb{R}^{n \times k}$, $\circ$ is the Hadmard product or the element-wise multiplication of two vectors, $\mathbb{1} \in \mathbb{R}^k$ is an all-one vector.

$$z = \text{SumPooling}(\tilde{U}^T x \circ \tilde{V}^T y, k)$$

where the function $\text{SumPooling}(x, k)$ means using a one-dimensional non-overlapped window with the size k to perform sum pooling over x. We name this model Multi-modal Factorized Bilinear pooling (MFB).

[1] Zhou Yuy, Jun Yuy* , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

➢ Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]

➢ **Model：**

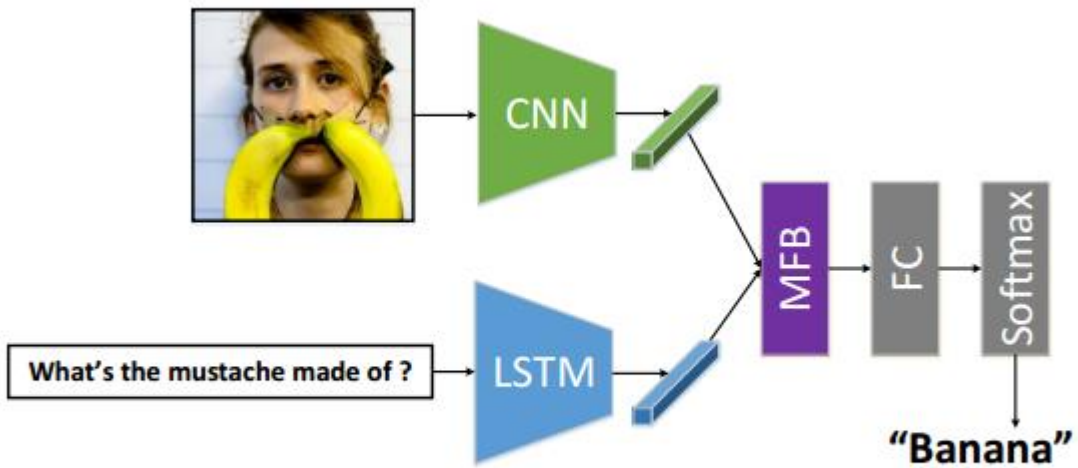


Figure 2. MFB baseline network architecture for VQA.

- *Image features*

152-layer ResNet model pre-trained on the ImageNet dataset.

↓

Resized to 448 × 448, and 2048-D *pool5* features(with normalization)

- *Questions*：

Questions are first tokenized into words,

↓

one-hot feature vectors with max length *T*.

