Second Order Enhanced Multi-glimpse Attention in Visual Question Answering

Qiang Sun[1][0000−0002−2877−8960], Binghui Xie[2][0000−0001−6533−9281], and Yanwei Fu[⋆][2][0000−0002−6595−6893]

[1] Academy for Engineering & Technology, Fudan University, Shinaghai, China
18110860051@fudan.edu.cn
[2] School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai
Key Lab of Intelligent Information Processing Fudan University Fudan University,
Shanghai, China {16307130163,yanweifu}@fudan.edu.cn

Abstract. Visual Question Answering (VQA) is formulated as predicting the answer given an image and question pair. A successful VQA model relies on the information from both visual and textual modalities. Previous endeavours of VQA are made on the good attention mechanism, and multi-modal fusion strategies. For example, most models, till date, are proposed to fuse the multi-modal features based on implicit neural network through cross-modal interactions. To better explore and exploit the information of different modalities, the idea of second order interactions of different modalities, which is prevalent in recommendation system, is re-purposed to VQA in efficiently and explicitly modeling the second order interaction on both the visual and textual features, learned in a shared embedding space. To implement this idea, we propose a novel Second Order enhanced Multi-glimpse Attention model (SOMA) where each glimpse denotes an attention map. SOMA adopts multi-glimpse attention to focus on different contents in the image. With projected the multi-glimpse outputs and question feature into a shared embedding space, an explicit second order feature is constructed to model the interaction on both the intra-modality and cross-modality of features. Furthermore, we advocate a semantic deformation method as data augmentation to generate more training examples in Visual Question Answering. Experimental results on VQA v2.0 and VQA-CP v2.0 have demonstrated the effectiveness of our method. Extensive ablation studies are studied to evaluate the components of the proposed model.

Keywords: visual question answering; multi-glimpse attention; second order

1 Introduction

Visual Question Answering (VQA) has been topical recently, as its solution has relied on the successful models in both computer vision and natural language

communities. VQA provides a simple and effective testbed to verify whether AI can truly understand the semantic meaning of vision and language. To this end, numerous efforts have been made towards improving the VQA models by better representations [1], attention [2–5] and fusion strategies.

Despite various fusion mechanisms have been proposed, most of them focus on the fusion features of cross-modality. The early proposed models fuse the cross-modal features with first order interaction such as concatenation [6, 7]. Recently the bi-linear based methods [8–10] have been proposed to capture the fine-grained cross-modal features with second order interaction. Multimodal Tucker Fusion (MUTAN) [10] proposes an effective bi-linear fusion for visual and textual features based on low-rank matrix decomposition. Furthermore, it is essential and necessary to extend the first order or bi-linear fusion models to high order fusion ones, in order to better grasp the rich and yet complex information, existed in both visual and textual features. On the other hand, the explicit high order fusion has been widely adopted in many applications, e.g., in recommendation tasks [11, 12], and yet to a lesser extent VQA. For example, DeepFM [11] adopts the factorized machine (FM) to construct the explicit second order features of deep features.

However it is nontrivial to apply the explicit high order method to VQA, as most of visual object features, in principle, are orderless with respect to the different semantic attributes. This is quite different from the attribute embeddings (*e.g.,* age, gender) of recommendation, which are arranged in a fixed order. To overcome this problem, the multi-glimpse attention strategy is re-introduced, and re-purposed as the ordered visual representations; and each glimpse is corresponding to one type of attribute. To this end, a novel Second Order enhanced Multi-glimpse Attention (SOMA) model is thus proposed to construct the explicit high order features from the multi-glimpse outputs and the question features. The SOMA calculates, in an embedding space, the interactions of features from both intra-modality (*i.e.* interaction between different glimpse outputs) and cross-modality (*i.e.,* interaction between glimpse output and question feature). To fully utilize the outputs of multi-glimpse attention, we feed each attended features to an independent prediction branch, ensuring that each glimpse has focused on the question-related objects

Furthermore, despite several large-scale VQA datasets have been contributed to the community, effectively learning a deep VQA network still suffers from the training data scarcity, and *long-tailed* distributed question-answer pairs. Particularly, as in [13], only a limited number of question and answer pairs appeared frequently, whilst most of the other ones have only sparse examples. To alleviate this problem, a novel data augmentation strategy has been proposed in this paper. Typically, data augmentation, e.g., cropping and resizing images, aims at synthesizing new instances by training examples. Most of previous data augmentation strategies are conducted in visual features space, rather than semantic space. Remarkably, as a task requiring the high-level reasoning, the VQA should demand the data augmentation method by integrating the semantic information of each modality. To this end, a data augmentation method – semantic deforma-

tion, is proposed in this paper, by randomly removing some visual object and adds some noise visual instance. The images are dynamically augmented by randomly removing some visual objects, to create the more diverse visual inputs. Such a technique is further adopted as a self-supervised mechanism to improve the learning process of attention.

