Visual Question Answering using Explicit Visual Attention

Vasileios Lioutas, Nikolaos Passalis and Anastasios Tefas
Dept. of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
Email: lioutasv@csd.auth.gr, passalis@csd.auth.gr, tefas@aiia.csd.auth.gr

***Abstract*—One of the most complex multi-model problems faced today is Visual Question Answering (VQA), which requires a machine to properly understand a question about a reference visual input, expressed in natural language, and then produce the answer to that question. In order to solve this problem and increase the probability of producing the correct answer, it is crucial to provide reliable attention information. However, existing methods only use implicitly trained attention models that are often unable to attend to the appropriate image region the question refers to, limiting their ability to provide the correct answer. To address this issue, we propose an explicitly trained attention model that is inspired by the theory of pictorial superiority effect. In this model, we use attention-oriented word embeddings that increase the efficiency of learning common representation spaces. The dataset that we use, the Visual7W dataset, is the only dataset that provides visual attention ground truth information. In this paper, we demonstrate the effectiveness of the proposed method over both implicit attention models and other state-of-art VQA techniques.**

I. Introduction

Over the past few years, high-level reasoning in terms of image comprehension has become one of the most challenging tasks in the field of Artificial Intelligence and it has received considerable attention, leading to a great amount of research actively pursuing it, not only in academia but in the industry as well. Natural Language Processing [1], [2], and Computer Vision [3], [4], as well as the rapidly increasing available computational power, have provided researchers the tools necessary to tackle the problem of building machines that interlink multiple modalities [5]–[7]. Visual Question Answering (VQA) [8]–[10], in particular, has become one of the most prominent multi-modal problems with a plethora of researchers working on it. VQA requires a machine to properly understand a question, posed in natural language, about a reference visual input, i.e., an image, and then to infer the correct answer.

Attention mechanisms attempt to provide fine-grained information with respect to a visual content and the task at hand [8], [11]. Those mechanisms work based on the assumption that image regions, which provide information relevant to the corresponding question, will develop stronger associations with that question [8], [11]. On the other hand, image regions that are irrelevant to the question will exhibit diminished associations to the corresponding question [8], [11]. Therefore, these attention models are trained implicitly, i.e., there is no prior information for the correct attention regions. However, it was recently shown that using explicitly trained attention models can significantly improve the accuracy for the task of automatic caption generation [12], which is a similar multi-modal problem. This is also demonstrated in Figure 1.

Fig. 1: Example of explicit attention. The proposed method provides better attention information given a question than the implicit attention model.

In this paper we proposed a novel explicitly-trained attention model for VQA tasks. Taking into consideration the pictorial superiority effect, which states that people tend to recall images better than words [13], [14], we propose using separate word embeddings for the attention model that are independent from the embeddings that are used for answering questions. This way, learning common representation spaces where each word is closer to the visual representation of its semantic content is facilitated. Finally, we use the Visual7W dataset [8], which is the only dataset that provides visual attention ground information, to evaluate the proposed method. To the best of our knowledge, this is the first work that exploits the visual attention ground information of the Visual7W to train more accurate attention models. Finally, using extensive experiments, it is demonstrated that the proposed method is more effective compared to implicit attention models or other proposed VQA techniques.

The rest of the paper is structured as follows. The related work is briefly discussed and compared to our approach in Section II. The proposed method is presented in Section III, while the experimental evaluation is provided in Section IV. Finally, conclusions are drawn and future work is discussed in Section V.

II. Related Work

The methods proposed for tackling VQA tasks can be divided into two categories: the fist one is composed of *generative* methods, where the answer is generated in free-form text, while the second one of *classification*-based methods, where the correct answer is chosen among a set of predefined answers. Most generative methods use recurrent

models, such as Long Short-Term Memory Units (LSTMs), to answer the question at hand [6], [15], [16]. Generating the question in free-form text also complicates the evaluation procedure, since multiple possible answers can be correct given the same question [17]. The classification-based methods extract features from the input modalities and then use a classifier to determine the correct answer [17]–[20]. Some of these methods also use recurrent models to provide better encoding of the input modalities [8], [16], [21]. However it worths mentioning that it was recently found that using a simple triplet (question-answer-image) model [17] can actually improve the precision of the model over most other more complicated methods proposed in the literature. In this work, we also use a triplet-based classification model (more details are given in Section III) that is extended with the proposed explicit attention model.

