Cross-modal Relational Reasoning Network for Visual Question Answering

Hongyu Chen Ruifang Liu*
Beijing University of Posts and Telecommunications
18801245963@163.com lrf@bupt.edu.cn

Bo Peng
Tencent
Boarpeng@tencent.com

Abstract

Visual Question Answering (VQA) is a challenging task that requires a cross-modal understanding of images and questions with relational reasoning leading to the correct answer. To bridge the semantic gap between these two modalities, previous works focus on the word-region alignments of all possible pairs without attending more attention to the corresponding word and object. Treating all pairs equally without consideration of relation consistency hinders the model's performance. In this paper, to align the relation-consistent pairs and integrate the interpretability of VQA systems, we propose a Cross-modal Relational Reasoning Network (CRRN), to mask the inconsistent attention map and highlight the full latent alignments of corresponding word-region pairs. Specifically, we present two relational masks for inter-modal and intra-modal highlighting, inferring the more and less important words in sentences or regions in images. The attention interrelationship of consistent pairs can be enhanced with the shift of learning focus by masking the unaligned relations. Then, we propose two novel losses $\mathcal{L}_{CMAM}$ and $\mathcal{L}_{SMAM}$ with explicit supervision to capture the fine-grained interplay between vision and language. We have conduct thorough experiments to prove the effectiveness and achieve the competitive performance for reaching 61.74% on GQA benchmark.

1. Introduction

Recently, with developments in deep learning models, we have witnessed great progress in both Computer Vision and Natural Language Processing (NLP). As cross areas between vision and language, many multi-modal learning tasks, such as image captioning, image-text matching, and visual question answering (VQA) have received increasing attention from the research community. Compared with other multi-modal tasks, VQA [4] needs to predict the correct answer when giving an image and a related question, which requires not only fine-grained semantic understanding of texts and images but also relational reasoning.

Extensive flexible learning and reasoning methods are built to tackle the VQA problem [3, 48]. A popular framework for VQA first represents images and questions as global features, then fuses them into a common space by bilinear fusion methods [9, 49, 5, 10, 23]. The fused features are fed into a classifier for answer prediction [3]. Attention mechanisms [47] are widely used to capture fine-grained cross-modal relations. Based on obtained relations, various reasoning methods like MAC [17] and Probabilistic Neural-symbolic Models [45] are proposed to execute sequential reasoning for the final answer.

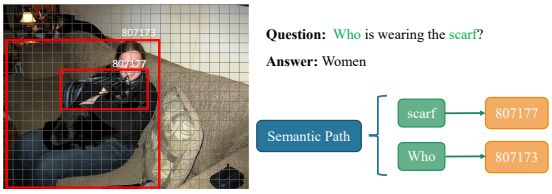

Figure 1. An example for visual question answering. The semantic path is provided by the GQA dataset. The picture is segmented into grids and represented by grid features. Regions with numbered red bounding boxes are related to green words in the question.

The work mentioned above has been proved to be effective and bridge the gap between perception and cognition. However, prior works mainly focus on the process of reasoning but ignore the importance of relations learning across different modalities. Superior fine-grained relations can boost the performance of the reasoning block. Common attention mechanisms for VQA task helps to learn the interrelationship between every possible word-region pairs without consideration of original explicit alignments between question words and image regions.

Intending to capture more accurate relations across visual and textual modalities and improve the performance of the reasoning process, we propose the Cross-modal Relational Reasoning Network (CRRN). Specifically, different from the common attention-based approaches, we additionally utilize the correspondence relations between questions and images to guide the attention map of cross-modal

and single-modal to be more accurate. Meanwhile, the explicit correspondence relations make the deep neural network more transparent and interpretable. Furthermore, explainable models such as NSM [19] rely on the probabilistic graph. It's a strong prior structured knowledge from the pre-trained scene graph generation model. Instead, we only utilize the coarse explicit alignments between objects and the corresponding words in sentences. It reduces the complexity of the pipeline. It should be noted that these coarse alignments are easy to obtain, either provided by the dataset or parsed by deep learning tools. As shown in Figure 1, the GQA dataset provides the alignments in the semantic path. The more fine-grained correspondence relations between visual and language features will be further extracted. Two relational masks named inter-modal mask and intra-modal mask are designed to highlight these consistent relations and benefit to the inter-modal and intra-modal interactions. Specifically, in the process of inter-modal interaction, we use the text feature to guide the fine tune of the image feature. Intuitively, the image feature should pay more attention to words that have a consistent relation with it. Therefore, the inter-modal mask is used to highlight the corresponding words in the question. In the single visual modal for image representation learning, we use an intra-modal mask to infer the more critical image regions, which helps to attend to more question-related visual features.

In this paper, a novel relational reasoning network is proposed in which the consistent relations across the two modalities are emphasized for better relations learning and relational reasoning. The proposed CRRN improves the accuracy of answer prediction while integrating the transparency and interpretability of the VQA system. The contributions can be summarized as follows:

(1) Two relational masks are designed to highlight correspondence relations and eliminate the interference information across the two modalities.

