An Improved Attention for Visual Question Answering

Tanzila Rahman[1,2] Shih-Han Chou[1,2] Leonid Sigal[1,2,3] Giuseppe Carenini[1]
[1]Department of Computer Science, University of British Columbia
Vancouver, BC, Canada
[2]Vector Institute for AI [3]Canada CIFAR AI Chair
{trahman8, shchou75, lsigal, carenini}@cs.ubc.ca

Abstract

We consider the problem of Visual Question Answering (VQA). Given an image and a free-form, open-ended, question, expressed in natural language, the goal of VQA system is to provide accurate answer to this question with respect to the image. The task is challenging because it requires simultaneous and intricate understanding of both visual and textual information. Attention, which captures intra- and inter-modal dependencies, has emerged as perhaps the most widely used mechanism for addressing these challenges. In this paper, we propose an improved attention-based architecture to solve VQA. We incorporate an Attention on Attention (AoA) module within encoder-decoder framework, which is able to determine the relation between attention results and queries. Attention module generates weighted average for each query. On the other hand, AoA module first generates an information vector *and an* attention gate *using attention results and current context; and then adds another attention to generate final* attended information *by multiplying the two. We also propose multimodal fusion module to combine both visual and textual information. The goal of this fusion module is to dynamically decide how much information should be considered from each modality. Extensive experiments on VQA-v2 benchmark dataset show that our method achieves better performance than the baseline method.*

1. Introduction

Different perceptual modalities can capture complementary information about aspects of an object, event or activity. As a result, multimodal representations are often shown to perform better in inference. Multimodal learning is widely used in the computer vision and forms basis for many visuo-lingual tasks, including image captioning [1, 23, 32], image-text matching [14, 30] and visual question answering [2, 17]). Visual question answering (VQA) is perhaps the most challenging, requiring detailed and intricate image and textual understanding (see Figure 1). Moreover, questions can be free-form and open-ended which requires VQA system to perform, simultaneously, a large collection of artificial intelligence tasks (*e.g.*, fine-grained recognition, object detection, activity recognition and visual common sense reasoning) to predict an accurate answer [2]. The answer format can also take different forms: a word, a phrase, yes/no, multiple choice, or a fill in the blank [28].

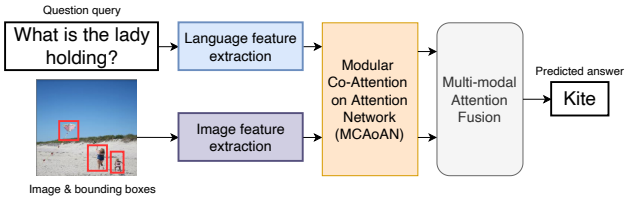


Figure 1: **Illustration of our proposed framework.** Given an image and a query question, we first extract visual and language features respectively. Our proposed Modular Co-Attention on Attention Network (MCAoAN) takes the features as inputs and refines both features jointly. The multimodal attention fusion fuses the refined visual and language features and then predicts final answer.

Inspired by the recent advantages of deep neural network, attention based approaches are widely used to solve many computer vision problems including VQA [1, 2, 36]. An attention based approach for VQA was first introduced by Shih *et al.* [27] and nowadays it has become an essential component in most of the architectures. Recent works [17, 36] include co-attention architecture to generate simultaneous attention in both visual and textual modality which increases prediction accuracy. The limitation of these, more *global*, co-attention methods, is their inability to model interactions and attention among individual image regions and segments of text (*e.g.*, at the word token level).

To address this problem, dense co-attention networks (*e.g.*, BAN [11], DCN [21]) have been proposed, where each image region is able to interact with any (and all) words in the question. As a result, the models can get better understanding and reason about the image-question relationships; this, in turn, results in improved VQA perfor-

mance. However, the bottleneck of these dense co-attention networks is the lack of self-attention within each modality, *e.g.*, region-to-region relationships in the image and word-to-word relationships in the question [34].