There is also a rich literature on using implicit attention models to improve visual analysis tasks. Implicit attention models usually work by learning weighting coefficients (or a probability distribution) over each image region (as expressed through the extracted feature maps, if a Convolutional Neural Network (CNN) is used) to improve the accuracy of the models for the task at hand, e.g., [11], [22]–[25]. Implicit attention have been also used to tackle VQA tasks [8]. Using explicit attention, i.e., using an attention module that was trained with ground truth human attention information, has been investigated for use for caption generation tasks in [12], and it was shown to improve the accuracy of the model over using plain implicit models. To the best of our knowledge we propose the first explicit attention model for dealing with VQA tasks.

III. Proposed Method

In this Section we introduce the used notation and we describe the proposed explicit attention model as well as the complete pipeline of the proposed visual question answering system in detail.

A. Explicit Attention Model

The gist of the proposed explicit attention model is to reduce the semantic gap between textual and image representations. To this end, we *directly* learn to attend the parts of an image that correspond to the given question in order to highlight only the regions that should be used to infer the answer. To further increase the flexibility of the attention mechanism, a separate word embedding model is used independently of the word embedding model that it is used to infer the answer to the question. This idea is inspired from the theory of pictorial superiority effect [13], [14], that states that "*human memory is extremely sensitive to the symbolic modality of presentation of event information*" [26]. It was also experimentally confirmed that decoupling the word representations used for visual tasks, e.g., for providing the attention, from the word representation used for the textual tasks, e.g., for answering the question at hand, improves the overall accuracy of the system.

The architecture of the proposed explicit attention model is summarized in Figure 2. Let $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ be a given question, where N is the total number of words in the question, $\mathbf{q}_i \in \mathbb{R}^{D_w}$ is the embedding vector for the i-th word, and

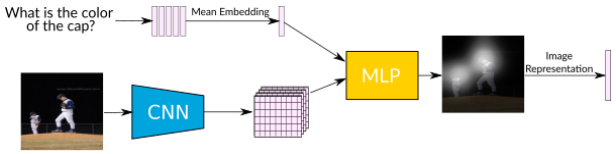

Fig. 2: The architecture of the proposed explicit attention model.

D_w is the dimensionality of the word embedding. Also, let $\mathbf{I}_m \in \mathbb{R}^{D_m \times D_m \times D_d}$ be the feature map used for providing the attention, where $D_m \times D_m$ is the size of the extracted feature map and D_d the number of filters used in the corresponding convolutional layer.

Given a question Q we first embed the words into a textual vector space using a word embedding model. After that, we extract the global question representation $\mathbf{Q}_f \in \mathbb{R}^{D_w}$ of the question Q by averaging the embeddings over all words of the question, where D_w is the dimensionality of the word embedding. The attention distribution $\mathbf{p_I}$ over the convolutional feature map $\mathbf{I}_m$ given the question Q is defined as:

$$\mathbf{h}_c = [\tanh(\mathbf{I}_m \times \mathbf{W}_I); \mathbf{1}_{D_m \times D_m \times 1} \times \tanh(\mathbf{Q}_f \times \mathbf{W}_Q)] \in \mathbb{R}^{D_m \times D_m \times 2D_c}, \quad (1)$$

$$\mathbf{p}_I = \text{softmax}(\text{relu}(\mathbf{h}_c \times \mathbf{W}_{h1}) \times \mathbf{W}_{h2}) \in \mathbb{R}^{D_m \times D_m}, \quad (2)$$