(2) A neural network is trained with novel objective functions to capture fine-grained relations with supervision.

(3) We improve relational reasoning by learning better cross-modal relations and achieve competitive performance on the GQA benchmark.

2. Related Work

2.1. Relational Reasoning

Relational reasoning tries to solve VQA by learning the relationships between individual visual regions and words [27, 12, 8, 33, 46, 25]. Image and text both contain rich information but reside in heterogeneous modalities. The designed models for the cross-modal task need not only to learn the features for images and texts to express their respective contents but also the correspondences between the detected visual objects and the textual items (words or phrases). Many studies have validated that the correspondences are helpful to model a more reliable relationship between images and texts [27, 35, 16, 3]. Recently, some studies explored to utilize semantic-enhanced strategies to learn the visual-semantic correspondence. Qi *et al*. [39] constructed pairwise combinations between regions and words to represent the correlations. They utilized KNN method to model these correlations for learning visual-semantic alignments. Huang *et al*. [16] used a multi-regional multi-label CNN to extract semantic concepts, and then used images and semantic concepts to generate the sentence representation. Hudson and Manning [19] proposed a model called Neural State Machine (NSM) for the visual questions that need compositionality and multi-step inference. A probabilistic graph is first predicted as a structured semantic representation of the image. Then, NSM executes sequential reasoning guided by the input question over the predicted graph. The proposed model achieves state-of-the-art results on VQA-CP [1] and GQA [18] datasets.

2.2. Attention Models for VQA

Co-attention based methods and enhanced embedding based methods, that are relevant to our work, will be briefly introduced for visual question answering. Exploring the relationship between a given image and a related question contributes to reasoning the right answer, which has been of key interest in the VQA task over the past few years. The mainstream method for this is multi-modal fusion. Images and questions are represented as global features [50], later fused to a common space by some effective multi-modal fusion methods [22, 9] for the right answer representation and prediction. To align the key part between textual and visual contents, a large amount of co-attention based methods are proposed for VQA [21, 48, 36]. Lu *et al*. [9] proposed a co-attention model to jointly reason for images and questions. Guo *et al*. [13] utilized the information in answer to re-attends the corresponding visual objects in images. Yu *et al*. [49] reduced the co-attention method into two steps, self-attention for a question embedding and the question-conditioned attention for a visual embedding. Nam *et al*. [34] proposed a multi-stage co-attention learning model to refine the attention based on the memory of previous attentions. Yu *et al*. [48] proposed a deep Modular Co-Attention Network to conduct dense interactions between each question word and each image region. Beyond the alignment of important objects and words between images and questions, VQA also requires full understandings of the contents of each modality. Some enhanced embedding based methods are proposed to add some complementary information to the original contents. Hu *et al*. [15] proposed a relation-wise dual attention network to extract the implicit connections between salient objects. Liu *et al*. [30] represented an image with a set of integrated visual regions and corre-

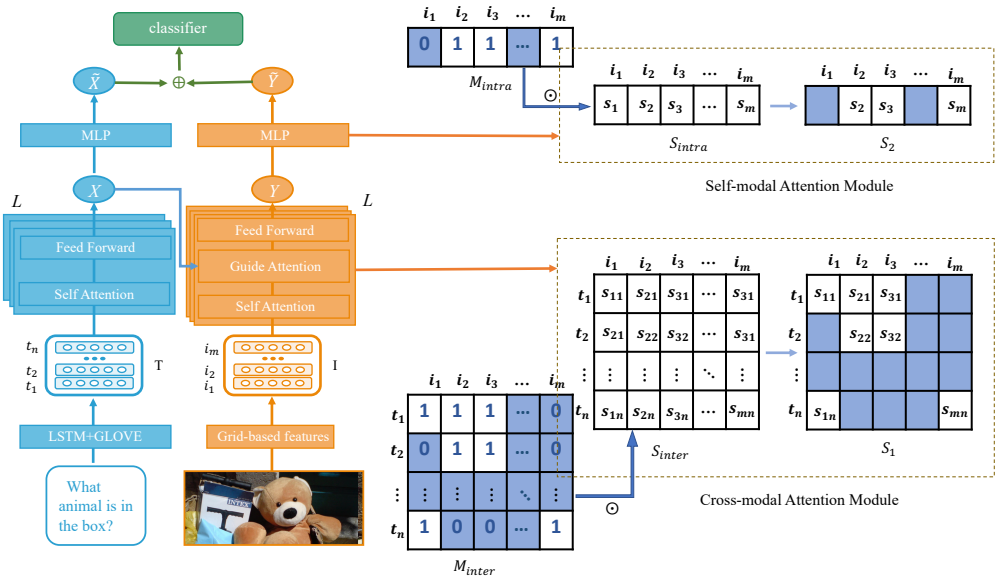

Figure 2. Overview of the proposed CRRN. The blue and orange blocks in the left of the figure are transformer-based encoders for questions and images, respectively. Two attention maps S_{inter} and S_{intra} indicating the latent relationship across two modalities and the single visual modality are obtained at the process of model training. Two relational masks M_{inter} and M_{intra} for inter-modal and intra-modal are adopted to highlight the interrelationship of word-image region pairs with explicit alignments.

sponding textual concepts.