↓

one-hot vectors are passed through an embedding layer

↓

fed into a two-layer LSTM networks with 1024 hidden units

↓

form a 2048-D feature vector for question representation

[1] Zhou Yuy, Jun Yuy∗ , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

- **Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]**
- **Model：**

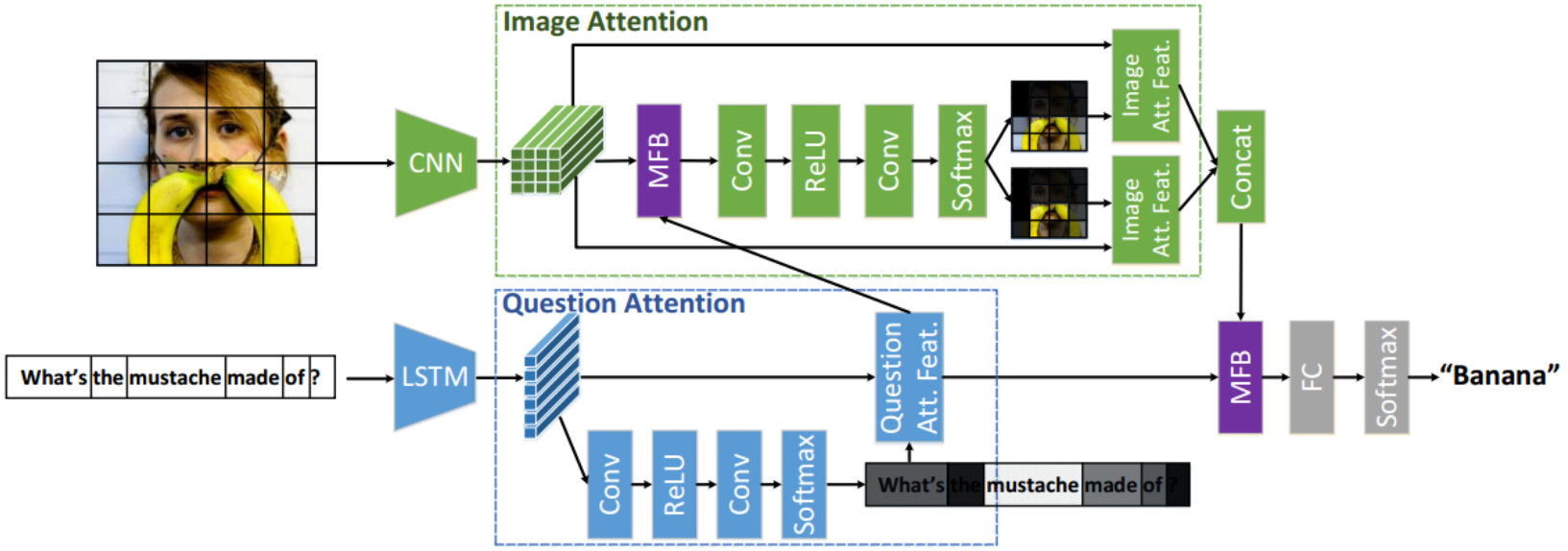


Figure 3. MFB with Co-Attention network architecture for VQA. Different from the network of MFB baseline, the images and questions are firstly represented as the fine-grained features respectively. Then, *Question Attention* and *Image Attention* modules are jointly modeled in the framework to provide more accurate answer predictions.

[2】 Zhou Yuy, Jun Yuy∗ , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

- **Multi-modal Factorized Bilinear Pooling with Co-Attention Learning [2]**

- **Results:**

Model	Acc.	Model Size
MCB[6] ($d = 16000$)	59.8	63M
MLB[12] ($d = 1000$)	59.7	25M
MFB($k = 1, o = 5000$)	60.4	51M
MFB($k = 5, o = 1000$)	**60.9**	46M
MFB($k = 10, o = 500$)	60.6	38M
MFB($k = 5, o = 200$)	59.8	22M
MFB($k = 5, o = 500$)	60.4	28M
MFB($k = 5, o = 2000$)	60.7	62M
MFB($k = 5, o = 4000$)	60.4	107M
MFB($k = 5, o = 1000$)	-	-
-w/o power norm.	60.4	-
-w/o ℓ_2 norm.	57.7	-
-w/o power and ℓ_2 norms.	57.3	-

Model	Att.	W.E.	Test-dev					Test-standard				
			OE				MC	OE				MC
			All	Y/N	Num	Other	All	All	Y/N	Num	Other	All
iBOWIMG [39]			55.7	76.5	35.0	42.6	-	55.9	78.7	36.0	43.4	62.0
DPPnet [23]			57.2	80.7	37.2	41.7	-	57.4	80.3	36.9	42.2	-
VQA team [3]			57.8	80.5	36.8	43.1	62.7	58.2	80.6	36.5	43.7	63.1
AYN [20]			58.4	78.4	36.4	46.3	-	58.4	78.2	36.3	46.3	-
AMA [31]			59.2	81.0	38.4	45.2	-	59.4	81.1	37.1	45.8	-
DMN+ [32]			60.3	80.5	36.8	60.3	-	60.4	-	-	-	-
MCB [6]			61.1	81.7	36.9	49.0	-	61.1	81.7	36.9	49.0	-
MRN [11]			61.7	**82.3**	**38.9**	49.3	-	61.8	82.4	38.2	49.4	66.3
MFB (Ours)			**62.2**	81.8	36.7	**51.2**	**67.2**	-	-	-	-	-
SMem [33]	✓		58.0	80.9	37.3	43.1	-	58.2	80.9	37.3	43.1	-
NMN [2]	✓		58.6	81.2	38.0	44.0	-	58.7	81.2	37.7	44.0	-
SAN [36]	✓		58.7	79.3	36.6	46.1	-	58.9	-	-	-	-
FDA [9]	✓		59.2	81.1	36.2	45.8	-	59.5	-	-	-	-
DNMN [1]	✓		59.4	81.1	38.6	45.4	-	59.4	-	-	-	-
HieCoAtt [18]	✓		61.8	79.7	38.7	51.7	65.8	62.1	-	-	-	-
RAU [22]	✓		63.3	81.9	39.0	53.0	67.7	63.2	81.7	38.2	52.8	67.3
MCB+Att [6]	✓		64.2	82.2	37.7	54.8	-	-	-	-	-	-
DAN [21]	✓		64.3	83.0	39.1	53.9	69.1	64.2	82.8	38.1	54.0	69.0
MFB+Att (Ours)	✓		64.6	82.5	38.3	55.2	69.6	-	-	-	-	-
MFB+CoAtt (Ours)	✓		**65.1**	83.2	38.8	**55.5**	**70.0**	-	-	-	-	-
MCB+Att+GloVe [6]	✓	✓	64.7	82.5	37.6	55.6	-	-	-	-	-	-
MLB+Att+StV [12]	✓	✓	65.1	**84.1**	38.2	54.9	-	65.1	84.0	37.9	54.8	68.9
MFB+CoAtt+GloVe (Ours)	✓	✓	**65.9**	84.0	**39.8**	**56.2**	**70.6**	65.8	83.8	**38.9**	56.3	70.5
MCB+Att+GloVe+VG [6]	✓	✓	65.4	82.3	37.2	57.4	-	-	-	-	-	-
MLB+Att+StV+VG [12]	✓	✓	65.8	83.9	37.9	56.8	-	-	-	-	-	-
MFB+CoAtt+GloVe+VG (Ours)	✓	✓	**66.9**	**84.1**	**39.1**	**58.4**	**71.3**	**66.6**	**84.2**	38.1	**57.8**	**71.4**