To overcome this, Yu *et al.* [34] proposed a deep Modular Co-Attention Network (MCAN) which consists of cascaded Modular Co-Attention (MCA) layers. MCA layer is obtained by combining two general attention units: self-attention (SA) and guided attention (GA). SA is able to capture intra-modal interactions (*e.g.*, region-to-region and word-to-word) while GA can capture cross-modal interactions (*e.g.*, word-to-region and region-to-word) by using multi-head attention architecture. While expressive and highly flexible, this form of attention still has a limitations. Specifically, the result is *always* a weighted combination of value pairs among which the model is attending. This maybe problematic when there is no closely related context over which the model is attending (*e.g.*, a word for which no context word or image region exists). In such a case attention would result in a noisy or, worse, distracting output vector that can negatively impact the performance.

Motivated by Huang *et al.* [10], in this paper we leverage the idea of Attention on Attention (AoA) module to address the above mentioned limitation. The AoA module is cascaded several times to form a novel Modular Co-Attention on Attention Network (MCAoAN) which is an improved extension to Modular Co-Attention Network (MCAN) [34]. The AoA module generates an information vector and an attention gate by using two separate linear transformations [10] which is similar to GLU [6]. Attention results and query context are concatenated together and through a linear transformation we can obtain an information vector. Similarly through another linear transformation followed by a sigmoid activation function we can obtain an attention gate. By applying element-wise multiplication, we finally obtain attended information which builds relation between multiple attention heads and keep only the most related one discarding all irrelevant attention results. As a result, the model is able to predict more accurate answer. We also propose a multi-modal fusion mechanism to dynamically modulate modality importance while combining image and language features.

Contributions. Our contributions are:

- We introduce an Attention on Attention module to form a Modular Co-attention on Attention Network (MCAoAN). MCAoAN captures intra- and inter-modal attention within and among visual and language modalities as well as able to mitigate information flow from irrelevant context.
- We also present a multimodal attention-based fusion mechanism to incorporate both image and question features. Our fusion network dynamically decides how to weight each modality to generate final feature representation to predict the correct answer.
- Extensive experiments on the VQA-v2 benchmark dataset [8] illustrate that the proposed method outperforms competitors, establishing significantly better performance than the baseline methods in visual question answering.

2. Related Works

In this section, we first briefly describe existing approaches for visual question answering and later review classical approaches to fuse image and question features.

2.1. Visual Question Answering

Antol *et al.* [2] first introduced the task of visual question answering (VQA), by combining computer vision with natural language processing, to mimic human understanding about a particular visual environment. The model used a CNN for feature extraction and an LSTM for language processing. The features were combined using element-wise multiplication in service of classifying the answers.

Over the last few years, a large number of deep neural networks have been proposed to improve the performance on VQA. Moreover, attention-based approaches became widely used to solve various sequence learning tasks, including VQA. The goal of attention module is to identify the most relevant part of image or textual content. Yang *et al.* [33] introduced an attention network to support multi-step reasoning for the image question answering task. A combination of bottom-up and top-down attention mechanism was presented in [1]. A set of salient image regions were proposed by bottom-up attention mechanism using Faster R-CNN [24]. On the other hand, task specific context was used to predict an attention distribution by top-down mechanism over the image regions. A model-agnostic framework is proposed by Shah *et al.* [26] which relies on cycle consistency to learn VQA model. Their model not only answers the posed question, but also generates diverse and semantically similar variations of questions conditioned on the answer. They enforce network to match the predicted answer with the ground truth answer to the original question. Wu *et al.* [31] propose a differential networks (DN), a novel plug and play module where differences between pair-wise features are used to reduce noise and learn inter-dependency between features. To extract image and text feature, Faster R-CNN [24] and GRU [5] are used respectively. Both features are refined by a differential module and finally combined to predict the answers.

Recently, co-attention based approaches are becoming popular. The goal of co-attention model is to learn image and question attention simultaneously. Lu *et al.* [17] introduced a co-attention network that jointly reasons about im-

age and question attention in a hierarchical fashion. Yu *et al.* [36] proposed an architecture to reduce irrelevant features by applying self attention for question embedding and question conditioned attention for image embedding. Multimodal attention is proposed in [18, 25] to focus on images, questions or answers feature simultaneously. Recently, bilinear attention is proposed in [7, 12, 35] to locate more accurate objects. A multi-step dual attention for multimodal reasoning and matching is presented in [20]. One major limitation of these co-attention based approaches is lack of dense interactions between different modalities. To overcome this limitation, dense co-attention based methods are proposed in [34, 11]. But dense co-attention can generate irrelevant vector in scenarios where nothing is related to the query. To overcome the problem, motivated by [10], in this paper we combine Attention-on-Attention (AoA) module with Modular co-attention network to improve existing architecture. Our revised attention mechanism delivers significantly better performance in VQA.