where $\mathbf{1}_{D_m \times D_m \times 1}$ is a matrix used for repeating $\tanh(\mathbf{Q}_f \times \mathbf{W}_Q)$ $D_m \times D_m$ times in $\mathbf{h}_c$, and $\mathbf{W}_I \in \mathbb{R}^{D_d \times D_c}$ and $\mathbf{W}_Q \in \mathbb{R}^{D_w \times D_c}$ are the projection weights used for constructing a common representation space (D_c is the dimensionality of this space). Equation (2) provides the attention distribution over the image regions (as expressed through the extracted feature map). Note than a simple Multilayer Perceptron (MLP) with D_h hidden units is used to this end, i.e., $\mathbf{W}_{h1} \in \mathbb{R}^{(2D_c) \times D_h}$ and $\mathbf{W}_{h2} \in \mathbb{R}^{D_h \times 1}$. We use the extracted attention distribution $\mathbf{p}_I \in \mathbb{R}^{D_m \times D_m}$ to provide the final attention representation:

$$\mathbf{I}_{m'} = \sum_{i=1}^{D_m} \sum_{j=1}^{D_m} \mathbf{p}_{I_{ij}} \mathbf{I}_{mij} \in \mathbb{R}^{D_d}. \quad (3)$$

In order to train the proposed explicit attention model, ground truth bounding boxes B_T that associate the correct answer with the different regions of the image are used. The ground truth attention target is set as:

$$\hat{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{||\boldsymbol{\alpha}||_0} \in \mathbb{R}^{D_m \times D_m}, \quad (4)$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{D_m \times D_m}]$ and

$$\alpha_t = \begin{cases} 1 & \text{if } t \text{ overlaps with any bounding box b } \in B_T \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

is the ground truth attention membership value of the t-th part of the extract feature map into the ground truth bounding box set B_T and $||\boldsymbol{\alpha}||_0$ is the number of 1s that exist in the membership vector. Note that we use nearest-neighbor

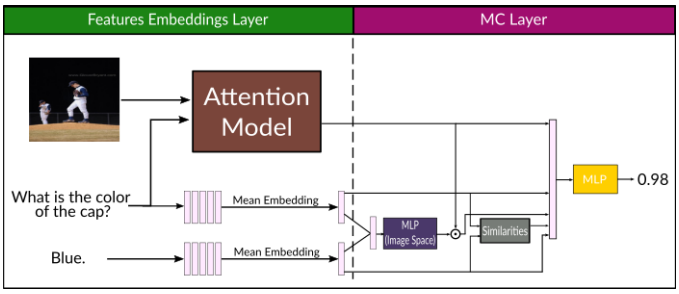

Fig. 3: The architecture of the proposed visual question answering model.

interpolation to assign each bounding box to the parts of the feature map that it belongs. Then, we train the model to attend the ground truth regions by minimizing the cross-entropy loss between the predicted attention distribution and the target attention distribution:

$$\mathcal{J}_{att} = -\sum_{i=1}^{D_m}\sum_{j=1}^{D_m} \hat{\alpha}_{ij} \log(p_{I\,ij}). \tag{6}$$

B. Visual Question Answering Model

Jabri *et al.* [17] proposed a simple baseline model for visual question answering and showed that using a binary classifier to predict whether a given question-image-answer triplet is correct can significantly improve the results over more complex models or trying to directly generate the correct answer using recurrent models. In this work we adopt a similar triplet-based evaluation setup. The architecture of the proposed visual question answering model that utilizes the proposed explicit attention model is shown in Figure 3. Note that the model consists of two parts, the *Feature Embedding layer* and the *Multiple Choice (MC) layer*.

In the feature embedding layer we extract representations from the input modalities. First, an explicit attention model is used to provide the attention vector. Then the question and the answer are encoded using the average embedding vector, similarly to the approach used in [17]. The notation $\mathbf{Q}_f$ and $\mathbf{A}_f$ is used to refer to these embedding vectors. However, note that in contrast to previous works we use separate embedding models for the textual tasks, i.e., predicting whether the given answer is correct, and for the visual tasks, i.e., providing the attention distribution.

