An Improve Attention based Architecture for Visual Question Answering

Tanzila Rahman
Department of Computer Science
University of British Columbia
trahman8@cs.ubc.ca

Abstract

We consider the problem of Visual Question Answering (VQA). Given an image and a free-form, open-ended question in natural language, the goal of VQA system is to provide accurate answer in natural language about the given image. The task is quite challenging because it requires simultaneous understanding of both visual and textual information. Recently, attention mechanism is widely used to generate correct answer by capturing global dependencies. In this paper, we propose an improve attention based architecture to solve VQA. We incorporate an Attention on Attention (AoA) module with existing 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 modules 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 them. We also propose a multimodal attention fusion module to combine both visual and textual information. The goal of this fusion module is to decide how much information should consider from each modality. Extensive experiment on VQA-v2 benchmark dataset shows that our method achieves the state-of-the-art performance.

1 Introduction

Signals from multiple modalities can capture complementary information about different aspects of an object, event or activity. Therefore, multimodal representations are capable of performing better during inference. Recently, multimodal learning is widely used in computer vision community where different modality can play integral role (e.g. visual captioning (Anderson et al., 2018; Xu et al., 2015; Rahman et al., 2019), image-text matching (Wang et al., 2018; Lee et al., 2018), visual question answering (Antol et al., 2015; Lu et al., 2016) and so on). Visual question answering (VQA) is one of the most challenging tasks among all multimodal based learning tasks that requires both image and textual understanding (see figure 1). Moreover, questions can be free-form and open-ended which requires a vast knowledge in artificial intelligence (e.g. fine-grained recognition, object detection, activity recognition, visual common sense reasoning and so on) to predict accurate answer (Antol et al., 2015). The answer can be a word, a phrase, yes/no, multiple choice answer, or a fill in the blank answer (Srivastava et al., 2019).

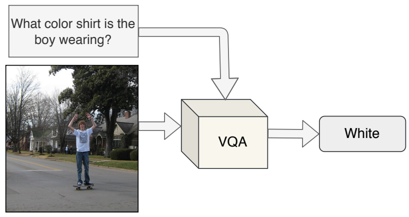


Figure 1: Illustration of visual question answering problem. Given an image and a query question, the goal of visual question answering is to predict accurate answer.

Inspired by the recent advantages of deep neural network, attention based approach are widely used to solve many computer vision problems including VQA (Anderson et al., 2018; Antol et al., 2015; Yu et al., 2018). To solve VQA, attention based approach was first introduced by Shih et al. 2016 and nowadays it becomes an essential component in most of the architectures. Recent works (Lu et al., 2016; Yu et al., 2018) include co-attention architecture to generate simultaneous attention in both visual and textual modality which increases prediction accuracy. There are significant limitations of this co-attention based architecture due to the lack of interaction between different modalities. As a result, co-attention module is unable to co-relate each image region with each word in the query question.

To solve this problem, dense co-attention network (e.g. BAN (Kim et al., 2018), DCN (Nguyen and Okatani, 2018)) has been proposed where each image region is able to interact with any word in the question. Therefore, the model can get better understanding about the image-question relationship and improve the performance of VQA system. However, the bottleneck of this co-attention network is lack of generating self attention between each modality such as region-to-region relationship in image and word-to-word relationship in question (Yu et al., 2019).

To overcome this bottleneck, Yu et al. 2019 propose a deep Modular Co-Attention Network (MCAN) which contains cascaded Modular Co-Attention (MCA) layers. MCA layer can be 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 inter modal interactions (e.g. word-to-region and region-to-word) by using multi-head attention. Each multi-head attention uses scaled dot-product attention function. But there could certain situation where attention results is not what the VQA system expects. This can be happened when there is nothing related to the query but still attention module generating an irrelevant vector leading wrong results.