[2] Zhou Yuy, Jun Yuy∗ , Jianping Fanz, Dacheng Tao (2017) Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In ICCV

Co-Attention

➢ Hierarchical Question-Image Co-Attention [3]

➢ Model:

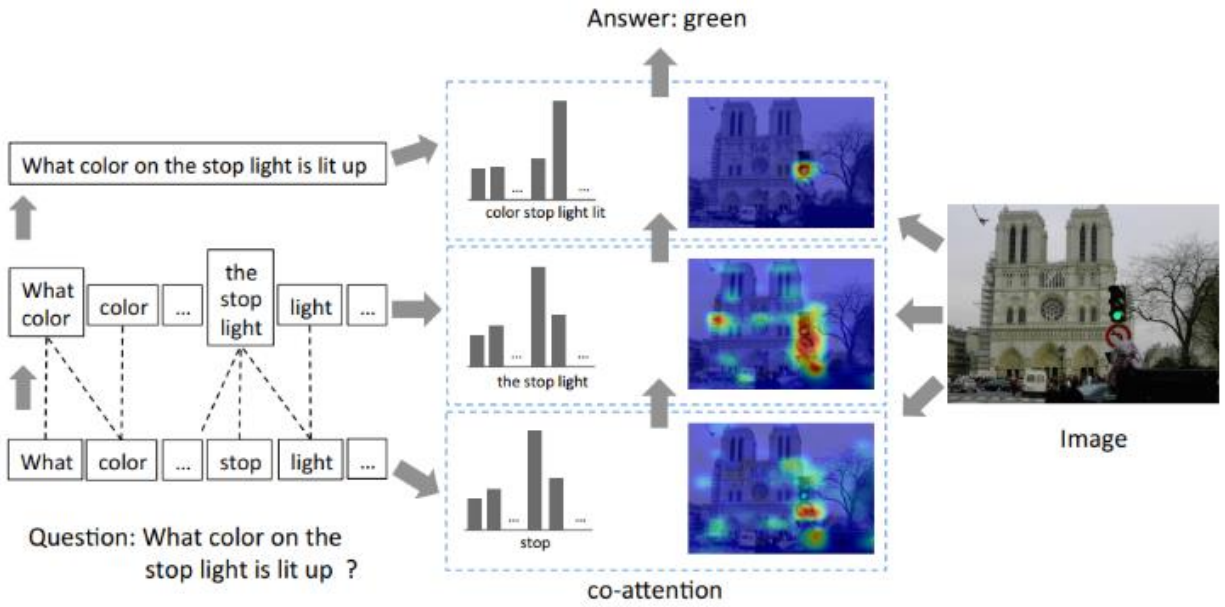


Figure 1: Flowchart of our proposed hierarchical co-attention model. Given a question, we extract its word level, phrase level and question level embeddings. At each level, we apply co-attention on both the image and question. The final answer prediction is based on all the co-attended image and question features.

1. Notation:

问题 $Q = \{q_1, \dots, q_T\}$，其中 q_t 是第 t 个单词的特征向量。我们用 q_t^w, q_t^p, q_t^s 分别表示 在位置 t 处的 Word embedding，phrase embedding 以及 question embedding。

图像特征表示为 $V = \{v_1, \dots, v_N\}$，其中，v_n 是空间位置 n 处的特征向量。

图像和问题的 co-attention features 在每一个层次，都可以表示为：$v^\wedge$, $q^\wedge$。

不同模块和层的权重可以表示为 **W**。

[1] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*†（2017）Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

➢ Hierarchical Question-Image Co-Attention [3]

➢ **Model:**

•

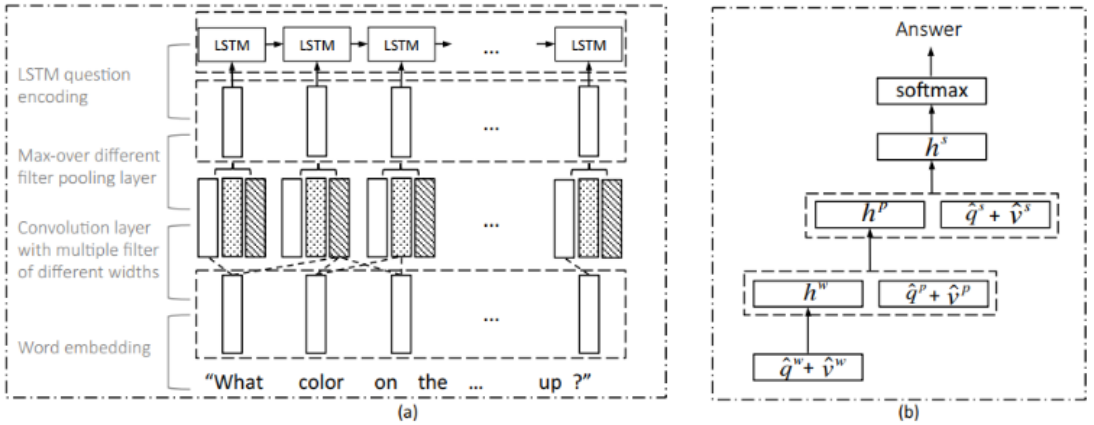


Figure 3: (a) Hierarchical question encoding (Sec. 3.2); (b) Encoding for predicting answers (Sec. 3.4).