Formally, in this paper, we propose a Second Order enhanced Multi-glimpse Attention (SOMA) to tackle the tasks of visual question answering. As shown in Fig. 1, the model has several key components, including multi-glimpse attention module, second order module and classifier. The multi-glimpse attention module has different attention preference to different semantic aspects of question in each glimpse, which makes the attended feature more robust. The second order module explicitly models the interaction both on intra-modality and cross-modality by embedding the visual and textual features into a shared space. The classifier is strengthened with branch loss, which is able to provide a more direct supervised signal for each glimpse and the second module.

To sum up, we have several contributions as follows. (1) A second order module to construct the explicit second order features from the outputs of multi-glimpse and question feature in a shared embedding space. (2) Branch loss as a prediction signal to make each glimpse have better learning ability and attention performance. (3) A semantic deformation method with semantic objects cropping, noise objects adding and negative sample loss regularization. (4) Extensive experiments and ablation studies have shown the effectiveness of SOMA and semantic deformation.

2   Related Work

Visual Question Answering (VQA). The goal of visual question answering is to predict an answer on the given question and image pair [13, 14]. The dominant methods solve this problem as a classification task. A canonical model has three main stages: visual and textual feature extraction[15, 1], attention[2, 16, 5] and fusion strategy. The textual features from questions are mainly extracted from RNN based methods or Transformer. Recently, the object visual features from Faster-RCNN are preferred to the grid visual features by ResNet. Extensive attention models are proposed to identify the question-related information in the image, including question guided attention [1, 15], co-attention [16, 17, 5], self-attention [5, 4] and stacked attention[2, 18]. The fusion of visual and textual features includes first order [6, 7] and high order solution[8, 9].

Attention. Attention mechanism is a key component in the canonical VQA model. Visual attention exploits the visual grounding information to identify the salient regions for questions in early works[1, 2]. Some co-attention models[16, 4] find textual attention is also beneficial to detect the related words in questions along with visual attention. Recently, models with stacked self-attention layers [19, 20, 4, 5] have achieved state-of-the-art results on VQA task. But the multi-layer architecture makes it require a large computation cost. Studies [8, 18] have shown multi-glimpse attention is more robust by generating more than

one attention map. However, the relation between different attention results has not been well studied yet.

Fusion. Fusion in VQA aims to combine visual and textual features. The two main factors of fusion are interaction granularity and orders. Coarse-grained first order fusion methods [6, 7] combine the aggregated visual feature and question feature by concatenation. The simple first order fusion is limited to model the complex interactions of two modalities. Coarse-grained second order fusion approaches [8–10] advocate the effective bi-linear pooling between aggregated visual and textual features. Fine-grained second order fusion approach BAN [3] applies bi-linear attention between visual objects and question words and uses the sum pooling to obtain the fusion feature. MFH [18] is the most related work to our paper. It first adopts bi-linear attention between grid visual features and question features to generate multi-glimpse output. Then it concatenates the multi-glimpse output into one visual feature for cross-modality bilinear fusion. In contrast, our approach projects the multi-glimpse output and question feature into a shared embedding space to gather the interaction of cross-modality and intra-modality simultaneously. Inspired by the success of explicit high order features in recommendation tasks[11, 12], We construct an explicit second order feature in the shared embedding space as fusion. Since our fusion is based on the result of multi-glimpse attention, its granularity is more flexible, which means it is fine-grained if each attention map is near a one-hot vector.

Data Augmentation. Due to the dynamic nature of vision and language combination, the current scale of VQA dataset is insufficient for the deep neural network based model. In image classification, the traditional data augmentation methods include cropping, resizing, flipping, rotation, mixup [21–23] on the input space. The manifold mixup method[23] is proposed to interpolate the training instances in the hidden layer and label space. Counterfactual Sample Synthesizing (CSS[24]) use critical objects masking to generate numerous samples for robust model training. Inspired by manifold mixup, we propose a semantic deformation method in the visual semantic space by instance-level cropping and noise adding.