2.3. Transformer-based Architecture for VQA

Transformer [44] is an encoder-decoder structure, which is formed by stacking several encoders and decoders. It consists of a self-attention module to learn the latent relationship among words and capture the internal structure of the sentence. BERT [7], evolved from the encoder of the transformer, has achieved great success in the field of language understanding. Inspired by the success of BERT, several transformer-like models [32, 31, 29, 2, 28, 42, 41] accompanied with large scale pretraining tasks (*e.g.* masked language modeling and masked visual-feature classification) have been used in cross-modal tasks such as VQA. ViLBERT [31] and LXMERT [42] apply a single-modal transformer to the image and sentence respectively, and then use a cross-modal transformer to combine the two modalities. VisualBERT [29], and Unicoder-VL [28] concatenate images and sentences into a single input of the transformer.

In addition, some individual modules in the transformer are also applied in cross-modal tasks. Yu *et al*. [48] and Peng *et al*. [37] used self-attention for each modality of vision and language. Then co-attention is adopted for modality fusion. $MCAN$ [48] used an encoder-decoder structure, performing self-attention for language modality as an encoder, followed by a decoder, where self-attention and co-attention for vision modality are used. By contrast, Peng *et al*. [37] stacked several layers. In each layer, co-attention is used first, followed by self-attention for each modality.

3. Method

In this section, we propose a novel framework CRRN for VQA which is depicted in Figure 2. The left is the common pipeline of Transformer-like architectures, where we briefly review in section 3.3. In the right of the figure, we depict two sub-modules closely related to the proposed method which will be introduced thoroughly in section 3.4.

3.1. Input Representation

Following recent approaches, grid features [20] from the pre-trained detector are the main choice for visual features, which make the model design and training process much simpler and perform competitively against their region-based counterparts. Every image is represented by a set of features $\{i_1, i_2, ..., i_m\}$, where m is the number of features and the dimension of each feature is set as d_v. The bounding box related to each feature is represented as the coordinates in the upper left and lower right corner. Each item in the bounding box $(x_{min}, y_{min}, x_{max}, y_{max})$ is a pixel value. We first normalize these values by the height and width of the image for numerical stability and then add their areas as a new feature. The obtained feature is $(\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, Area)$. We embed the 5-d feature to a high-dimensional representation $P \in \mathbb{R}^{m \times d_v}$ as positional features. The visual features and positional features are concatenated together and passed through a linear

layer. The output is the final image representation.

As for textual features, we trim all questions to a maximum length of n. Each word in the question is embedded into a 300-dimensional Glove Vector [38], which is pretrained on a large scale dataset. Then word embeddings $\{t_1, t_2, ..., t_n\}$ where $n \in [1, 29]$ is the length of the question, are passed through LSTM network with d_h hidden units. Finally, we utilize the output features of all words to obtain a question representation $T \in \mathbb{R}^{n \times d_h}$.

3.2. Encoders and Attention Method

We adopt two transformer encoders for each modality. As shown in Figure 2, each layer in the language encoder consists of a multi-head self-attention and a feed-forward sub-layer. For the image encoder, it should be noted that guide-attention is added after the self-attention. It uses the output of the previous layer and the textual features as input. The co-attention mechanism is achieved in this sub-layer and the textual content is used to guide the fine-tuning of every visual feature. More specifically, as shown in the scaled dot-product attention mechanism [44]:

$$\mathbb{A}(Q, K, V) = softmax(\frac{QK^\top}{\sqrt{d_k}})V, \tag{1}$$

where $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{m \times d_k}$, $V \in \mathbb{R}^{m \times d_k}$ mean a set of query, key, and value vectors. n and m are the numbers of query, key vectors, and d_k is the dimension of the key vectors. Each visual feature is used as a query, and all the token-level features in the question are a set of key vectors. After getting the score of the current visual feature and each question token feature, the visual feature is represented as the weighted sum of these token-level features. We simply stack L layers in-depth for each encoder. Then we adopt Multi-Layer Perception (MLP) on the outputs X and Y from encoders to obtain aggregated representations of the whole questions and images represented as $\widetilde{X}$ and $\widetilde{Y}$. Following Anderson *et al*. [43], we view VQA as a multi-label classification task. So, we apply an element-wise product on $\widetilde{X}$ and $\widetilde{Y}$ and get the fused features $Z \in \mathbb{R}^D$, where D is the dimension of Z. The fused features are fed into weight-sharing classifier $W \in \mathbb{R}^{D \times C}$ with a sigmoid function to predict score $\hat{s} \in C$ for each candidate answer:

$$\hat{s} = sigmoid(W^\top Z). \tag{2}$$

Then binary cross-entropy based loss [43] is employed for classification as our loss function:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{M}\sum_{j=1}^{N} s_{ij} log(\hat{s}_{ij}) - (1 - s_{ij}) log(1 - \hat{s}_{ij}), \tag{3}$$

where M and N indicate the number of training questions and candidate answers. s and $\hat{s}$ are ground truth and predicted scores respectively.