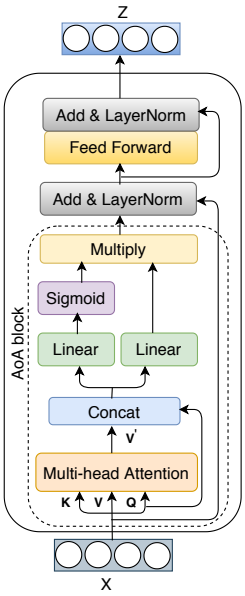
2.2. Fusion Strategies for VQA

To combine multi-modal features, sophisticated fusion technique is required. Depending on the type of fusion, existing VQA models can be divided into two categories: linear and bilinear [31]. Linear models use simple fusion approaches to combine image and question features. Simple element-wise summation and element-wise multiplication are used in [17, 33] and [15, 20] respectively. On the other hand, bilinear model uses more fine-grained approache to fuse image and question features. Fukui *et al.* [7] used outer product to fuse multi-modal features. A low-rank projection followed by an element-wise multiplication is used by Kim *et al.* [12]. A Multi-modal Factorized Bilinear (MFB) pooling approach with co-attention learning is proposed in [35]. Wu *et al.* [31] proposed a Differential Networks (DN) based Fusion (DF) approach which first calculates differences between image and textual feature elements and then combines the differential representations to predict final answer.

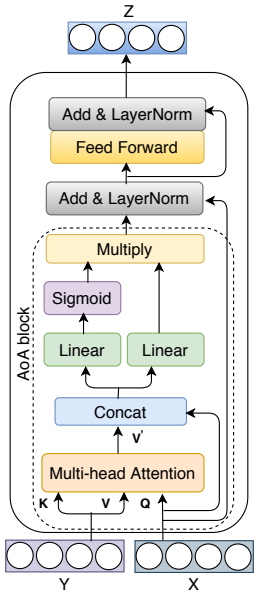
In this paper, we propose an attention-based multi-modal fusion to combine image and question features by dynamically deciding how much weight to put on each modality; the weighted features are used to predict final answer.

3. Our Approach

Motivated by [10], in this paper we present Modular Co-Attention on Attention Network (MCAoAN) module which is an extension of Modular Co-Attention Network (MCAN) [34]. MCAoAN consists of Modular Co-Attention on Attention (MCAoA) layer which is a composition of two primary attention units: Self Attention on Attention (SAoA) and Guided Attention on Attention (GAoA) unit. In this section, we first discuss SAoA and GAoA units in Section 3.1 followed by Modular Co-Attention on Attention (MCAoA) layer in Section 3.2. Lastly we present our MCAoAN with multimodal fusion mechanism in Section 3.3 and Section 3.4 respectively.



(a) Self Attention on Attention block



(b) Guided Attention on Attention block

Figure 2: **Illustration of the two basic attention units.** (a) Self Attention on Attention block (SAoA), which takes input feature X and output attended feature Z for X; and (b) Guided Attention on Attention block (GAoA),which takes two input features X and Y and generate attended feature Z for the input X guided by Y feature. Here X and Y represents image and question features respectively.

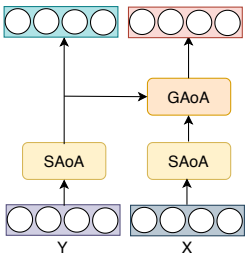


Figure 3: **Illustration of Modular Co-Attention on Attention (MCAoA) layer.** It consists of two attention units: Self Attention on Attention (SAoA) unit and Guided Attention on Attention (GAoA) unit where Y and X denotes question and image features respectively.