Self-Supervised Learning. The intrinsic structure information in the domain data can be utilized as an extra supervised signal for machine learning. In computation vision, the relative position of image patches [25], colorization [26], inpainting [27] and jigsaw problem [28] are formulated as surrogate tasks. In NLP tasks, the language model skip-gram [29, 30] learns the word embedding via context prediction in NLP tasks. Particularly, it adopts negative sampling to distinguish the learned vector from noise distribution. For the semantic deformation examples, we propose a hinge loss on the attention score of noise instance as an extra supervised signal by the assumption that a noise instance in VQA should be ignored with high possibility.
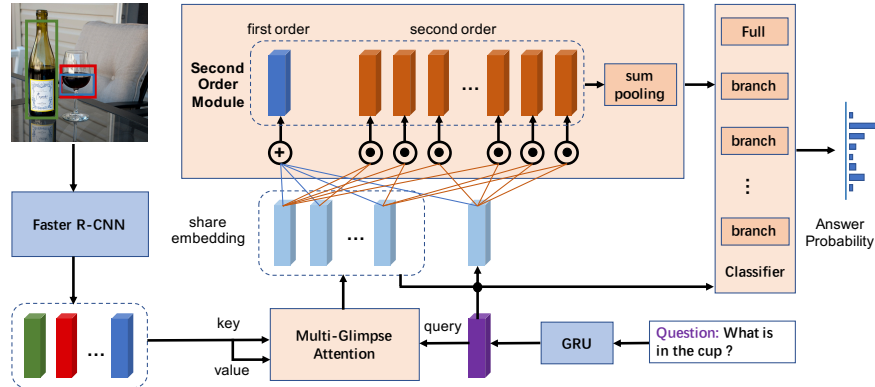
Fig. 1. The framework of SOMA. The main components of SOMA are multi-glimpse attention, second order module and classifier. Extracted visual features and question feature are fed to the multi-glimpse attention module to generate the attended visual features. The attended visual features and question feature are taken as inputs of the second order module. Finally, the attended visual feature, second order feature and the question feature are put into the classifier.

3 Approach

Overview. We formulate the visual question answering task, as a classification problem to calculate the answer a possibility $p(a \mid \mathbf{Q}, \mathbf{I})$ conditioned on the question $\mathbf{Q}$ and image $\mathbf{I}$. In this paper, we propose a novel framework – Second Order enhanced Multi-glimpse Attention (SOMA). SOMA is composed of three components: multi-glimpse attention module, second order module and classifier. The whole pipeline is illustrated in Figure 1. Multiple attended visual features are generated through the multiple-glimpse attention module with different semantic similarity preferences. The question embedding and attended visual features are fed into the second order module to produce the second order feature. Then the second order feature, attended visual features and question embedding are further passed to the classifier. During training, in addition to the full prediction, a branch prediction is used as an extra supervised signal for each glimpse in the classifier.

3.1 Feature Extraction

Typically, we have the image $\mathbf{I}$, question $\mathbf{Q}$ into the vision feature set $\mathbf{V}$ and the question embedding $\mathbf{q}$. Thus the original task of calculating $p(a|\mathbf{Q}, \mathbf{I})$ is translated into obtaining $p(a|\mathbf{q}, \mathbf{V})$.

Visual features. The visual feature set $\mathbf{V} = \{v_1, \ldots, v_k\}$, $v_i \in \mathbb{R}^{D_v}$ is the output of Faster R-CNN as described in Bottom-up [1]. The Faster R-CNN model is pre-trained on Visual Genome[31] and the object number k is fixed at 36 in our experiments. Thus, in our case, we denote the extracted visual object

set as:

$$\mathbf{V} = \text{RCNN}(\mathbf{I}, \theta_{\text{RCNN}}). \tag{1}$$

Question feature. The question embedding $\mathbf{q} \in \mathbb{R}^{D_t \times 1}$ is obtained from a single layer GRU. The words in the question are first transformed into a vector by GloVe. Then the word vectors are fed into the GRU in sequence. The last hidden state vector is taken as the question embedding. We represent the question embedding as:

$$\mathbf{q} = \text{GRU}(\mathbf{Q}, \theta_{\text{GRU}}). \tag{2}$$

3.2   Multi-glimpse Attention

To answer a question about an image, the attention map in one glimpse is used to identify the visual grounding objects. In multi-glimpse attention mechanism, each glimpse may have different semantic similarity preference, some prefer to attend the question-related colors, some prefer to attend the question-related shapes and so on. We adopt the multi-glimpse attention mechanism to make the attention results more robust and diverse. First, we project the visual feature set $\mathbf{V} \in \mathbb{R}^{k \times d_v}$ and question embedding $\mathbf{q} \in \mathbb{R}^{1 \times d_t}$ into a shared embedding space by $\mathbf{W}_v \in \mathbb{R}^{D_v \times D_h}$ and $\mathbf{W}_t \in \mathbb{R}^{D_t \times D_h}$ respectively. The two latent features are further combined through element products and then to generate the attention weight $A \in \mathbb{R}^{k \times m}$ as:

$$\mathbf{A} = \text{softmax}\left(\left(\text{ReLU}\left(\mathbf{1}\left(\mathbf{qW}_t\right)\right) \odot \text{ReLU}\left(\mathbf{VW}_v\right)\right)\mathbf{W}_G\right) \tag{3}$$

where $\mathbf{1} \in \mathbb{R}^{k \times 1}$ is an all-one vector by using k ones to expand the $\mathbf{q}$. $\mathbf{W}_G \in \mathbb{R}^{d_h \times m}$ and m is the number of glimpses. The softmax function is performed on the first dimension to generate weights on the k objects for each glimpse.