3.3. Extraction of Relational Masks

As is depicted in Figure 1, GQA provides semantic paths to highlight some words and the location of corresponding image regions for answering a visual question. For example, when asking 'Who is wearing the scarf' according to the figure below, the highlighted words are 'scarf' and 'who'. The corresponding image regions marked by the red bounding box are numbered as '807177' and '807173', respectively. We utilize the coarse explicit alignment such as ('scarf', '807177') and ('person', '807173') to explore more fine-grained consistent relations between the textual and visual features.

Specifically, given an image I with m grid features and a question T with n words, we first extract a set of coarse alignments in the semantic path. For one of the alignments represented as (t_i, B), t_i indicates the i_{th} word in the question and B indicates the bounding box of the corresponding image regions annotated manually. We can obtain an inter-modal consistent relational mask $M_{inter} \in \mathbb{R}^{n \times m}$ defined as the following way:

$$M_{inter}(ij) = \begin{cases} 0 & P(B, b_j) = 0, \\ 1 & P(B, b_j) > 0, \end{cases} \tag{4}$$

where 1 and 0 represent that there is with and without a relation between t_i and j_{th} grid feature i_j. P is a function to compute the value of intersection-over-union (IoU) between the ground truth bounding box B and the bounding box b_j of grid feature i_j. Therefore, $M_{inter}(ij) = 1$ indicates the grid has an intersection with B and it is reasonable to assume that the grid feature i_j has a consistent relationship with the word t_i. If a word in the question does not present in the semantic path, it has no explicit relation with any of the grid features. M_{inter} represents the fine-grained relation between any word and grid feature. Meanwhile, to further highlight the important grid features, an intra-modal relational mask $M_{intra} \in \mathbb{R}^m$ is designed as follows:

$$M_{intra}(j) = \begin{cases} 0 & M_{inter}(:, j) = 0, \\ 1 & else, \end{cases} \tag{5}$$

where $M_{inter}(:, j) = 0$ indicates for every word t_i in the question, $M_{inter}(ij) = 0$. Instead, if the grid feature has a consistent relation with word t_i ($M_{inter}(ij) = 1$), it should be attached more importance and marked as '1' in M_{intra}.

3.4. Relational Learning with Masks

To enhance the attention interrelationship of textual and visual pairs with consistent relations, we use the relational masks in section 3.3 as supervision. We will introduce the Cross-modal Mask Attention Module (CMAM) for intra-modal and Self-modal Mask Attention Module (SMAM) for inter-modal. The two modules help the VQA system

capture important contents in two modalities with the shift of learning focus. Specifically, we use the fine-grained consistent relations to guide the model to attend to the related image regions or question words by enlarging their attention weights.

$CMAM$ As illustrated in Figure 2, after L stacking layers of image encoder, we get the final attention map about every visual and textual feature pair named as $S_{inter} \in \mathbb{R}^{k \times m \times n}$. k is the number of heads in multi-head attention, m and n are the number of visual features and word tokens in the question-image pair. Generally, the attention map demonstrates the latent relationship across two modalities learned by the model. To enhance the consistent relations of corresponding word-region pairs and to make the model more transparent, the inter-modal relational mask M_{inter} is applied to mask the inconsistent attention weights in S_{inter}. We additionally define an auxiliary objective function $\mathcal{L}_{CMAM}$ to help to train the whole model.

$$\begin{aligned} \mathcal{L}_{CMAM} &= \sum_{i=1}^{p} max(\beta_1 - max(S_1), 0), \\ S_1 &= S_{inter} \odot M_{inter}, \end{aligned} \tag{6}$$

where p is the total number of training samples (image-question pairs) within a semantic path in the dataset and β_1 is the threshold of the maximum value in S_1.

After the masking operation is adopted, we can obtain a new version of attention map S_1. The attention weights in S_1 imply the interrelationship of visual and textual pairs with consistent relations. The maximum value of S_1 indicates the network's ability to model these alignments. To extend the degree of this ability, we employ a hyper-parameter β_1 to be the upper bound and encourage the model to approach it based on the idea of hinge loss. specifically, when the maximum value is larger than β_1, the model learns the relations thoroughly and there is no need to optimize, thus the loss is 0. On the contrary, if the model is not well learned, the gap between the threshold and the maximum value of S_1 will be utilized to conduct optimization.