3.1. SAoA and GAoA Units

Our SAoA unit (see Figure 2(a)) is an extension of multi-head self attention mechanism [34]. Multi-head attention consists of N parallel heads where each head can be represented as a scaled dot product attention function as follows:

$$\mathrm{f_{att}} = \mathrm{f}(Q, K, V) = \texttt{Softmax}\left(\frac{QK}{\sqrt{d}}\right)V, \qquad (1)$$

where attention function $\mathrm{f}(Q, K, V)$ operates on Q, K and V corresponds to query, key and value respectively. The output of this attention function is the weighted average vector V'. To do so, first we calculate the similarity scores between Q and K; and normalize the scores with $\texttt{Softmax}$. The normalized scores are then used together with V to generate weighted average vector V'. Here, d is the dimension of Q and K; both dimensions are the same.

The multi-head attention module always generates weighted vector, no matter whether it finds any relation between Q and K/V or not. So this approach can easily mislead or generate wrong answer for VQA. Therefore, following [10], we incorporate another attention function over the multi-head attention module to measure the relation between attention results (V') and the query(Q). The final AoA block will generate an information vector (I) and attention gate (G) through two separate linear transformations which can be represented as follows:

$$\mathrm{I} = \mathrm{W_Q Q} + \mathrm{W_{V'} V'} + \mathrm{b_I}, \qquad (2)$$

$$\mathrm{G} = \sigma(\mathrm{W_G Q} + \mathrm{W_{G'} V'} + \mathrm{b_G}), \qquad (3)$$

Here, $\mathrm{W_Q}$, $\mathrm{W_{V'}}$, $\mathrm{W_G}$, $\mathrm{W_{G'}} \in \mathbb{R}^{d \times d}$ and $\mathrm{b_I}$, $\mathrm{b_G} \in \mathbb{R}^d$. d is the dimension of Q and V' where $V' = \mathrm{f_{att}}$ and σ denotes sigmoid function. AoA block adds another attention via element-wise multiplication between both information vector and attention gate. Moreover, SAoA uses a point-wise feed-forward layer after the AoA block, considering only input features $X = [x_1, x_2, ..., x_m] \in \mathbb{R}$.

Following [34], we also propose another attention unit called guided attention on attention (GAoA) unit (see Figure 2(b)). Unlike SAoA unit, GAoA uses AoA block and a point-wise feed-forward layer along with two input features X and $Y = [y_1, y_2, ..., y_n] \in \mathbb{R}$ where X is guided by Y. In both attention unit, feed forward layer takes the output feature of AoA block and apply two fully connected layers along with ReLU and dropout function (i.e. $\texttt{FC}(4d) - \texttt{ReLU} - \texttt{dropout}(0.1) - \texttt{FC}(d)$).

3.2. MCAoA layers

Modular Co-Attention on Attention (MCAoA) layer (see Figure 3) consists of two attention units discussed in Section 3.1. Here X and Y represents image and question feature respectively. MCAoA layer is designed to handle multimodal interactions. We use cascaded MCAoA layers, *i.e.*, output from previous MCAoA is fed as input to the next MCAoA layer. For both input features, MCAoA layer first uses two separate SAoA units to caption intra-modal interactions for X and Y separately and then uses GAoA unit to capture inter-modal relationships where Y guides X feature.

3.3. MCAoAN

In this section, we discuss our proposed modular co-attention on attention network (MCAoAN) (see Figure 4) which is motivated by [34]. First we have to pre-process the inputs (*i.e.*, image and query question) into appropriate feature representations. We use Faster R-CNN [24] with ResNet-101 as its backbone which is pretrained on Visual Genome dataset [13] to process input images. The intermediate feature of the detected object from Faster R-CNN is considered as visual feature representation. Following [34], we also consider a threshold value to obtain dynamic number of detected objects, *e.g.*, x_i is corresponds to i-th object feature. The final image feature is represented by a feature matrix X.

The input query question is first tokenized and later trimmed to maximum 14 words. The pre-trained GloVe embedding [22] is used to transformed each word into a vector representation. This results a final representation of size $n \times 300$ for a sequence of words where $n \in [1, 14]$ denotes the number of word in the sequence. The word embedding is fed to a one layer LSTM network [9] and generate final query feature matrix Y by capturing the output features of all words.