It then calculates the question attended visual features $\mathbf{G} \in \mathbb{R}^{m \times d_v}$ as a product of the attention weights and the original visual feature set,

$$\mathbf{G} = \mathbf{A}^T \mathbf{V}. \tag{4}$$

3.3   Second Order Module

We introduce the denotation for the second order module. Particularly, we introduce a score prediction task over a scalar variable set $\{x_1, x_2, \ldots, x_n\}$ as below,

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \tag{5}$$

where $\hat{y}$ is the predicted score, w_0 is the bias, $\sum_{i=1}^{n} w_i x_i$ represents the score from first order interaction and the last term denotes the impact of second order interaction. The inner product $< v_i, v_j >$ represents the coefficient for the interaction of variable x_i and x_j.

we propose a second order interaction module for the question and visual features as shown in Figure 1. The question feature is first projected into the

visual feature space. The concatenation of question feature and the attended visual features are further transformed into a shared embedding space as:

$$\mathbf{E} = \mathrm{ReLU}([\mathbf{G}; \mathrm{ReLU}(\mathbf{q}\mathbf{W}_{qv})]\mathbf{W}_{ve}) \tag{6}$$

where $\mathbf{W}_{qv} \in \mathbb{R}^{d_t \times d_v}$, $\mathbf{W}_{ve} \in \mathbb{R}^{d_v \times d_e}$ and d_e is the dimension of the latent space.

We construct the explicit second-order feature $\mathbf{s}$ over a vector variable set $\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2, \ldots, \mathbf{e}_{m+1}]$ as below:

$$\mathbf{s} = \sum_{i=1}^{m+1} \mathbf{e}_i + \sum_{i=1}^{m} \sum_{j=i+1}^{m+1} \mathbf{e}_i \circ \mathbf{e}_j \tag{7}$$

where $\circ$ denotes Hadamard product. The first term represents the impact of first order features and the second term reflects the importance of second order interactions. For simplicity and efficiency, the coefficients of this vector version FM are all fixed at 1. We argue that a proper embedding space learned by $\mathbf{W}_{ve}$ can alleviate this impact.

3.4   Classifier

The classifier takes the question embedding, multi-glimpse outputs and second order feature as inputs. It contains two subcomponent types: branch prediction module and full prediction module. The branch prediction is used for each glimpse or the second order feature. The full prediction takes all the glimpse outputs and second order features as inputs.

Branch prediction. To encourage each glimpse and the second order module to gather the information for answering, we feed each of them into an independent branch prediction module. In the branch prediction module the visual feature and question feature are first transformed into a hidden space, then projected by a fully connected layer to the answer space as follows:

$$\begin{aligned} \mathbf{h}_x &= \mathrm{ReLU}\left((\mathrm{ReLU}\left(\mathbf{q}\mathbf{W}_{qh}\right) \circ \mathrm{ReLU}\left(\mathbf{v}_x\mathbf{W}_{xh}\right))\right) \\ \hat{\mathbf{a}}_x &= \mathrm{sigmoid}\left(\mathbf{h}_x\mathbf{W}_{xa}\right) \end{aligned}$$

where $\mathbf{v}_x \in \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m, \mathbf{s}\}$, $\mathbf{W}_{qh} \in \mathbb{R}^{d_t \times d_h}$, $\mathbf{W}_{xh} \in \mathbb{R}^{d_v \times d_h}$, $\mathbf{W}_{xa} \in \mathbb{R}^{d_h \times d_a}$.
Full prediction. To fully utilize all the information in each branch, we concatenate all the hidden features in branches into $\mathbf{h}$, then map it into the answer space by a linear transformation.

$$\begin{aligned} \mathbf{h} &= [\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_{m+1}] \\ \hat{\mathbf{a}} &= \mathrm{sigmoid}\left(\mathbf{h}\mathbf{W}_{ha}\right) \end{aligned}$$

where $\mathbf{h} \in \mathbb{R}^{(m+1) \times d_h}$ and $\mathbf{W}_{ha} \in \mathbb{R}^{(m+1)d_h \times d_a}$.
Loss function. The total loss for prediction is composed of two parts: loss for branch prediction and loss for full prediction. The branch loss is scaled by a factor α_b.

$$L = L_f + \alpha_b L_b \tag{8}$$

Both the full prediction loss and branch prediction loss adopt binary cross-entropy (BCE) as the loss function.

$$L_f = \text{BCE}\,(\hat{a}, a)$$

$$L_b = \sum_{x=1}^{m+1} \text{BCE}\,(\hat{a}_x, a)$$

3.5   Data Augmentation by Semantic Deformation

We observe that, with a high probability, humans are able to answer questions when some objects in the image are occluded, or some un-related 'noise' objects are existing in the image. Inspired by this, we propose an object-level data augmentation method – semantic deformation. Essentially, it contains two key steps, *i.e.*, semantic objects cropping and semantic objects adding.