2. Question Hierarchy:

给定 the 1-hot encoding of the question words Q, 首先将单词映射到单词空间，以得到：Q^w. 为了计算词汇的特征，我们采用在单词映射向量上采用 1-D 卷积。在每一个单词位置，我们计算 the Word vectors with filters of three window sizes 的内积：unigram, bigram and trigram. 对于第 t 个单词，在窗口大小为 s 时的卷积输出为：

$$\hat{\boldsymbol{q}}_{s,t}^{p} = \tanh(\boldsymbol{W}_c^s \boldsymbol{q}_{t:t+s-1}^{w}), \quad s \in \{1,2,3\} \tag{1}$$

W_c^s 是权重参数。单词级别的向量 Q^w 是 **approximately 0-padding** before feeding into bigram and trigram convolutions to maintain the length of the sequence after convolution. 给定卷积的结果，我们然后 在每一个单词位置，跨越不同的 n-grams 采用 max-pooling 以得到 phrase-level features：

$$\boldsymbol{q}_t^p = \max(\hat{\boldsymbol{q}}_{1,t}^p, \hat{\boldsymbol{q}}_{2,t}^p, \hat{\boldsymbol{q}}_{3,t}^p), \quad t \in \{1,2,\ldots,T\} \tag{2}$$

[1] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*† （2017） Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

➢ Hierarchical Question-Image Co-Attention [3]

➢ Model:

3. Co-Attention:

我们提出两种协同显著的机制（two co-attention mechanism），第一种是 **parallel co-attention**，同时产生 image 和 question attention。第二种是 alternating co-attention，顺序的产生 image 和 question attentions。如图2所示，这些 co-attention mechanisms 可以在所有问题等级上执行。

【Parallel Co-Attention】这种 attention 机制尝试同时对 image 和 question 进行 attend。我们通过计算 图像 和 问题特征在所有的 image-locations and question-locations 进行相似度的计算。具体来说，给定一个图像特征图 V，以及 问题的表达 Q，放射矩阵 （the affinity matrix）C 可以计算如下：

$$\boldsymbol{C} = \tanh(\boldsymbol{Q}^T \boldsymbol{W}_b \boldsymbol{V}) \tag{3}$$

其中，$\mathbf{W}_b$ 包括了权重。在计算得到 affinity matrix 之后，计算 image attention 的一种可能的方法是：simply maximize out the affinity over the locations of other modality, i.e.

$$\boldsymbol{a}^v[n] = \max_i(\boldsymbol{C}_{i,n}) \text{ and } \boldsymbol{a}^q[t] = \max_j(\boldsymbol{C}_{t,j})$$

并非选择 the max activation，我们发现如果我们将这个 affinity matrix 看做是一个 feature，然后学习去预测 image 和 question attention maps 可以提升最终的结果：

$$\begin{aligned} \boldsymbol{H}^v &= \tanh(\boldsymbol{W}_v \boldsymbol{V} + (\boldsymbol{W}_q \boldsymbol{Q})\boldsymbol{C}), \quad \boldsymbol{H}^q = \tanh(\boldsymbol{W}_q \boldsymbol{Q} + (\boldsymbol{W}_v \boldsymbol{V})\boldsymbol{C}^T) \\ \boldsymbol{a}^v &= \text{softmax}(\boldsymbol{w}_{hv}^T \boldsymbol{H}^v), \quad \boldsymbol{a}^q = \text{softmax}(\boldsymbol{w}_{hq}^T \boldsymbol{H}^q) \end{aligned} \tag{4}$$

其中 $\mathbf{W}_v$ 和 $\mathbf{W}_q$，$\mathbf{w}_{hv}$，$\mathbf{w}_{hq}$ 是权重参数。$\mathbf{a}^v$ 和 $\mathbf{a}^q$ 是每一个图像区域 $\mathbf{v}_n$ 和 单词 $\mathbf{q}_t$ 的 attention probability。放射矩阵 **C** 将 question attention space 转换为 image attention space. 基于上述 attention weights，图像 和 问题 attention vectors 可以看做是 image feature 和 question feature 的加权求和：

$$\hat{\boldsymbol{v}} = \sum_{n=1}^{N} a_n^v \boldsymbol{v}_n, \quad \hat{\boldsymbol{q}} = \sum_{t=1}^{T} a_t^q \boldsymbol{q}_t \tag{5}$$

Co-Attention

➢ Hierarchical Question-Image Co-Attention [3]

➢ Model:

- 【Alternating Co-Attention】分步的协同 attention，简单来讲，包括三个步骤：

1）summarize the question into a single vecror **q**;

2）attend to the image based on the question summary **q** ;

3）attend to the question based on the attended image feature.