Both input features are passed to the encoder-decoder module which contain cascaded MCAoA layers. Similar to [34], encoder learns attention question features Y_L by stacking L number of SAoA units. On the other hand, decoder learns attended image features X_L by stacking L number of GAoA units by using query features Y_L.

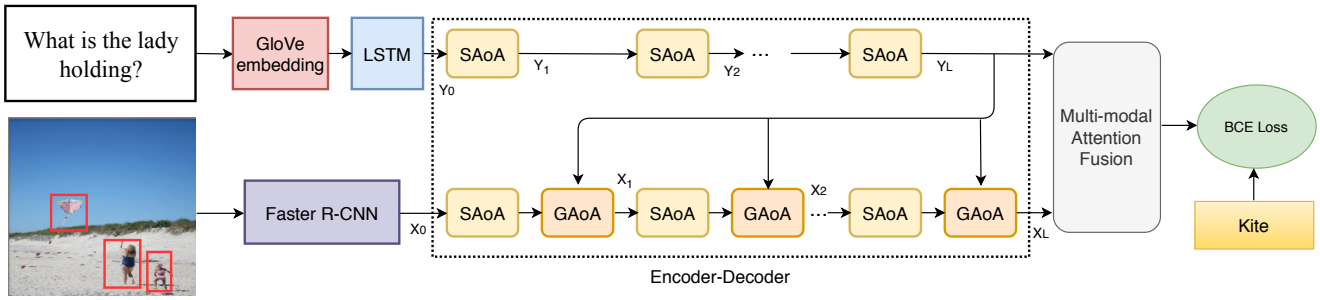


Figure 4: **Illustration of overall architecture of proposed Modular Co-Attention on Attention Network (MCAoAN).** The network takes image and question feature as inputs. Image features are the intermediate features extracted from a Faster R-CNN [24] model and each work from the question is transformed to a vector using 300-D GloVe word embedding [22] followed by a LSTM unit [9]. Both features are fed to an Encoder-Decoder module consists of cascaded MCAoA layers and generate X_L and Y_L feature representations. X_L and Y_L denotes image and question feature respectively and combined together to generate desire answer by a multi-modal fusion module.

3.4. Multi-modal Fusion.

The outputs (i.e image features $X_L = [x_1, x_2, ..., x_m] \in \mathbb{R}^{m \times d}$ and question features $Y_L = [y_1, y_2, ..., y_n] \in \mathbb{R}^{n \times d}$) from encoder-decoder contains attended information about image and query regions. Therefore, we need to apply an appropriate fusion mechanism to combine both feature representation. In this paper, we propose two kind of multi-modal fusion networks (see Figure 5) to aggregate features of both modality: (1) Multi-modal Attention Fusion and (2) Multi-modal Mutan Fusion. Following [34], we first use two layers of MLP (i.e. FC(d)- ReLU - Dropout(0.1) - FC(1)) for both X_L and Y_L; and generate attended features X' and Y' as follows:

$$\mathrm{X}' = \sum_{i=1}^{m} \mathtt{Softmax}(\mathtt{MLP}(\mathrm{X_L}))\, \mathrm{x}_i, \tag{4}$$

and

$$\mathrm{Y}' = \sum_{i=1}^{n} \mathtt{Softmax}(\mathtt{MLP}(\mathrm{Y_L}))\, \mathrm{y}_i, \tag{5}$$

Now we have rich attended features from both modality and at the same time each modality should use to generate attention with one another for better prediction. Therefore, we have to decide, how much information should use from each modality. Following [19], in multi-modal attention fusion, we apply concatenation on X′ and Y′ followed by a series of fully-connected layers (*i.e.*, $\mathtt{FC}(1024) - \mathtt{Dropout}(0.2) - \mathtt{FC}(512) - \mathtt{Dropout}(0.2) - \mathtt{FC}(2) - \mathtt{Softmax}$) (see Figure 5 (a)). The output of these operations is a 2-dimensional feature vector that corresponds to the importance of two modality for answer prediction. After that, we generate weighted average of attended feature (i.e. A and B) for each modality similar to eq. 4 and 5. A and B is combined with attended visual and textual features X' and Y'. Finally, fused feature is fed to a $\mathtt{LayerNorm}$ to stabilize the training followed by a fully connected layer and sigmoid activation to generate predicted answer Z. We use binary cross-entropy loss (BCE) to train the network.