$SMAM$ Similar to the $CMAM$, the model is expected to learn an accurate attention distribution within the single visual modality, and the image regions within consistent relations are expected to be captured. Specifically, the final image visual features Y are obtained at the last layer of the image encoder. Then we conduct a two-layer MLP for Y to get the attended features $\widetilde{Y}$:

$$\begin{aligned} S_{intra} &= softmax(MLP(Y)), \\ \widetilde{Y} &= \sum_{i=1}^{n} (S_{intra})_i y_i, \end{aligned} \tag{7}$$

where $S_{intra} \in \mathbb{R}^{n \times 1}$ is the learned attention map about visual features. To highlight the important visual features in the attention map, the intra-modal relational mask M_{intra} is adopted, and the corresponding objective function L_{SMAM} is designed for the model training.

$$\begin{aligned} \mathcal{L}_{SMAM} &= \sum_{i=1}^{p} max(\beta_2 - \sum_{j=1}^{q} S_2, 0), \\ S_2 &= S_{intra} \odot M_{intra}, \end{aligned} \tag{8}$$

where q is the number of visual features within relation in M_{intra}. β_2 indicates the threshold of the sum of S_2.

As the definition of softmax function, it's easy to know that the sum of the original visual attention map S_{intra} equals 1. After masking the irrelevant attention weights by M_{intra}, the sum of S_2 indicates how much attention the model assigns to those visual features within a consistent relation to the question words. The sum should be a large value and close to 1 if the model captures most of the relative visual features. Therefore, similar to $\mathcal{L}_{CMAM}$, $\mathcal{L}_{SMAM}$ employs a hyper-parameter β_2 to extend the model's learning capacity to the maximum degree. Only when the sum of S_2 is smaller than β_2, the loss function $\mathcal{L}_{SMAM}$ works. finally, the integrated loss function $\mathcal{L}$ is as follows:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{CE} + \gamma_2 \mathcal{L}_{CMAM} + \gamma_3 \mathcal{L}_{SMAM}, \tag{9}$$

where γ_1, γ_2 and γ_3 are the parameters to combine three loss functions.

4. Experiments

4.1. Dataset

We evaluate the performance of our proposed method on the common benchmark GQA [18]. GQA covers 113,018 photo-realistic images. The questions are divided into five different types including Choose, Logical, Compare, Verify, and Query. The dataset is popular with deep reasoning tasks: over 85% of questions with 2 or 3 reasoning steps and 8% of questions with 4+ reasoning steps. The GQA is also annotated with scene graphs extracted from the Visual Genome [26] and functional programs that specify reasoning operations for each pair of image and question. The task of GQA is the same as VQA (i.e., answer single-image related questions), but GQA requires more reasoning skills including spatial reasoning, relational reasoning, logic, and comparisons. 22M questions in the dataset are generated from the ground truth image scene graph to explicitly control the quality of questions. We use the more common "balanced" version that has been designed to reduce biases within the answer distribution (similar in motivation to the VQA2 dataset [11]) and includes 1.7M questions split into 70%/10%/10% for training, validation, and test sets, respectively. We adopt the standard accuracy metric and the more

detailed type-based diagnosis supported by GQA to evaluate the results.

4.2. Experimental Setup

Our framework is implemented using PyTorch and trained with Adam optimizer [24] on 4 NVIDIA GTX 1080ti. We use Ubuntu 16.04.6 LTS with a CPU @ 3.70GHz, and the total memory is 32GB.

The hyper-parameters of our model used in the experiments are set as follows. The number of words in questions is padded to 29. For visual features, the region-based features for comparison are from a ResNet-101 [40] model and the grid features are from a ResNeXt-152 [14] model. Both models are pre-trained on the Visual Genome dataset [26]. The number of the region-based and grid-based features are set to 100 and 510 respectively. The dimension of input object features, input word features, and fused multimodal features are 2048, 512, and 1024. And the number of attention heads is 8. We follow the suggestions in official GQA guidelines to take testdev as our validation set. For the results on testdev split, we train the model on train split with the latent dimension of the multi-head attention being set to 512; for the result on GQA test split, we train the model on the train and val splits with a larger latent dimension of 1024. The threshold β_1 in CAM is set to a series of constants and a variable relevant to the length of the question. The other threshold β_2 mentioned in SMAM in sub-section 3.4 are set to higher values which range from 0.6 to 1. For the final combined loss function, as is described in Equation 9, γ_1 and γ_3 are both set to 1 and γ_2 changes from 0 to 0.001 during the training process. The best base learning rate for SAM_{inter} and SAM_{intra} are set to 2.5te$-$5 and 1e$-$4, where t is the current epoch number starting from 1. All the models are trained up to 11 epochs with the same batch size 64. For single-model settings, to have a fair comparison, we consider all models are similar to ours. They did not use the ensemble submissions to the GQA challenge and a much stronger prior structured knowledge such as scene graphs of images.

To gain further insight into the relative contributions of different aspects of our model to its overall performance, we conducted multiple ablation experiments. Specifically, we validate the importance of using grid features and two attention sub-modules. All the results reported in our ablation study are based on a small model architecture with 8 multi-head attention heads and 64-dimensionality of each head. To reduce time consumption, Only train split was used for training and test-dev split was used for validation.