Following Yu et al. 2019, in this paper we propose a cascaded Modular Co-Attention on Attention Network (MCAoAN) based on Attention on Attention (AoA) module (Huang et al., 2019) by adding another attention on top of multi-head attention. MCAoAN is an extension of Modular Co-Attention Network (MCAN) (Yu et al., 2019). AoA module generates an information vector and an attention gate by using two separate linear transformations (Huang et al., 2019) which is similar to GLU (Dauphin et al., 2017). 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 most related one discarding all irrelevant attention results. Hence, the model is able to predict more accurate answer.

Contributions. Our contributions includes:

- We propose a Modular Co-Attention on Attention Network (MCAoAN) based on Attention on Attention module to capture related attention results from both visual and textual modality simultaneously.
- We also present multimodal attention-based fusion mechanism to incorporate both image and question features. Our fusion network decides, how to weight each modality to generate final feature representation to predict correct answer.
- Extensive experiments on the VQA-v2 benchmark dataset (Goyal et al., 2017) shows that our proposed method outperform other state-of-the-art methods in visual question answering.

2 Related Works

Antol et al. 2015 first introduced the task of visual question answering by combining the concept of computer vision and natural language processing to mi-mick human understanding about a particular environment. The model uses a CNN for feature extraction and a LSTM for language processing. The features are combined using element wise multiplication and classifies one of the answers. Over the last few years, a large number of deep neural network has been proposed to improve the performance of VQA system. Moreover, attention based approaches are widely using to solve various sequence learning tasks including VQA. The goal of attention module is to identify the most salient part of image or textual content. Yang et al. 2016 introduced an attention network to support multi-step reasoning for the image QA task. A combination of bottom-up and top-down attention mechanism is presented in (Anderson et al., 2018). A set of salient image regions are proposed by bottom-up attention mechanism using Faster R-CNN. On the other hand, task specific context is used to predict an attention distribution by top-down mechanism over the image regions. A model-agnostic framework is proposed in (Shah et al., 2019) which relies on cycle consistency without any additional annotation to learn

VQA model. Their model not only answer the question but also generate 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.

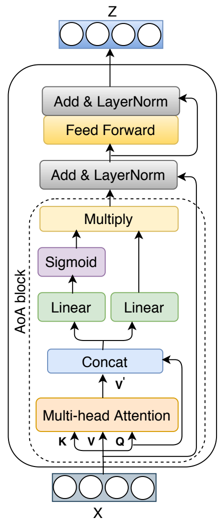
Recently, co-attention based approaches are becoming popular to the computer vision and NLP researchers. The goal of co-attention model is to learn image and question attention simultaneously. Lu et al. 2016 proposed a co-attention network that jointly reasons about image and question attention in a hierarchical fashion. Yu et al. 2018 reduced irrelevant features by applying self attention for question embedding and question conditioned attention for image embedding. A multi-steps dual attention for multimodal reasoning and matching is presented in (Nam et al., 2017). One major limitation of these co-attention based approaches is lack of dense interactions between different modality. To cope up this limitation, dense co-attention based methods are proposed in (Yu et al., 2019; Kim et al., 2018). Motivated by Yu et al. 2019, in this paper we proposed an improved architecture based on AoA module (Huang et al., 2019) to deliver significantly better performance on VQA system.

3 Our Approach

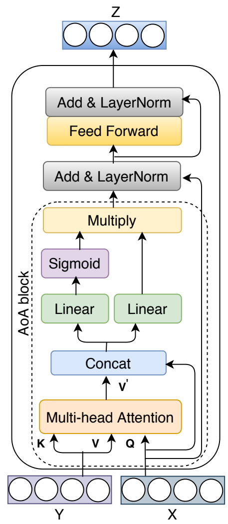
Motivated by Yu et al. 2019 and Huang et al. 2019, in this paper we present Modular Co-Attention on Attention Network (MCAoAN). MCAoAN consists of Modular Co-Attention on Attention (MCAoA) layer which is a modular composition of two primary attention units: self Attention on Attention (SAoA) and guided Attention on Attention (GAoA) unit. In this section, we first discuss SAoA and GAoA unit in sec. 3.1 followed by Modular Co-Attention on Attention (MCAoA) layer in sec. 3.2. Lastly we present our MCAoAN with multimodal attention-based fusion mechanism in sec. 3.3.