Semantic Objects Cropping. The size k of visual object set $\mathbf{V}$ from Faster R-CNN is usually very large to make sure that it contains the necessary objects for question answering. If we randomly remove a small number of k_r objects from the original visual object set, the remaining object set will still contain the clues for answering with high probability. We choose the k_r over a uniform distribution over 1 to R_{max}, where the R_{max} is the maximum number of objects that can be removed.

$$k_r = \text{uniform}(1, R_{max})$$

$$\mathbf{V}_{selected} = \text{select}(\mathbf{V}, k - k_r)$$

Semantic Objects Adding. We add k_a semantic noise objects to the visual object set from a randomly picked image. The number k_a of noise objects is a uniform distribution over 1 to A_{max}. The selected visual object set and the added noise object set can be merged into a new semantic image by concatenation.

$$k_a = \text{uniform}(1, A_{max})$$

$$\mathbf{V}_{add} = \text{select}(\mathbf{V}', k_a)$$

$$\mathbf{V}_{new} = \text{concate}(\mathbf{V}_{\text{selected}}, \mathbf{V}_{\text{add}})$$

Negative Example Loss. Intuitively, the added noise objects are unrelated to the question and visual context with a big chance. The irrelevance can be utilized as a self-supervised signal to guide the model where not to look. We apply a negative example loss to punish the model when the attention score on the added noise object surpasses a threshold.

$$L_{neg} = \sum_{g=1}^{m} \sum_{i=k-k_r+1}^{k-k_r+k_a} \max(0, \mathbf{A}_{i,g} - \tau) \tag{9}$$

where $\mathbf{A}_{i,g}$ is the attention score on the i-th object in the g-th glimpse, τ is the threshold attention value for noise objects. The negative loss is added to the total loss by a factor of α_{neg}.

$$L = L_f + \alpha_b L_b + \alpha_{neg} L_{neg} \tag{10}$$

where α_b is the coefficient.

4 Experiments

4.1 Datasets

We evaluate our model both on VQA v2.0 [32] and VQA-CP v2.0 [33]. VQA v2.0 contains 204k images from MS-COCO dataset [34] and 1.1M human-annotated questions. The dataset is built to alleviate the language bias problem existing in VQA v1.0. VQA v2.0 dataset makes the image matter by building complementary pairs as $< Question_A, Img_A, Ans_A >$ and $< Question_A, Img_B, Ans_B >$, which share the question but with different images and answers. The dataset is divided into 3 folder: 443K for training, 214K for validation and 453K for testing. VQA-CP v2.0 generates the new training and testing splits with changing priors from VQA dataset. The changing priors setting requires the model to learn the ground concept in images rather than memorizing the dataset bias. For each image, question pair, there are 10 human annotated answers. The evaluation metric for the predicted answer is defined as below:

$$\text{Acc}(ans) = \min \left\{ \frac{\#humans \quad that \quad said \quad ans}{3}, 1 \right\} \tag{11}$$

4.2 Implementation details

Model setting. The hyper-parameters of the proposed model in the experiments are as follows. The dimension of visual features d_v, question feature d_t, second order feature d_e and hidden feature d_h are set to 2048, 1024, 2048 and 2048 respectively. The number of candidate answers is set to 3129 according to the occurrence frequency. The glimpses number in attention is $m \in \{1, 2, 4, 6, 8\}$. We empirically set the branch loss factor α_b to 0.2.

Training setting. In training, we choose the Adamax optimizer with learning rate $min(t \times 10^{-3}, 4 \times 10^{-3})$ for the first 10 epochs and then decayed by 1/5 for every 2 epochs. The model is trained by 13 epochs with a clip value of 0.25 and batch size of 256. When tested at VQA v2.0 test-dev and test-std split, we train the model on training, validation and extra genome dataset. The performance on VQA v2.0 validation dataset is evaluated by the model trained on training split. The result on VQA-CP v2.0 test split is evaluated by the model trained on the training split.

Semantic Deformation setting. We denote the maximum number of objects removed and objects added as R_{max} and A_{max} respectively. They are both set to 4 by default. The negative sample loss factor α_{neg} and threshold τ is set to 1.0 and 0.18 respectively.

4.3 Results and Analysis

Results on VQA v2.0. First, we evaluate SOMA model on VQA v2.0 dataset. The results of our model and other attention based methods are summarized

Table 1. The Results of SOMA and other previous state-of-the-art methods on VQA v2.0 test-dev and test-std splits. The accuracy of each answer type on test-dev split is listed separately.