After extracting feature vectors from the input modalities we use an MLP to predict whether the given question-answer-image triplet is correct. Therefore, the MC layer outputs a scalar value that indicates the correctness of the given input question-answer-image triplet. Instead of directly feeding the extracted feature vectors into the used MLP we also calculate the similarity and the distance between the representation of the image, the question and the answer. Therefore, the following vector is fed into the final classifier:

$$[\mathbf{Q}_f; \mathbf{A}_f; \mathbf{Q}_f \odot \mathbf{A}_f; \|\mathbf{Q}_f - \mathbf{A}_f\|; \mathbf{I}_{m'}; \mathbf{I}_{m'} \odot \mathbf{z}] \tag{7}$$

where $\odot$ is the Hadamard product operator, $\mathbf{I}_{m'} \in \mathbb{R}^{D_d}$ denotes the attention representation vector extracted from the attention model and $\mathbf{z}$ is the result of the transformation layer that transforms the concatenated vector of the question and the answer into a common representation space. The output of this transformation layer is computed as:

$$\mathbf{t}_{qa} = [\mathbf{Q}_f; \mathbf{A}_f] \in \mathbb{R}^{2D_w} \tag{8}$$

$$\mathbf{z}^{(n)} = \sigma(\mathbf{t}_{qa}\mathbf{W}_{qa}^{(n)} + \mathbf{b}_{qa}^{(n)}) \in \mathbb{R}^{D_d} \tag{9}$$

where $\mathbf{W}_{qa}^{(n)}$ and $\mathbf{b}_{qa}^{(n)}$ are the parameters of the transformation layer and $\sigma(\cdot)$ denotes the sigmoid activation function.

After computing the aforementioned input vector we use an MLP with 8096 hidden units, rectifier activation functions in the hidden layer and sigmoid activation function for the final output to predict the correctness score for the input triplet. The proposed model was optimized by minimizing the binary logistic loss.

IV. Experiments

We evaluate the proposed model on the Visual7W Telling dataset [8], which is a subset of the Visual Genome dataset [27]. The dataset contains 69,817 training questions, 28,020 validation questions and 42,031 test questions. Each question comes with 4 possible answers of which only one is correct. The negative choices are human-generated and the performance is measured by the percentage of correctly answered questions. In addition, this dataset contains visual bounding boxes for the images that are associated with the answer of each question (attention ground truth information). This allows us to train explicit attention models with the supplied annotations. Note that only a fraction of the questions are annotated with bounding boxes that can be used for training the explicit attention model (30,491 training questions, 12,103 validation questions and 18,253 test questions).

For developing the proposed model we used the theano library [28], and the Lasagne framework [29]. We use the Adam optimizer with the default settings [30]. For the explicit attention model we use a learning rate of 0.001 and for the multiple choice answering model a learning rate of 0.0001. The batch size for training both models is set to 16. We apply dropout with probability of 0.2 and batch normalization on MC Layer. The explicit attention model was trained for 5 epochs and the multiple choice answer model was trained for 12 epochs using the training and validation sets. For extracting the convolutional feature map we used the pre-trained deep residual network, ResNet-152 [24]. The feature maps were extracted from the last convolutional layer of the network. For extracting textual representations we used pre-trained GloVe embedding vectors (Common Crawl (42B tokens), 300d) [1]. Note that the GloVe embeddings were used only for initialization and then they were optimized during the training. For measuring the performance of the developed model we followed the procedure described in [8], using the toolbox supplied by the authors of [8].