3.1 SAoA and GAoA Units

Our SAoA unit (see figure 2(a)) is an extension of multi-head self attention mechanism (Yu et al., 2019). Multi-head attention consists of N number of parallel heads where each head can be represented as a scaled dot product attention function as follows:



(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.

$$f_{att} = f(Q, K, V) = Softmax(\frac{QK}{\sqrt{d}})V \quad (1)$$

Where attention function $f(Q, K, V)$ 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 Softmax. The normalized scores are then used with V to generate weighted average vector V'. Here, d is the

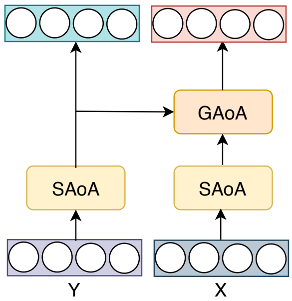


Figure 3: MCAoA layer where Y and X denotes question and image features respectively.

dimension of Q and K where both are 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 Huang et al. 2019, 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:

$$I = W_Q Q + W_{V'} V' + b_I \tag{2}$$

$$G = \sigma(W_G Q + W_G V' + b_G) \tag{3}$$

Here, W_Q, $W_{V'}$, W_G, $W_G \in \mathbb{R}^{d \times d}$ and b_I, b_G $\in \mathbb{R}^d$. d is the dimension of Q and V' where V' = 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 Yu et al. 2019, 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. $\mathbf{FC(4d)} - \mathbf{ReLU} - \mathbf{dropout(0.1)} - \mathbf{FC(d)}$).

3.2 MCAoA layers

Modular Co-Attention on Attention (MCAoA) layer (see figure 3) consists of two attention units discussed in sec. 3.1. 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 feature, MCAoA layer first uses two separate SAoA unit to caption intra-modal interactions for X and Y separately and later uses GAoA unit to capture inter-modal relationships where Y guides X feature. Here X and Y represents image and question feature respectively.

3.3 MCAoAN

In this section, we discuss our proposed modular co-attention on attention network (MCAoAN) (see figure 4) which is motivated by Yu et al. 2019. First we have to pre-process the inputs (i.e. image and query question) into appropriate feature representations. We use Faster R-CNN (Ren et al., 2015) with ResNet-101 as its backbone which is pretrained on Visual Genome dataset (Krishna et al., 2017) to process input images. The intermediate feature of the detected object from Faster R-CNN is considered as visual feature representation. Following Yu et al. 2019, 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 pretrained GloVe embedding (Pennington et al., 2014) 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 (Hochreiter and Schmidhuber, 1997) 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 Yu et al. 2019, 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.

Multi-modal Attention Fusion. The outputs

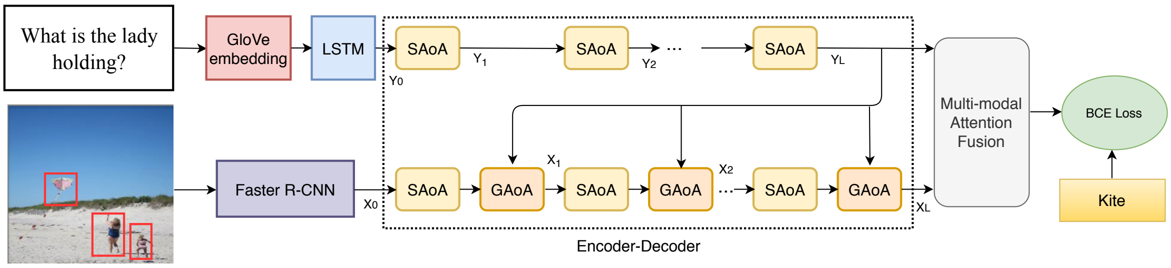


Figure 4: Overall architecture of our proposed Modular Co-Attention on Attention Networks (MCAoAN).