4.3. Ablation Study on GQA Testdev

To gain further insight into the relative contributions of different aspects of our model to its overall performance, we conducted multiple ablation experiments. Specifically, we validate the importance of using grid features and two attention sub-modules. All the results reported in our ablation study are based on a small model architecture with 8 multi-head attention heads and 64-dimensionality of each head. To reduce time consumption, Only train split was used for training and the test-dev split was used for validation.

Visual features	accuracy(%)
frcn	51.49
frcn+fr-bbox	56.28
grid	58.35
grid+gr-box	**59.34**

Table 1. Ablation studies on various visual features on the GQA testdev split. 'frcn' indicates region-based features named faster-rcnn features, 'fr-bbox' and 'gr-box' indicate bounding box features for fatser-rcnn and grid features, respectively.

Method	Binary	Open	Acc
base	78.62	42.98	59.34
base + CMAM/β_1=0.2	78.81	43.56	60.03
base + CMAM/β_1=0.4	78.99	44.03	**60.16**
base + CMAM/β_1=0.6	78.76	43.04	59.44
base + CMAM/β_1=0.8	78.57	43.32	59.50
base + CMAM/β_1=1.0	78.49	43.38	59.49
base + CMAM/β_1=1.0-0.033*len	78.76	42.63	59.21
base + SMAM/β_2=0.6	78.93	43.92	60.12
base + SMAM/β_2=0.8	78.86	44.38	**60.33**
base + SMAM/β_2=1.0	79.41	43.52	60.07
base + C-SMAM/c=0	**79.53**	43.48	60.03
base + C-SMAM/c=8	78.85	44.28	60.30
base + C-SMAM/c=10	78.97	**44.71**	**60.58**

Table 2. Ablation studies on CAM and SAM applied to the baseline. 'Acc' indicates the overall accuracy. 'base' indicates baseline with grid features. 'CMAM', 'SMAM' and 'C-SMAM' indicate the cross-modal mask attention module, the self-modal mask attention module, and the combination of both, respectively.

The effectiveness of different visual features As shown in Table 1, different visual features have different impacts on the same baseline. First, compared with the widely used bottom-up region features, grid features help the GQA boost the accuracy by 6.86% (row 1 & 3). The relevant bounding box information of grid features helps to slightly improve the performance by 0.99% (row 3 & 4), which is in contrast with great increments of 4.79% (row 1 & 2) in region-based features. The reason for the difference may be that the bounding boxes of region-based features can bring more information on spatial relations. While the order of grid features sequence naturally encodes the relative position of the corresponding image region. Thus, we use visual features of 2^{th} row (best for region-based) and 4^{th} row (best for grid) to perform a more in-depth study. A fair compar-

Model	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human [18]	91.20	87.40	98.40	98.80	97.20	-	89.30
Global Prior [18]	42.94	16.62	51.69	88.86	74.81	93.08	28.90
Local Prior [18]	47.90	16.66	54.04	84.33	84.31	13.98	31.24
BottomUp [3]	66.64	34.83	78.71	96.18	84.57	5.98	49.74
MAC [17]	71.23	38.91	81.59	96.16	84.48	5.34	54.06
GRN	77.53	43.35	88.63	96.18	84.71	6.06	59.37
Dream	77.84	43.72	91.71	96.38	85.48	8.40	59.72
LXRT [42]	77.76	44.97	92.84	96.30	85.19	8.31	60.34
MetaNetwork [6]	78.90	44.89	92.49	96.19	84.55	5.54	60.83
SK T-Brain*	79.12	44.76	92.61	96.35	85.63	8.56	60.87
CRRN	**77.89**	**45.12**	**93.75**	96.25	85.11	5.44	**61.74**

Table 3. GQA scores for the single-model settings, including official baselines and some state-of-the-art submissions

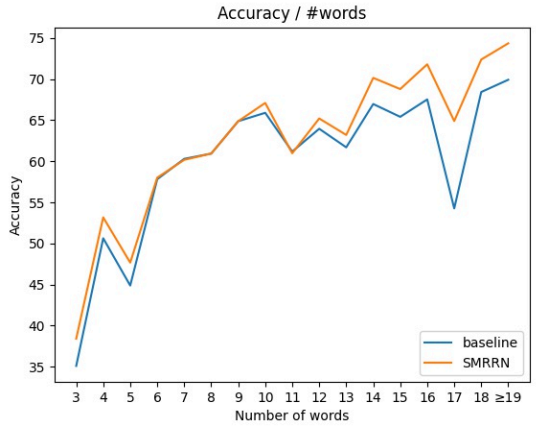

Figure 3. Accuracy as a function of the number of question words. Performance is reported on the GQA Test-Dev split.

ison between the two will be conducted through the rest of the ablation study.