Method	test-dev				test-std
	Y/N	No.	Other	All	All
VQA team-LSTM+CNN[32]	-	-	-	-	54.22
MCB[8] reported in [32]	-	-	-	-	62.27
Bottom-up[1]	81.82	44.21	56.05	65.32	65.67
MF-SIG-VG[35]	81.29	42.99	55.55	64.73	-
CoR-3[36]	84.98	47.19	58.64	68.19	68.59
MFH[18]	-	-	-	66.12	-
MuRel[37]	84.77	49.84	57.85	68.03	68.41
DCN[17]	83.51	46.61	57.26	66.87	66.97
Counter[38]	83.14	**51.62**	58.97	68.09	68.41
SOMA(ours)	84.86	47.59	59.06	68.38	68.67
SOMA + Semantic Deformation	**84.87**	47.71	**59.12**	**68.43**	**68.71**

in Table 1. Bottom-up model is the winner of VQA v2.0 Challenge 2017 which utilizes the visual features from Faster R-CNN. Multimodal Compact Bilinear Pooling (MCB) [8] adopts count-sketch projection to calculate the outer product of visual feature and textual feature in a lower dimensional space. Multimodal Factorized High-order Pooling (MFH) [18] cascades multiple low rank matrix factorization based bilinear fusion modules. MuRel[37] adopts bilinear fusion to represent the interactions between question and visual vectors. Dense Co-Attention Network (DCN) is composed of co-attention layers for visual and textual modalities. Counter [38] is specialized to count objects in VQA by utilizing the graph of objects. In contrast to MFH, our model SOMA projects the visual and textual features into a shared embedding space and models the interaction of intra-modality and cross-modality simultaneously in the second order module. The results on VQA v2.0 show that SOMA improves the Bottom-up baseline with a margin of 3% overall, which demonstrates the effectiveness of the second order module. Furthermore, we apply the semantic deformation strategy in training and the results show that performance has been boosted on all answer types. Since we train the model with semantic deformation in the same epochs, the improvement is totally a benefit for free.

Results on VQA-CP v2.0. In this experiment, we compare SOMA with other competitors on VQA-CP v2.0 dataset. Ground visual Question Answering model (GVQA) disentangles the recognition of visual concepts from answer identification. Bilinear Attention Network (BAN) [3] develops an effective way to utilize multiple bilinear attention maps in a residual way. Bottom-up + AttAlign aligns the model attention with human attention to increase the robustness. Bottom-up + AdvReg trains a VQA model and a question-only model. It uses the question-only model as an adversary to discourage the VQA model to keep the language bias in its learned question feature. Table 2 shows that SOMA outperforms the bottom-up model with a margin of 2.3% in total. And it is only below to Bottom-

Table 2. The Results of SOMA and other previous state-of-the-art methods on VQA-CP v2.0 test splits. Models with * have been trained by [39].

Method	test			
	Y/N	Num	Other	All
SAN[2]	38.35	11.14	21.74	24.96
GVQA[33]	57.99	13.68	22.14	31.30
RAMEN[39]	-	-	-	39.21
BAN[3]*	-	-	-	39.31
Bottom-up[1]*	-	-	-	38.01
Bottom-up + AttAlign[40]	42.5	11.4	43.8	38.5
Bottom-up + AdvReg[41]	**65.5**	**15.5**	35.5	**41.2**
SOMA(ours)	43.0	12.9	**47.3**	40.8

up + AdvReg with a minor gap. To be noticed, Bottom-up + AdvReg is specially designed to prevent the model from overfitting the bias. While SOMA achieves this score with no special design and it can perform well on both VQA v2.0 and VQA-CP v2.0 dataset.
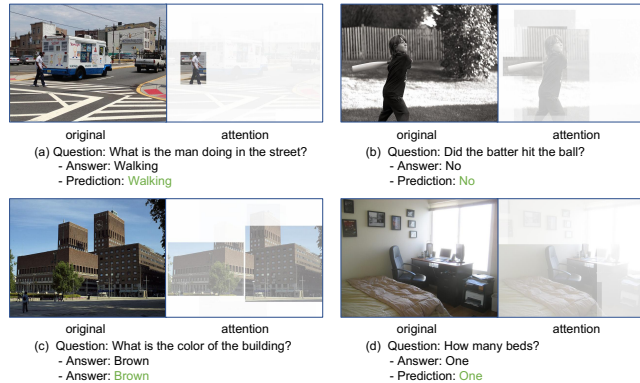

Fig. 2. Qualitative examples of the prediction results on VQA v2.0 dataset for model SOMA. In each example, the left part is the original image and the right part is the illustration of attention. Below the image is the question, ground-truth answer and predicted answer respectively.

Qualitative Results. To better reveal the insight of our model, we give some qualitative results. Particularly, to qualitatively analyze SOMA, we visualize the input image, question and predicted answer in Figure 2. The examples have shown that SOMA is able to attend to the question related region in the image during answering. This also validates the efficacy of our model.