我们定义 attention operation **x^ = A(X; g)**，将图像特征 X 以及 从问题得到的 attention guidance **g** 作为输入，然后输出 the attended image vector。这些操作可以表达为：

$$
\begin{aligned}
\boldsymbol{H} &= \tanh(\boldsymbol{W}_x\boldsymbol{X} + (\boldsymbol{W}_g\boldsymbol{g})\mathbb{1}^T) \\
\boldsymbol{a}^x &= \text{softmax}(\boldsymbol{w}_{hx}^T\boldsymbol{H}) \\
\hat{\boldsymbol{x}} &= \sum a_i^x \boldsymbol{x}_i
\end{aligned}
\tag{6}
$$

其中，空心符号**1** 是元素全为 1 的向量。

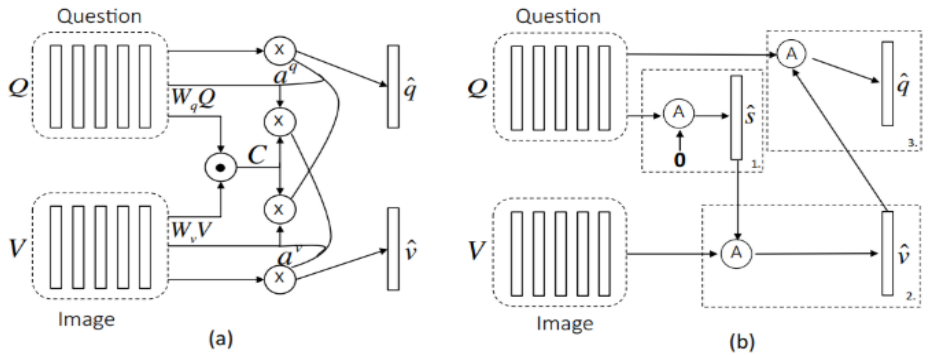


Figure 2: (a) Parallel co-attention mechanism; (b) Alternating co-attention mechanism.

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*†（2017）Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

- **Hierarchical Question-Image Co-Attention [3]**
  - **Results:**

Table 1: Results on the VQA dataset. "-" indicates the results is not available.

	Open-Ended					Multiple-Choice				
	test-dev				test-std	test-dev				test-std
Method	Y/N	Num	Other	All	All	Y/N	Num	Other	All	All
LSTM Q+I [2]	80.5	36.8	43.0	57.8	58.2	80.5	38.2	53.0	62.7	63.1
Region Sel. [20]	-	-	-	-	-	77.6	34.3	55.8	62.4	-
SMem [24]	80.9	37.3	43.1	58.0	58.2	-	-	-	-	-
SAN [25]	79.3	36.6	46.1	58.7	58.9	-	-	-	-	-
FDA [11]	**81.1**	36.2	45.8	59.2	59.5	**81.5**	39.0	54.7	64.0	64.2
DMN+ [23]	80.5	36.8	48.3	60.3	60.4	-	-	-	-	-
Ours^{p}+VGG	79.5	**38.7**	48.3	60.1	-	79.5	39.8	57.4	64.6	-
Ours^{a}+VGG	79.6	38.4	49.1	60.5	-	79.7	**40.1**	57.9	64.9	-
Ours^{a}+ResNet	79.7	**38.7**	**51.7**	**61.8**	**62.1**	79.7	40.0	**59.8**	**65.8**	**66.1**

Table 2: Results on the COCO-QA dataset. "-" indicates the results is not available.

Method	Object	Number	Color	Location	Accuracy	WUPS0.9	WUPS0.0
2-VIS+BLSTM [17]	58.2	44.8	49.5	47.3	55.1	65.3	88.6
IMG-CNN [15]	-	-	-	-	58.4	68.5	89.7
SAN(2, CNN) [25]	64.5	48.6	57.9	54.0	61.6	71.6	90.9
Ours^{p}+VGG	65.6	49.6	61.5	56.8	63.3	73.0	91.3
Ours^{a}+VGG	65.6	48.9	59.8	56.7	62.9	72.8	91.3
Ours^{a}+ResNet	**68.0**	**51.0**	**62.9**	**58.8**	**65.4**	**75.1**	**92.0**

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*† (2017) Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

Multi-level Attention Networks for Visual Question Answering[4]

Background

- effective semantic embedding and fine-grained visual understanding;
- 人类语言问题以明确的查询意图传达强大的高级语义，而具有数万个像素的真实世界图像则相对低级且抽象，由于众所周知的语义差距，这对深度图像理解提出了巨大的挑战；
- 视觉问题回答需要细粒度的空间推理，因为某些答案只能从高度本地化的图像区域推断出“What”和“Where”的问题

Main Work

- 我们通过共同学习multi-level attention来解决自动视觉问题回答的挑战，这可以同时减少从视觉到语言的语义鸿沟，并有益于VQA任务中的细粒度推理；
- 我们引入了一种新颖的视觉注意空间编码方法，通过双向RNN模型从有序图像区域中提取上下文感知视觉特征

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*†（2017）Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

➢ Multi-level Attention Networks for Visual Question Answering[4]