On the other hand, we also leverage a powerful fusion technique, MUTAN fusion [3], to integrate image and question features (see figure 5 (b)) in multi-modal mutan fusion. The network is similar to the above model but replacing the concatenation to MUTAN fusion with fully-connected layers (*i.e.*, $\mathtt{Dropout}(0.2) - \mathtt{FC}(2) - \mathtt{Softmax}$).

L	All	Other	Y/N	Num
L = 2	81.88	74.47	96.11	69.00
L = 4	83.34	**76.48**	96.65	71.00
L = 6	**83.45**	76.45	**96.83**	**71.44**
L = 8	82.20	75.42	95.87	68.53

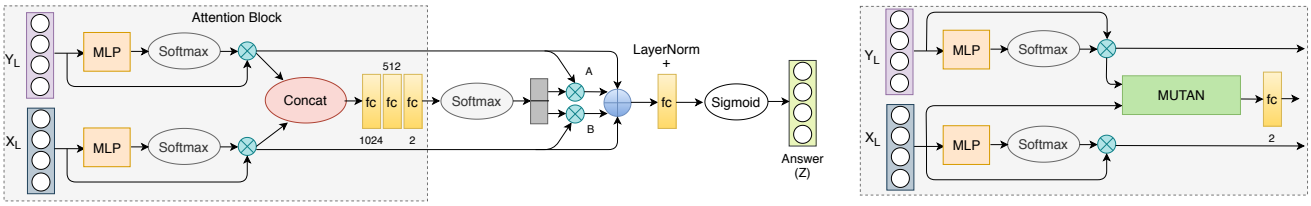
Table 1: **Experimental results with different L.** Here we use a range of values from 2 to 8 on validation set. Best performance is achieved with $L = 6$. Therefore, in this paper we choose $L = 6$ for our work.

4. Experiments

In this section we first describe the dataset (see Section 4.1) used in our experiments. Then we present experimental setup and implementation details in Section 4.2. In Section 4.3, we include a number of ablations to show the effectiveness of our proposed model. Lastly, we discuss experimental results in Section 4.4.

4.1. Datasets

To evaluate our method, in this paper we use VQA-v2 benchmark dataset [8] which consists of images from MS-COCO dataset [16] with human annotated question-answer pairs. There are 3 questions for each image and 10 answers



(a) Multi-modal Attention Fusion

(b) Attention Block for Multi-modal Mutan Fusion

Figure 5: **Illustration of proposed multimodal fusion network.** (a) Multi-modal attention fusion where we apply simple concatenation to combine initial attended features from both image and language modalities and apply series of fully connected layer to generate weighted features. The final weighted features represents how much importance should we give on each modality. (b) Multi-modal mutan fusion, another version of multi-modal fusion where we incorporate mutan fusion instead of concatenation keeping rest of the network similar to multi-modal attention fusion.

Methods	**All**	**Other**	**Y/N**	**Num**
MCAN [34]	81.20	73.73	95.86	67.30
Ours (MCAoAN)	82.91	75.92	96.47	70.38
Ours (MCAoAN + Mutan)	83.00	76.13	96.36	**70.42**
Ours (MCAoAN + Multi-modal Attention Fusion)	**83.25**	**76.51**	**96.58**	70.40

Table 2: **Visual Question Answering results using VQA-v2 dataset**. Comparison of our proposed approach with state-of-the-art method on validation set. Here we also show each component in our proposed method contribute to increase the performance of VQA system.