The effectiveness of $CMAM$ Since we have verified the superiority of grid features over region-based features on the baseline, for time consideration, we evaluate the effectiveness of $CMAM$ and $SMAM$ only based on the grid features. Different thresholds β_1 represent different levels of restrictions to the maximum value of attention scores related to textual-visual consistent relations. 'len' indicates the length of the question in a sample pair. Thus, we analyze the performance of $CMAM$ when β_1 is with different settings. In addition to a constant value, with an assumption that the maximum scores may be influenced by the length of questions, we also design a linear function as a threshold that uses the number of question words as a factor of penalty. We observe that with a fixed value of β_1 (β_1=0.4) to the attention scores, the $SMAM$ has a better performance. Using the threshold varied with the question length leads to a slight drop of 0.13% on the overall accuracy. Adding $CMAM$ leads to higher overall accuracy for grid features (row 1 & 3). The results indicate that it is efficient to emphasize the consistent relations between visual and textual features, helping to model a more accurate fine-grained interaction between the two modalities.

The effectiveness of $SMAM$ As shown in Table 2, compared with the baseline, additionally adding $SMAM$ and set β_2 to 0.8 achieves 0.99% improvements on the overall accuracy (row 1 & 9). It can be observed that when β_2 is set to 1.0, the overall accuracy is degraded from 60.33 to 60.07 (row 9 & 10). In this case, the model learns an attention map that only focuses on visual features having explicit consistent relations with textual features. However, some visual features may have implicit alignments with the textual contents and can not be ignored when enhancing the alignments of corresponding word-region pairs with explicit supervision.

The combination of $CMAM$ **and** $SMAM$ Using the best settings of $CMAM$ and $SMAM$, we combine both modules to see overall performance. Surprisingly, it leads to a drop of 0.3% accuracy when combining the $CMAM$ at the beginning of the training process compared to the result of only using $SMAM$ (60.03 $vs.$ 60.33). It is partly due to that the suitable learning rates for the $CMAM$ and $SMAM$ are different. The learning rate for the former is smaller. The learning rate has experienced two decays in our experiment (epoch 8 and 10). Therefore, we conduct experiments to join the $CMAM$ with the $SMAM$ at different training epochs. As shown the bottom three rows in Table 2, c is the epoch SAM_{inter} begins to join the training and the weight γ_2 for loss function $\mathcal{L}_2$ is changed to 0.001 after epoch c, which is set to 0 before. The proposed method achieves a good result with an appropriate combination of the two sub-modules, which leads to a 1.24% improvement in the overall accuracy. Specially, accuracy on open-domain questions ('Open' in Table 2) achieves an improvement of 1.73% against the baseline, which demonstrates the proposed model's superior ability to deal with more complicated scenes and questions.

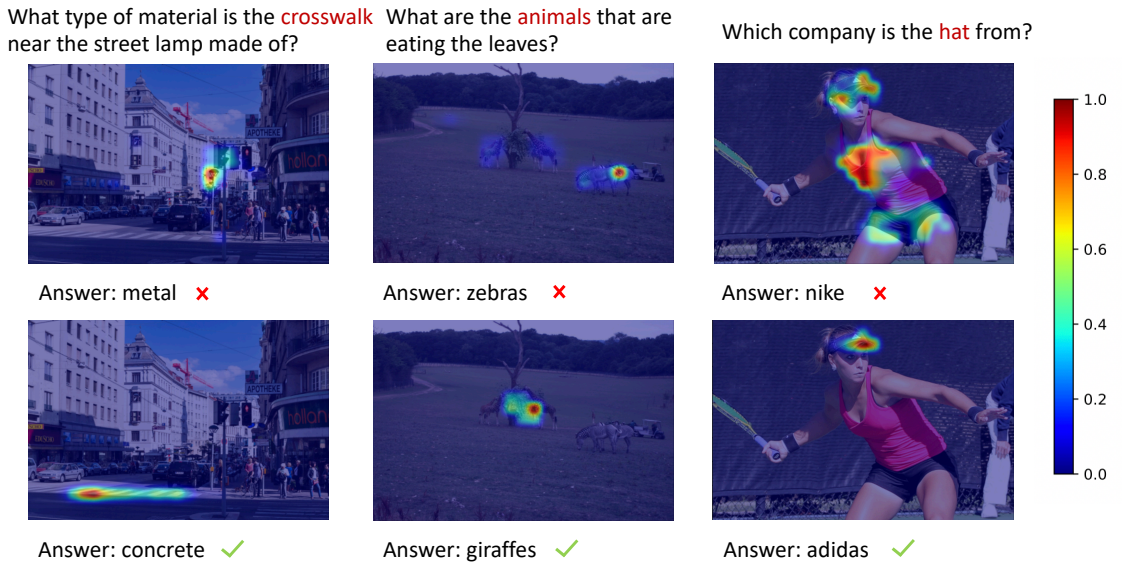

Figure 4. Visualization of the learned attention for some typical examples in the GQA. Each example shows different attentions for the same image, learned by the baseline and the proposed CRRN. We also display questions above the image and answers predicted by both models below the image. The highlighted part in the images and words in red denote the attended visual and textual contents, respectively.