➢ *Model*

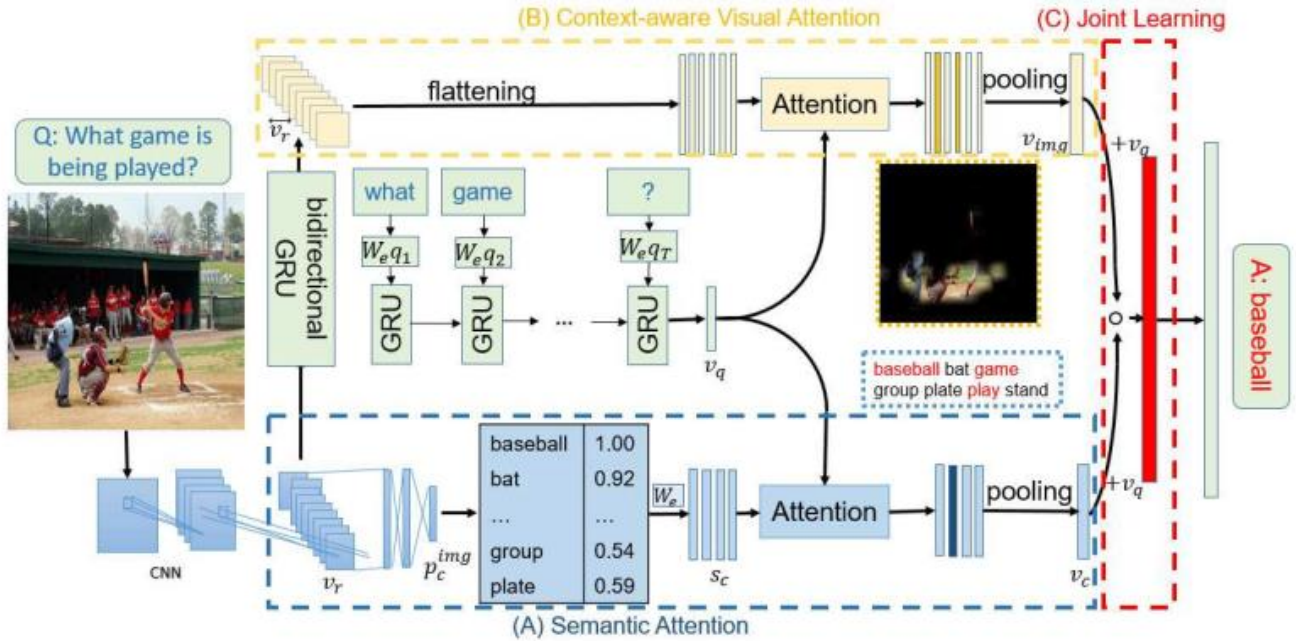


Figure 2. Overall framework of multi-level attention networks. Our framework consists of three components: (A) semantic attention, (B) context-aware visual attention and (C) joint attention learning. Here, we denote by v_q the representation of the question Q, by v_{img}, v_c the representation of image content on the visual and semantic level queried by the question, respectively. v_r and p_c^{img} is the activation of the last convolutional layer and the probability layer from the CNN.

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*† (2017) Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

➢ Multi-level Attention Networks for Visual Question Answering[4]

➢ *Model*

Semantic Attention

1. 通过深度卷积神经网络训练概念检测器，它可以产生图像的语义概念概率；

$$p_c^{img} = f_c(I).$$

2. 训练一个注意力网络来衡量词汇和问题中每个概念之间的语义相关性。
使用以下等式来表示问题编码模型：

$$x_t = W_e^q q_t,$$
$$h_t = GRU(x_t, h_{t-1}),$$
$$v_q = h_T.$$

$$s_c = W_e^c(W_e^q c),$$

$$p_c^q = \text{sigmoid}(v_q \cdot s_c), \quad (6)$$
$$M_c = p_c^{img} p_c^q, \quad (7)$$

where the operator · represents the dot product of two vectors, p_c^q is the relevance score measuring the semantic similarity between the question Q and the concept c, M_c is the semantic attention weights over concepts. Finally, we represent the high-level semantic information of image I queried by question Q by a weighted sum over all concepts representation, which is given by:

$$v_c = \sum_{i=1}^{C} M_c(i) s_c(i). \quad (8)$$

我们对概念和问题使用相同的词汇表和嵌入矩阵，因此它们可以共享相同的语义表示。 具体而言，我们通过双层堆叠嵌入层用语义向量sc表示概念c。 第一层设计为与问题模型共享相同的词嵌入层，第二层用于将概念向量投影到具有问题表示的相同维度中，由下式给出：

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*†（2017）Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Co-Attention

➢ Multi-level Attention Networks for Visual Question Answering[4]

➢ *Model*

Context-aware Visual Attention

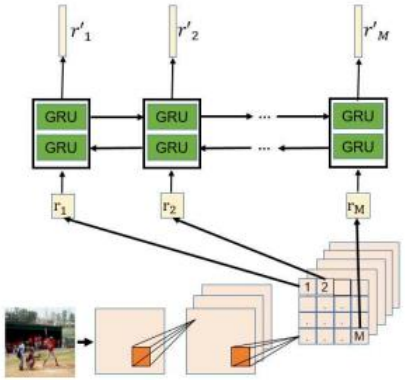


Figure 3. An illustration of the context-aware visual representation for image regions by bidirectional GRU. Regions in convolutional feature maps are encoded into GRU with the order of left-to-right and top-to-bottom.

$$\overleftrightarrow{v_r} = GRU^f(v_r) + GRU^b(v_r) \quad (9)$$

where $\overleftrightarrow{v_r}$ is the context-aware visual representation of image region r. The new feature vectors contain not only the visual information of corresponding regions but also the contextual information from surrounding regions. We set the dimension of the hidden state in each GRU to the same with the question vector.