Methods	**All**	**Other**	**Y/N**	**Num**
Bottom-up [29]	65.32	56.05	81.82	44.21
MFH [36]	68.76	59.89	84.27	49.56
BAN [11]	69.52	60.26	85.31	50.93
BAN+Counter [11]	70.04	60.52	85.42	54.04
MuRel [4]	68.03	57.85	84.77	49.84
MCAN [34]	70.63	60.72	86.82	53.26
Ours (MCAoA)	**70.90**	**60.97**	**87.05**	**53.81**

Table 3: **Experimental results with other state-of-the-art models on Test-dev.**

Methods	**All**	**Other**	**Y/N**	**Num**
Bottom-up [29]	65.67	56.26	82.20	43.90
BAN+Counter [11]	70.35	-	-	-
MuRel [4]	68.41	-	-	-
MCAN [34]	70.90	-	-	-
Ours (MCAoA)	**71.14**	**61.18**	**87.25**	**53.36**

Table 4: **Experimental results with other state-of-the-art models on Test-std.**

per questions. The dataset has three parts: train set (80k images with 444k QA pairs), validation set (40k images with 214k QA pairs) and test set (80k images with 448k QA pairs). Moreover, test set is splited into two subsets: test-dev and test-standard where both are used for online evaluation performance. For measuring the overall accuracy, three types of answer are considered: Number, Yes/No and other.

4.2. Experiment and Implementation Details

To evaluate our method, we follow the experimental protocol proposed by [34]. The number of head in multi-head attention is 8. The latent dimension for both multi-head and AoA block is 512. Therefore, the dimension of each head is $512/8 = 64$. The size of the answer vocabulary is 3129.

To train the MCAoA network we use Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We train our network up to 13 epoch with batch size 64 which takes around 24hrs to complete the training. The learning rate set to $min(2.5Te^{-5}, 1e^{-4})$ where T represents current epoch. Learning rate starts to decay by $1/5$ every 2 epochs when $10 \leq T$.

4.3. Ablation studies

We run a number of experiments to show the effectiveness of our proposed method and results of these experiments are presented in Table 1 and 2.

Number of Cascaded Layer (L): MCAoA layers consist of L number of stacked SAoA and GAoA units. From Table 1, we can see, initially, with the increasing value of L, performance of the model is also increasing – up to $L = 6$. After that the performance is saturated. We use $L = 6$ in our final model. We use *validation set* for this experiment with the default hyperparameters of [34].

Figure 6: **Illustration of some qualitative results from validation set using MCAN [34] and our method.** First we present ground-truth (GT) annotations followed by the predicted answers of state-of-the-art method and our proposed method. Here Q and A represents query question and generated answer respectively. Moreover, red text indicates predicted wrong answer for the corresponding question.

Effectiveness of Each Individual Component: In this paper, our improved architecture has two important components: (1) MCAoAN network which consists of SAoA module and GAoA module and (2) Multi-modal fusion to incorporate image and language features. Here, we describe two different fusion mechanism : Mutan fusion and Multi-modal attention fusion. Table 2 shows experimental results of these individual components and compare with existing MCAN [34] on *validation set*. From the table, we see that incorporating SAoA and GAoA module with MCAN improves the performance of VQA system.

Moreover, we argue that a sophisticated way to aggregate language and visual features to support multi-modal reasoning is essential to further boost the performance. Table 2 also shows the comparison of different fusions with the MCAoA only where the former achieves better performance. More specifically, our proposed MCAoAN with both multi-modal fusion modules outperforms the baseline about 2% accuracy on the whole *validation set*. This shows that the fusion module is important to combine vision and language representations. The proposed both fusion modules are suitable for VQA tasks. Among them multi-modal attention fusion performs the best. Beside that, Table 2 also shows that each individual component within our proposed method is important to increase the performance of VQA system.

4.4. Experimental Results

We evaluate our model on VQA-v2 dataset and compare with other state-of-the-art methods. We re-run the PyTorch implementation provided by [34][1] and compare the results with our proposed method. Table 3 and 4 shows experimental results using test-dev and test-std respectively using online evaluation [2]. Offline evaluation only supports on validation split (see table 2). Figure 6, shows some qualitative results using our method on validation set. From the experimental results, we can see that our proposed method

[1] https://github.com/MILVLG/mcan-vqa

[2] https://evalai.cloudcv.org/web/challenges/challenge-page/163/overview

Figure 7: **Qualitative results with multi-modal fusion.** The first row is the input images, questions and ground truth answers. The second row is the baseline model MCAN [34]. The third row is the proposed model, MCAoAN w/ multi-modal fusion. The probabilities on the image and in front of the question represent the weight from each modality. We also show the attention across bounding boxes and words. In the image, the brighter area with green bbox has higher weight. For questions, the darker color of the word, the higher attention score.