The evaluation results are shown in Table I. The accuracy of the models for each question type is shown in columns 2-7, while the overall accuracy is shown in the last column. The proposed explicit attention model achieves higher overall

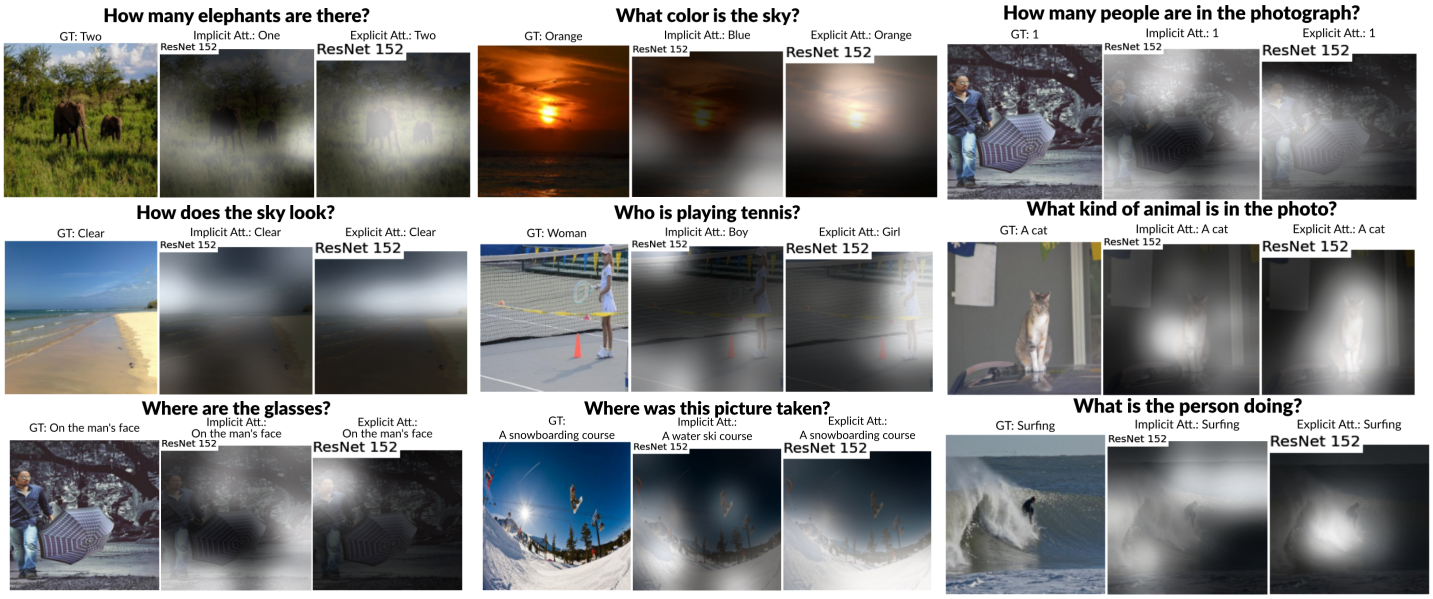

Fig. 4: Comparing between implicit attention and explicit attention models.

TABLE I: Comparing the proposed explicit attention method to implicit attention

Method	What	Where	When	Who	Why	How	Overall
Implicit	0.617	0.706	0.801	0.693	0.602	0.532	0.634
Proposed	**0.642**	**0.748**	**0.825**	**0.729**	**0.623**	**0.536**	**0.659**

TABLE II: Comparing the proposed method to other baseline and state-of-the-art VQA techniques

Method	Overall
Human (Question + Image) [8]	0.966
Logistic Regression (Q + I) [8]	0.352
LSTM (Q + I) [15]	0.521
LSTM-Att [8]	0.556
MCB [31]	0.622
Triplet MLP [17]	**0.671**
Proposed	**0.659**

question answering accuracy over the baseline implicit attention model. Using explicit attention increases the answering accuracy for every question type (especially for the "what" and "why" questions where providing reliable attention is crucial). This fact is also confirmed in Figure 1 and Figure 4, where the implicit attention model and the proposed explicit attention model are compared using some of the test question and images. It is evident that the proposed method significantly improves the attention accuracy. The explicit attention also significantly improves the "How many"-type questions. This can be better understood from the question "How many elephants are there", where it is evident that attending to the correct region of the image is vital for correctly answering the question. Similar conclusions can be drawn for the rest of the images. Finally, the proposed method is compared to other baseline and state-of-the-art VQA techniques in Table II. The proposed method achieves the second higher VQA accuracy. We tried to use the network architecture proposed in the Triplet MLP method (which provides the best baseline accuracy) [17], but we were unable to reproduce the reported results, since not all the details of the used setup are reported. However, combining the explicit attention model with the exact setup used in [17] is expected to further improve the accuracy (as already demonstrated in Table I using a weaker baseline model).