与通过两个向量的点积测量问题和概念词之间的语义相似性的semantic attention不同，我们对齐问题和每个区域通过两个向量的元素乘法，然后将它们馈送到多层感知器（MLP），这种设计使得能够通过MLP中的参数优化来自动学习attention功能。

1. 使用在上一步中获得的上下文感知视觉特征来表示局部区域，而不是在卷积神经网络中来自每个区域的独立表示，其通常缺乏不同区域之间的相互作用；
2. 我们使用逐元素乘法而不是逐元素加法来对齐每个区域的问题特征和视觉特征，这克服了多模态特征汇集中的尺度不一致问题。

$$h = \tanh\left((W_Q v_q + b_Q) \otimes (W_I \overleftrightarrow{v_r} + b_I)\right),$$
$$M_r = \mathrm{softmax}(W_p h + b),$$

$$v_{img} = \sum_{i=1}^{R} M_r(i) \overleftrightarrow{v_r}(i).$$

在实践中，我们重复上述过程，如[34]中所述，使用问题特征和参加区域特征作为指导，忽略了这里的细节以便简洁。

Joint Attention Learning

我们使用问题作为查询来搜索不同级别的图像信息。 在低级视觉特征中，我们通过visual attention关注与问题相关的区域，而在高级语义特征中，我们通过semantic attention关注与问题相关的概念。 两个级别的注意力通过融合他们的代表性表达而结合在一起。 特别是，我们首先将问题向量添加到从不同层提取的 attended image features，然后我们使用逐元素乘法将两种类型的注意力组合在一起。最后，我们将关节特征馈送到softmax层以预测预定义候选答案集A的概率。具有最高概率的候选者被确定为最终答案，其由下式给出：

$$u = (v_q + v_{img}) \circ (v_q + v_c),$$
$$p_a = \mathrm{softmax}(Wu + b),$$

[3] Jiasen Lu* , Jianwei Yang* , Dhruv Batra*† , Devi Parikh*† （2017）Hierarchical Question-Image Co-Attention for Visual Question Answering. In NIPS

Recursive-Attention

Recursive Visual Attention in Visual Dialog[5]

- *Model*

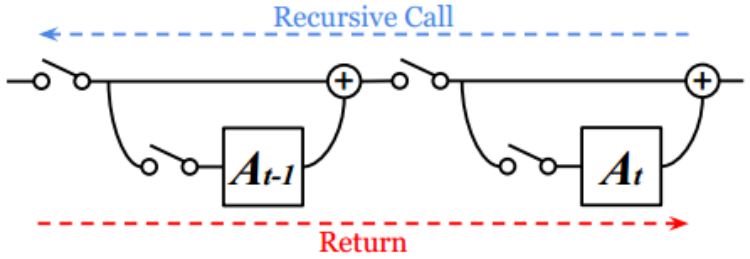


Figure 2. A high-level view of Recursive Visual Attention. The right-to-left direction (dashed blue) represents recursive call, and the left-to-right direction (dashed red) represents visual attention return. The *cond* variable controls the switch on the trunk, while λ controls the switch on the branch (see Algorithm 1). A_t represents the attended feature $\text{ATT}(\mathcal{V}, \mathcal{Q}, t)$.

Algorithm 1 Recursive Visual Attention

function $\text{RVA}(\mathcal{V}, \mathcal{Q}, \mathcal{H}, T, t)$
  $cond, \lambda, T \leftarrow \text{INFER}(\mathcal{Q}, \mathcal{H}, T, t)$
  if *cond* **then**
    return $\lambda \text{ATT}(\mathcal{V}, \mathcal{Q}, t)$
  else
    return $\text{RVA}(\mathcal{V}, \mathcal{Q}, \mathcal{H}, T, t-1) + \lambda \text{ATT}(\mathcal{V}, \mathcal{Q}, t)$
  end if
end function