(i.e image features $X_L = [x_1, x_2, ..., x_m] \in \mathbb{R}^m$ and question features $Y_L = [y_1, y_2, ..., y_n] \in \mathbb{R}^n$) 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 a multi-modal attention fusion network (see figure 5) to aggregate features of both modality. Following Yu et al. 2019, 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:

$$X' = \sum_{i=1}^{m} softmax(MLP(X_L))x_i \tag{4}$$

and

$$Y' = \sum_{i=1}^{n} softmax(MLP(Y_L))y_i \tag{5}$$

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 (Mees et al., 2016), we apply concatenation on X' and Y' followed by a series of fully-connected layers (i.e. $FC(1024) - Dropout(0.2) - FC(512) - Dropout(0.2) - FC(2) - softmax$). 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 together and feed to a $LayerNorm$ to stabilize the training followed by a fully connected layer and sigmoid activation to generate final predicted answer Z. We use binary cross-entropy loss (BCE) to train the network.

4 Experiments

In this section, first we describe the dataset (see sec. 4.1) used in our experiments. Then we present experimental setup and implementation details in sec. 4.2. Lastly, we discuss experimental results in sec. 4.3.

4.1 Datasets

To evaluate our method, in this paper we use VQA-v2 benchmark dataset (Goyal et al., 2017) which consists of images from MS-COCO dataset (Lin et al., 2014) with human annotated question-answer pairs. There are 3 questions for each image and 10 answers 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 Experimental Setup and Implementation Details

To evaluate our method, we follow the experimental protocol proposed by Yu et al. 2019. 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. 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 $10 \leq T$.

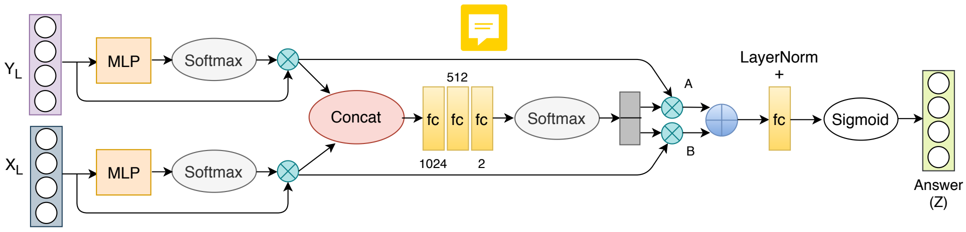


Figure 5: Illustration of our proposed multimodal attention fusion network.

Methods	**All**	**Other**	**Y/N**	**Num**
MCAN (Yu et al., 2019)	81.20	73.73	95.86	67.30
Ours (MCAoA)	82.91	75.92	96.47	70.38
Ours (MCAoA + Multi-modal Attention Fusion)	**83.25**	**76.51**	**96.58**	**70.40**

Table 1: Experimental results and comparison of our proposed method with other state-of-the-art method on validation set.

Methods	**All**	**Other**	**Y/N**	**Num**
Bottom-up (Teney et al., 2018)	65.32	56.05	81.82	44.21
MFH (Yu et al., 2018)	68.76	59.89	84.27	49.56
BAN (Kim et al., 2018)	69.52	60.26	85.31	50.93
BAN+Counter (Kim et al., 2018)	70.04	60.52	85.42	54.04
MCAN (Yu et al., 2019)	70.63	60.72	86.82	53.26
Ours (MCAoA)	**70.90**	60.97	**87.05**	**53.81**
Ours (MCAoA + Multi-modal Attention Fusion)	70.84	**61.01**	86.96	53.45

Table 2: Experimental results with other state-of-the-art models on Test-dev.