V. Conclusions

In this paper, we demonstrated that using an explicitly trained attention model the VQA accuracy can be significantly improved compared to other implicit attention models and VQA techniques. In addition, we developed a mechanism that it is inspired by the pictorial superiority effect and further improves answering accuracy. This paper paves the way for multiple interesting future directions, including techniques that could be used to combine multiple attention models, similar to other ensemble models [32]. In addition, in the proposed method the attention model is not trained if a question does not contain ground truth bounding boxes. Exploiting the information contained in these image-question pairs, in a way similar to the implicit attention, can lead to a hybrid implicit-explicit attention model that can further improve the visual question answering accuracy. Furthermore, a pyramid Bag-of-Features (BoF)-based representation can be extracted, e.g., using the techniques proposed in [33], and [34], to provide fine-grained visual information and further increase the VQA accuracy.

VI. Acknowledgments

Nikolaos Passalis was financially supported by the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (PhD Scholarship No. 1215).

REFERENCES

[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[2] N. Passalis and A. Tefas, "Bag of embedded words learning for text retrieval," in *Proceedings of the 23rd International Conference on Pattern Recognition*, 2016, pp. 2416–2421.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] N. Passalis and A. Tefas, "Concept detection and face pose estimation using lightweight convolutional neural networks for steering drone video shooting," in *Proceedings of the European Signal Processing Conference*, 2017, pp. 71–75.

[5] A. N. Venkitasubramanian, T. Tuytelaars, and M.-F. Moens, "Wildlife recognition in nature documentaries with weak supervision from subtitles and external data," *Pattern Recognition Letters*, vol. 81, pp. 63–70, 2016.

[6] Q. Wu, C. Shen, A. v. d. Hengel, P. Wang, and A. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," *arXiv preprint arXiv:1603.02814*, 2016.

[7] A. S. Toor and H. Wechsler, "Biometrics and forensics integration using deep multi-modal semantic alignment and joint embedding," *Pattern Recognition Letters*, 2017.

[8] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded Question Answering in Images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," *CoRR*, vol. abs/1505.00468, 2015.

[10] M. Ren, R. Kiros, and R. S. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *CoRR*, vol. abs/1505.02074, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.

[12] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention correctness in neural image captioning," *CoRR*, vol. abs/1605.09553, 2016.

[13] Nelson Douglas L., Reed Valerie S., and Walling John R., "Pictorial superiority effect." *Journal of Experimental Psychology: Human Learning and Memory*, vol. 2, no. 5, pp. 523–528, 1976.

[14] P. Miller, "The processing of pictures and written words: A perceptual and conceptual perspective," *Psychology*, no. 2, pp. 713–720, 2011.

[15] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1–9.

[16] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[17] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting visual question answering baselines," *CoRR*, vol. abs/1606.08390, 2016.

[18] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[19] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," *arXiv preprint arXiv:1506.00333*, 2015.

[20] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[21] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[25] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.

[26] J. Yuille, *Imagery, Memory and Cognition (PLE: Memory): Essays in Honor of Allan Paivio*, ser. Psychology Library Editions: Memory. Taylor & Francis, 2014.

[27] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016.

[28] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688.

[29] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri *et al.*, "Lasagne: First release." Aug. 2015.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[31] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[32] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[33] N. Passalis and A. Tefas, "Learning bag-of-features pooling for deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[34] ——, "Neural bag-of-features learning," *Pattern Recognition*, vol. 64, pp. 277–294, 2017.