Table 3. Ablation experiments results on VQA v2.0 validation split. SOMA w/o SO denotes the model without utilizing the second order feature. SOMA w/o BL denotes the model without using branch loss. The comparisons are performed on all the models with the glimpse number of 4.

Model	Y/N	Num	Other	ALL
SOMA w/o SO	82.98	45.00	57.31	65.34
SOMA w/o BL	82.96	44.64	57.04	65.15
SOMA (full)	**83.20**	**45.07**	**57.46**	**65.51**

(a) Acc of All   (b) Acc of Y/N   (c) Acc of Num   (d) Acc of Other
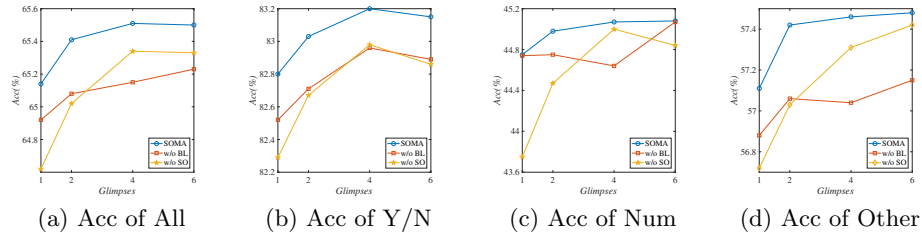
Fig. 3. Accuracies of model SOMA and its variants over different glimpses $G \in \{1, 2, 4, 6\}$ on VQA v2.0 validation split.

4.4   Ablation Study

Component Study. To investigate the contribution of each component, we train a full SOMA model with the glimpse number of 4 as a baseline. Then we propose two variants of SOMA. (1) SOMA w/o SO denotes the model does not contain the second order module for the multi-glimpse attention output. (2) SOMA w/o BL indicates the model does not conclude the branch loss of each glimpse. As shown in Table 3, the overall performance of SOMA w/o SO and SOMA w/o BL drops 0.17% and 0.36% respectively. Figure 3 further shows that the full model outperforms the variants on all glimpse $G \in \{1, 2, 4, 6\}$ in all answer types. We notice that the overall performance of the full model with the glimpse number of 2 is even better than the variants with the glimpse number of 4 or 6.

Table 4. Performance and model size of SOMA over the number of glimpses. Accuracy denotes the prediction accuracy on VQA v2.0 validation split. Params represent the total parameter size of the model. FLOP denotes float point operation cost. Computation cost is evaluated when the number of visual objects is 36 and the question contains 7 words over the glimpse number $G \in \{1, 2, 4, 6\}$.

Glimpse	Accuracy	#Params	#FLOPs
1	65.14	82M	420M
2	65.41	105M	475M
4	65.51	151M	586M
6	65.50	198M	696M

Performance and Cost. It is important to investigate the relationship between performance and cost, especially in real word application. Table 4 quantitatively shows the accuracy, model size and computation cost (FLOPs) trends over the glimpse number. The result shows that SOMA achieves the best performance when the number of glimpse is 4.

4.5 Experiments of Data Augmentation

Table 5. The performance of SOMA with semantic deformation on VQA v2.0 val split. SOMA indicates the baseline with 4 glimpses. SOMA + C indicates cropping on the input visual features of SOMA. SOMA + CA denotes cropping and noise adding. SOMA + CAN represents cropping and noise adding with negative example loss.

Model	SOMA	SOMA+C	SOMA+CA	SOMA+CAN
All	65.51	65.52	65.60	65.61

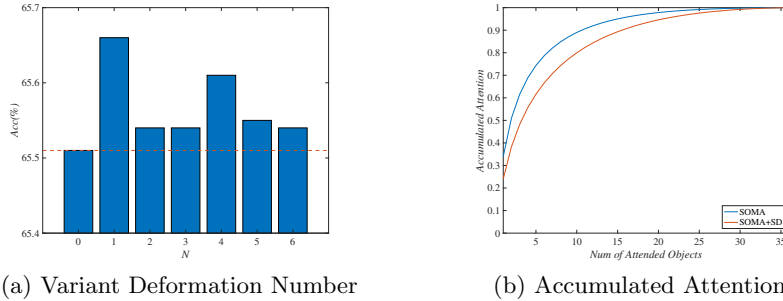(a) Variant Deformation Number   (b) Accumulated Attention

Fig. 4. (a) Performance of Semantic Deformation for different deformation number N, which denotes the maximum object removed and added number. The red dash line denotes the baseline which is the case $N = 0$. (b) Accumulated attention for the most attended M objects.