Besides, we draw the curve of accuracy (Figure 3) for questions with different lengths. The proposed CRRN surpasses the baseline nearly for all lengths of questions, with high improvements for questions with more words, and the largest gain is for questions with 17 words.

4.4. Compared with SOTA

To demonstrate the effectiveness of our method for VQA, we compare it with previous state-of-the-art methods or submissions on the GQA dataset. In Table 3, our proposed method achieves 61.74% overall accuracy on the test split, outperforming all of the compared fusion-based methods, attention-based methods, and other reasoning methods. Compared with the state-of-the-art method LXRT [42], our method gains an improvement of 1.4% for the overall accuracy. While both consider interactions between images and questions, the proposed CRRN insightfully utilizes correspondence relations to supervise the model to learn meaningful visual regions in images. It demonstrates the significance of attention supervision. Several prior works have argued for the great success of large-scale pre-trained models [31, 29, 28]. Our blocks are easy to be incorporated with these models. They can indeed be highly beneficial to create models that are more capable and interpretable.

4.5. Visualization

In Figure 4, compared to the baseline, our model accurately attends to the image regions closely related to the questions. Take the sample at the first column as an example, the proposed model focuses on the keyword "crosswalk" and the question is correctly understood. And then, the crosswalk in the image is attended according to the semantic of the question. We can observe that the consistent relations between the image regions and the question words are important for answer prediction. From the sample in the third column, we can find that the proposed CRRN surprisingly has the capacity of attention concentration. It can be observed that the baseline and CRRN both capture the right word 'hat' in the question and the corresponding image region with a high attention weight. However, the baseline also attends much to some irrelevant regions like the women's chest and thighs, which interferes with the prediction of the answer. Instead, the proposed model focuses only on the relevant areas, the women's hat, and therefore gets the right answer 'adidas'.

5. Conclusions

In this paper, we propose a cross-modal relational reasoning network (CRRN) for VQA. The network learns a more accurate fine-grained interrelationship across modals. Novel relational masks are designed from explicit alignments between textual and visual contents. Related objective functions are applied to supervise the learning process. The proposed method contributes to capturing the latent relations and boosting the performance of VQA. The experimental results confirm the effectiveness of our work.

Acknowledgments This work was supported by the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (No.2019BDKFJJ002).

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. pages 2425–2433, 2015.

[5] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639, 2017.

[6] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Aviv Eisenschtat and Lior Wolf. Linking Image and Text with 2-Way Nets. *arXiv e-prints*, page arXiv:1608.07973, Aug. 2016.

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. pages 457–468, 2016.

[10] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling, 2016.

[11] Yash Goyal, Tejas Khot, Douglas Summersstay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[12] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[13] Wenya Guo, Ying Zhang, Xiaoping Wu, Jufeng Yang, Xiangrui Cai, and Xiaojie Yuan. Re-attention for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 91–98, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *IJCAI*, pages 789–795, 2019.

[16] Y. Huang, Q. Wu, C. Song, and L. Wang. Learning semantic concepts and order for image and sentence matching. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[17] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning, 2018.

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[19] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine, 2019.

[20] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[21] Jinhwa Kim, Jaehyun Jun, and Byoungtak Zhang. Bilinear attention networks. pages 1564–1574, 2018.

[22] Jinhwa Kim, Sangwoo Lee, Donghyun Kwak, Minoh Heo, Jeonghee Kim, Jungwoo Ha, and Byoungtak Zhang. Multimodal residual learning for visual qa. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[23] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling, 2017.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv: Learning*, 2014.

[25] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *CoRR*, abs/1411.7399, 2014.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv e-prints*, page arXiv:1602.07332, Feb. 2016.

[27] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching, 2018.

[28] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[29] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[30] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. Aligning visual regions and textual concepts for semantic-grounded image representations. In *Advances in Neural Information Processing Systems*, pages 6850–6860, 2019.

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[32] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[33] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2623–2631, 2015.

[34] Hyeonseob Nam, Jungwoo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. pages 2156–2164, 2017.

[35] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. *arXiv e-prints*, page arXiv:1611.00471, Nov. 2016.

[36] Duy-Kien Nguyen and Takayuki Okatani. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. *arXiv e-prints*, page arXiv:1804.00775, Apr. 2018.

[37] Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter- modality attention flow for visual question answering, 2019.

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. pages 1532–1543, 2014.

[39] Jinwei Qi, Yuxin Peng, and Yuxin Yuan. Cross-media multi-level alignment with relation attention network, 2018.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. 2015:91–99, 2015.

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[43] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. *arXiv e-prints*, page arXiv:1708.02711, Aug. 2017.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008, 2017.

[45] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural-symbolic models for interpretable visual question answering. *arXiv preprint arXiv:1902.07864*, 2019.

[46] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2019.

[47] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020.

[48] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering, 2019.

[49] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks*, 29(12):5947–5959, 2018.

[50] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv: Computer Vision and Pattern Recognition*, 2015.