Methods	**All**	**Other**	**Y/N**	**Num**
Bottom-up (Teney et al., 2018)	65.67	-	-	-
BAN+Counter (Kim et al., 2018)	70.35	-	-	-
MCAN (Yu et al., 2019)	70.90	-	-	-
Ours (MCAoA)	71.14	61.18	87.25	**53.36**
Ours (MCAoA + Multi-modal Attention Fusion)	**71.16**	**61.23**	**87.29**	53.19

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

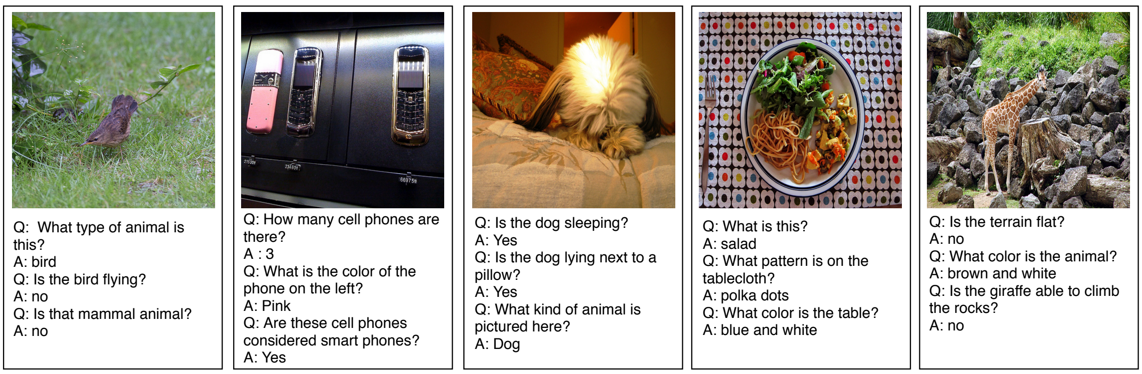


Figure 6: Illustration of some qualitative results from test set using our method. Here Q and A represents query question and generated answer respectively.

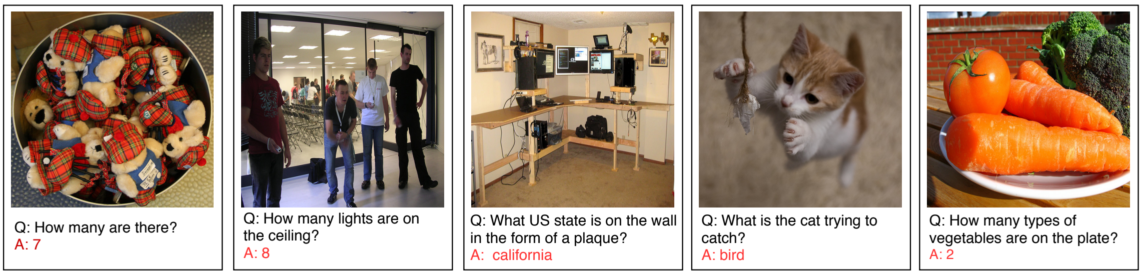


Figure 7: Illustration of some failure cases from test set using our method. Here Q and A represents query question and predicted wrong answer (mark as red) respectively.

4.3 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 (Yu et al., 2019)[1] and compare the results with our method on validation set in table 1. Table 2 and 3 shows experimental results using test-dev and test-std respectively using online evaluation [2]. Offline evaluation only supports on validation split. Figure 6, shows some qualitative results using our method on test set. From the experimental results, we can see that our proposed method outperforms other state-of-the-art methods on VQA. Beside that, we also present some typical failure cases using our method in figure 7.

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 a novel multi-modal fusion architecture. Experimental results show that each component within our model improve the performance of VQA system. Moreover, The final network achieves state-of-the-art performance on VQA-v2 dataset.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

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

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

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

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

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

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

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

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

[1] https://github.com/MILVLG/mcan-vqa
[2] https://evalai.cloudcv.org/web/challenges/challenge-page/163/overview

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

Oier Mees, Andreas Eitel, and Wolfram Burgard. 2016. 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.

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

Duy-Kien Nguyen and Takayuki Okatani. 2018. 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.

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

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

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

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

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

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

Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. 2018. 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.

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

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

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

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959.