Data Augmentation Evaluation. To analyze semantic deformation, we propose serval variants and perform the ablation study on VQA v2.0 validation split. We train a SOMA model with the glimpse number of 4 as the baseline. Three variants are proposed by taking first n steps in semantic objects cropping, semantic objects adding and negative example loss applying gradually. The results in Table 5 show that semantic cropping, noise object adding are all beneficial to improve the performance of the baseline. And negative example loss is effective when the noise object adding strategy is used. When all the three techniques are used, the trained model achieves the best performance of 65.61% on the validation split. Furthermore, we conduct a series of experiments with

different maximum numbers for objects removed and added in semantic deformation. Figure 4(a) shows that the model with semantic deformation can beat the baseline with a slighter margin when the maximum number is from 1 to 6. Figure 4(b) indicates that the accumulated attention of SOMA model grows slower when trained with semantic deformation, which means the model gains robustness by attending to more related objects.
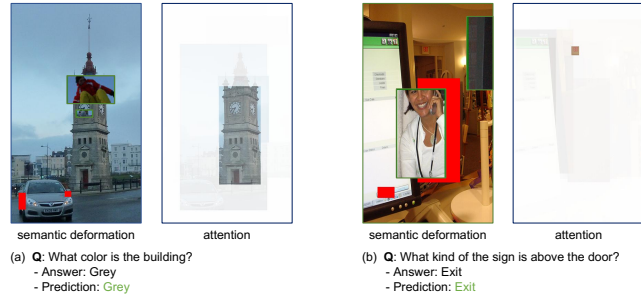

<div align="center">
semantic deformation   attention   semantic deformation   attention

(a) **Q**: What color is the building?   (b) **Q**: What kind of the sign is above the door?
    - Answer: Grey   - Answer: Exit
    - Prediction: Grey   - Prediction: Exit
</div>

Fig. 5. Qualitative examples of Semantic Deformation. The left image is the training image from semantic deformation. The red box denotes the bounding box of removed semantic objects. And the patch with green frames represents the added noise objects. The right image is the visualization results of attention maps.

Data Augmentation Example. To qualitatively analyze why semantic deformation works, we visualize a randomly generated image from semantic deformation as in Figure 5. For simplicity, we do not plot all of the 36 bounding boxes. We only show the bounding boxes of removed semantic objects and added semantic objects. Actually, the 36 bounding box has a lot of overlaps which make the visual feature set with high redundancy. Intuitively, we can see that the model is able to answer the question with a high possibility from the necessary grounding visual information.

In this paper, we propose a Second Order enhanced Multi-glimpse Attention (SOMA) model for Visual Question Answering. SOMA adopts a second order module to explicitly model the interaction on both the intra-modality and cross-modality in the shared embedding for multi-glimpse outputs and question feature. The branch loss is added to enhance each glimpse for better feature learning and attention ability. Furthermore, we advocate a novel semantic deformation method as data augmentation for VQA, which can generate the new image in the semantic space by semantic object cropping and semantic object adding. A negative example loss is introduced to provide a self-supervised signal for where not to look. The experiments on VQA v2.0 and VQA-CP v2.0 have shown the effectiveness of SOMA and semantic deformation. In feature works, we would like to design a better strategy for noise objects picking and apply semantic deformation to more multi-modal tasks.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. Volume 3. (2018) 6
2. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 21–29
3. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems. (2018) 1564–1574
4. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6639–6648
5. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6281–6290
6. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4613–4621
7. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
9. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
10. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proc. IEEE Int. Conf. Comp. Vis. Volume 3. (2017)
11. Guo, H., Tang, R., Ye, Y., Li, Z., He, X.: Deepfm: a factorization-machine based neural network for ctr prediction. arXiv preprint arXiv:1703.04247 (2017)
12. Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., Sun, G.: xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2018) 1754–1763
13. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2015) 2425–2433
14. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in neural information processing systems. (2014) 1682–1690
15. Kazemi, V., Elqursh, A.: Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162 (2017)
16. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems. (2016) 289–297
17. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6087–6096

18. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems **29** (2018) 5947–5959

19. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

20. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems. (2019) 13–23

21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

22. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

23. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Courville, A., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. arXiv preprint arXiv:1806.05236 (2018)

24. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10800–10809

25. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1422–1430

26. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision, Springer (2016) 649–666

27. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2536–2544

28. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision, Springer (2016) 69–84

29. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119

31. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123** (2017) 32–73

32. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. Volume 1. (2017) 3

33. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 4971–4980

34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755

35. Zhu, C., Zhao, Y., Huang, S., Tu, K., Ma, Y.: Structured attentions for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1291–1300
36. Wu, C., Liu, J., Wang, X., Dong, X.: Chain of reasoning for visual question answering. In: Advances in Neural Information Processing Systems. (2018) 273–283
37. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. arXiv preprint arXiv:1902.09487 (2019)
38. Zhang, Y., Hare, J., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. arXiv preprint arXiv:1802.05766 (2018)
39. Shrestha, R., Kafle, K., Kanan, C.: Answer them all! toward universal visual question answering models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10472–10481
40. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2591–2600
41. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: Advances in Neural Information Processing Systems. (2018) 1541–1551