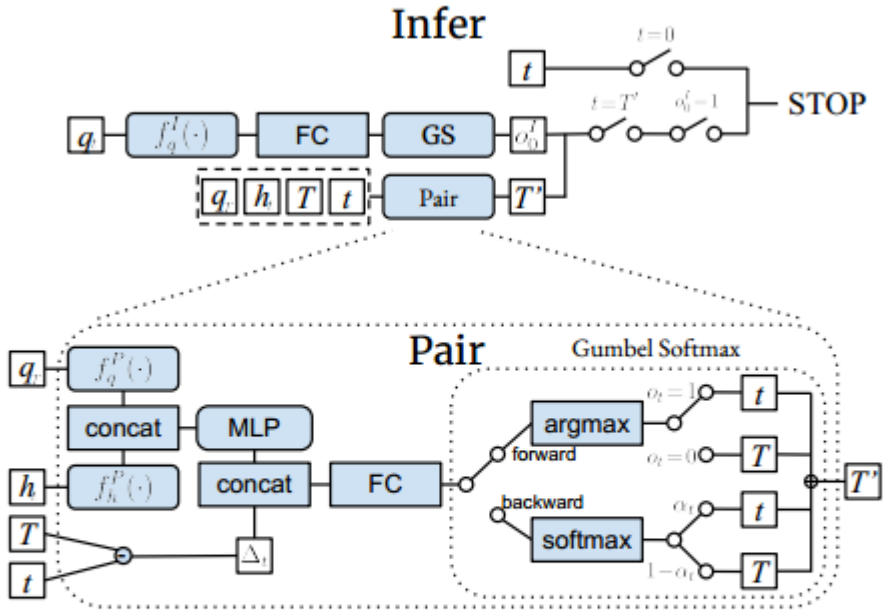


Figure 3. The flowchart of INFER and PAIR modules. We only show the computation of recursion termination as an example for simplicity and clarity. GS: Gumbel softmax. MLP: multilayer perceptron. FC: fully connected layer. $f(\cdot)$: non-linear transformation.

[5] Yulei Niu1 Hanwang Zhang2 Manli Zhang1 Jianhong Zhang1 Zhiwu Lu1 Ji-Rong Wen （2019）Recursive Visual Attention

Recursive-Attention

Recursive Visual Attention in Visual Dialog[5]

- *Result*

Table 1. Retrieval performance of discriminative models on the validation set of VisDial v0.9. RPN, Rv and FL indicate the usage of region proposal network, recursive image attention, and visual feature filter, respectively.

Model	MRR	R@1	R@5	R@10	Mean
LF [9]	0.5807	43.82	74.68	84.07	5.78
HRE [9]	0.5846	44.67	74.50	84.22	5.72
HREA [9]	0.5868	44.82	74.81	84.36	5.66
MN [9]	0.5965	45.55	76.22	85.37	5.46
HCIAE [20]	0.6222	48.48	78.75	87.59	4.81
AMEM [27]	0.6227	48.53	78.66	87.43	4.86
CoAtt [33]	0.6398	50.29	80.71	88.81	4.47
CorefNMN [17]	0.641	50.92	80.18	88.81	4.45
RvA w/o RPN	0.6436	50.40	81.36	89.59	4.22
RvA w/o Rv	0.6551	51.81	82.35	90.24	4.07
RvA w/o FL	0.6598	52.35	82.76	90.54	3.98
RvA	**0.6634**	**52.71**	**82.97**	**90.73**	**3.93**

Table 2. Retrieval performance of discriminative models on the test-standard split of VisDial v1.0. † indicates that the model uses ResNet-152 features.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF [9]	0.5542	40.95	72.45	82.83	5.95	0.4531
HRE [9]	0.5416	39.93	70.45	81.50	6.41	0.4546
MN [9]	0.5549	40.98	72.30	83.30	5.92	0.4750
CorefNMN† [17]	0.615	47.55	78.10	88.80	4.40	0.547
RvA w/o RPN	0.6060	46.25	77.88	87.83	4.65	0.5176
RvA w/o Rv	0.6226	47.95	79.75	89.08	4.37	0.5319
RvA w/o FL	0.6294	48.68	80.18	89.03	4.31	0.5418
RvA	**0.6303**	**49.03**	**80.40**	**89.83**	**4.18**	**0.5559**

Table 3. Retrieval performance of generative models on the validation set of VisDial v0.9. ‡ indicates that the model is trained using reinforcement learning.

Model	MRR	R@1	R@5	R@10	Mean
LF [9]	0.5199	41.83	61.78	67.59	17.07
HRE [9]	0.5237	42.29	62.18	67.92	17.07
HREA [9]	0.5242	42.28	62.33	68.17	16.79
MN [9]	0.5259	42.29	62.85	68.88	17.06
CorefNMN [17]	0.535	43.66	63.54	69.93	15.69
HCIAE [20]	0.5386	44.06	63.55	69.24	16.01
CoAtt [33]	0.5411	44.32	63.82	69.75	16.47
CoAtt‡ [33]	**0.5578**	**46.10**	**65.69**	71.74	14.43
RvA w/o RPN	0.5417	43.75	64.21	71.85	11.18
RvA w/o Rv	0.5523	45.15	65.06	72.87	**10.64**
RvA w/o FL	0.5547	45.43	65.24	72.92	10.67
RvA	0.5543	45.37	65.27	**72.97**	10.71

[5] Yulei Niu1 Hanwang Zhang2 Manli Zhang1 Jianhong Zhang1 Zhiwu Lu1 Ji-Rong Wen （2019）Recursive Visual Attention