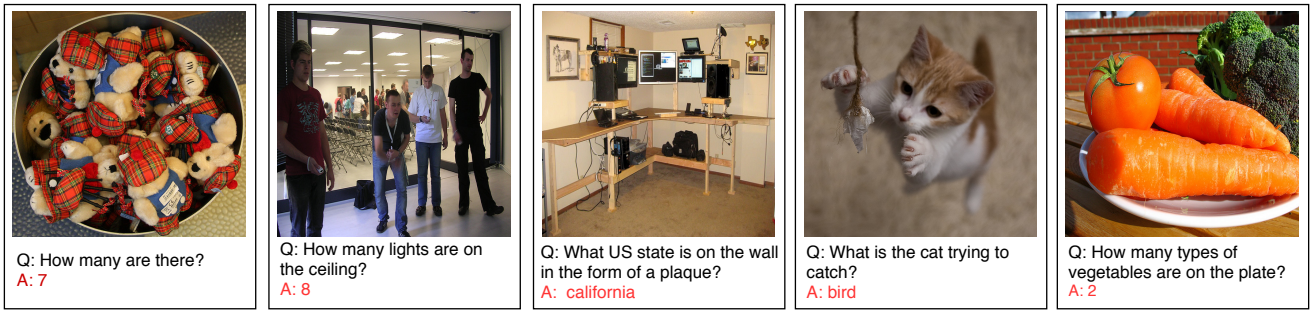


Figure 8: **Illustration of some failure cases using our method.** Here Q and A represents query question and predicted wrong answer (mark as red) respectively.

outperforms other baseline methods on VQA. In Figure 7, we also visualize multi-modal fusion to compare how correctly MCAN [34] and our proposed multi-modal attention fusion can able to focus on image and question elements. The brighter bounding-box along with green color within the image and darker color in question represents higher attention score. We can see that our proposed method is able to focus more on correct answer. Beside that, Figure 8 shows typical failure cases using our method.

5. Conclusion

In this paper, we propose an improved end-to-end attention based architecture for visual question answering. Our proposed method includes modular co-attention on attention module with multi-modal fusion architecture. In this paper, we propose two version of multi-modal fusion : multi-modal attention fusion and multi-modal mutan fusion. Experimental results show that each component within our model improve the performance of VQA system. Moreover, The final network achieves significant performance on VQA-v2 dataset.

References

[1] 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. 1, 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 5

[4] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 6

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[6] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 933–941. JMLR. org, 2017. 2

[7] 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. *arXiv preprint arXiv:1606.01847*, 2016. 3

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2, 5

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4, 5

[10] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 2, 3, 4

[11] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 1, 3, 6

[12] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 3

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4

[14] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 1

[15] Ruiyu Li and Jiaya Jia. Visual question answering with question representation update (qru). In *Advances in Neural Information Processing Systems*, pages 4655–4663, 2016. 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[17] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016. 1, 2, 3

[18] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. *arXiv preprint arXiv:1711.06794*, 2017. 3

[19] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156. IEEE, 2016. 5

[20] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 3

[21] Duy-Kien Nguyen and Takayuki Okatani. 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*, pages 6087–6096, 2018. 1

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4, 5

[23] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8908–8917, 2019. 1

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 4, 5

[25] Idan Schwartz, Alexander Schwing, and Tamir Hazan. High-order attention models for visual question answering. In *Advances in Neural Information Processing Systems*, pages 3664–3674, 2017. 3

[26] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6649–6658, 2019. 2

[27] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016. 1

[28] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. Visual question answering using deep learning: A survey and performance analysis. *arXiv preprint arXiv:1909.01860*, 2019. 1

[29] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018. 6

[30] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 1

[31] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Ruifan Li. Differential networks for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8997–9004, 2019. 2, 3

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1

[33] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 2, 3

[34] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 2, 3, 4, 5, 6, 7, 8

[35] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 3

[36] 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 and learning systems*, 29(12):5947–5959, 2018. 1, 3, 6