Explicit Ensemble Attention Learning for Improving Visual Question Answering

Vasileios Lioutas[a], Nikolaos Passalis[a,*], Anastasios Tefas[a]

[a] *Dept. of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece*
lioutasv@csd.auth.gr, passalis@csd.auth.gr, tefas@aiia.csd.auth.gr

Abstract

Visual Question Answering (VQA) is among the most difficult multi-modal problems as it requires a machine to be able to properly understand a question about a reference image and then infer the correct answer. Providing reliable attention information is crucial for correctly answering the questions. However, existing methods usually only use implicitly trained attention models that are frequently unable to attend to the correct image regions. To this end, an explicitly trained attention model for VQA is proposed in this paper. The proposed method utilizes attention-oriented word embeddings that allows efficiently learning the common representation spaces. Furthermore, multiple attention models of varying complexity are employed as a way of realizing a mixture of experts attention model, further improving the VQA accuracy over a single attention model. The effectiveness of the proposed method is demonstrated using extensive experiments on the Visual7W dataset that provides visual attention ground truth information.

1. Introduction

Due to the recent developments in the Natural Language Processing and Computer Vision areas, in combination with the rapidly increasing computational power, significant research efforts have been focusing on tackling the problem of building machines that interlink multiple modalities [34, 35, 33]. One of the most prominent multi-modal problems is the task of Visual Question Answering (VQA) [42, 4, 29], which has become one of the most active

[*] Corresponding author:
Email address: passalis@csd.auth.gr (Nikolaos Passalis)

research directions, not only in academia but also in the industry. VQA requires a machine to be able to properly understand a question about a reference image and then infer the correct answer.

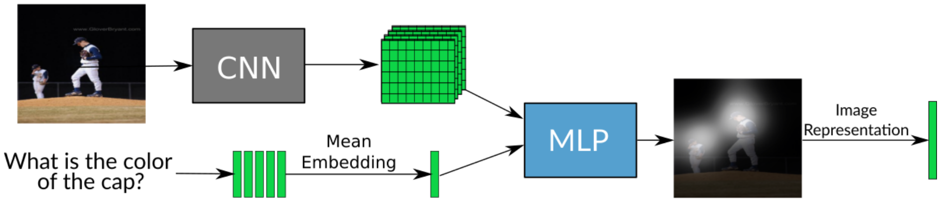


Figure 1: The architecture of the proposed explicit attention model.

To provide fine-grained information regarding the visual content, *attention mechanisms* have been developed [42, 17]. These methods work under the assumption that the image regions, which contain relevant information to the question at hand, will eventually be strongly associated with the corresponding questions, while the irrelevant regions will exhibit diminishing association. However, these attention mechanisms are trained *implicitly*. It was recently demonstrated that utilizing explicitly trained attention models can improve the accuracy of automatic caption generation [19]. Even though a recently released VQA dataset, the Visual7W dataset [42], contains attention ground truth information for some of the available questions, no technique has yet exploited this information to train more accurate attention models and evaluate their performance using the Visual7W dataset.

The paper proposes a novel explicit attention model for VQA tasks. Inspired by the theory of the pictorial superiority effect, we propose employing separate word embeddings for the attention model that is independent from the embeddings, which are used for answering the questions. The theory of pictorial superiority effect refers to the phenomenon of humans remembering images easier compared to words [24, 23]. Thus, it is easier to learn common representation spaces, where each word is closer to the visual representation of its semantic content. Finally, recognizing the difficulty of training reliable attention models we use multiple attention models of varying complexity as a way of realizing a mixture of experts attention model [22] that is able to provide more accurate answers than a single attention model. We demonstrate the effectiveness of the proposed method, over both implicit attention models as well as other state-of-the-art VQA techniques, using the Visual7W dataset [42].

The rest of the paper is structured as follows. The prior work is discussed in Section 2 and the proposed method is presented in Section 3. The experimental evaluation is presented in Section 4 and conclusions are drawn in Section 5.

2. Related Work

Visual Questioning Answering (VQA) methods fall into two categories: a) the *generative* methods, in which the answer is generated in free-form text, and b) the *classification*-based methods, in which the correct answer is chosen among a set of predefined answers. The generative methods usually employ recurrent models, such as Long Short-Term Memory Units (LSTMs), to generate the answer to the question [35, 21, 3]. However, generating the answer in free-form text significantly complicates the evaluation procedure, since there are multiple correct answers for the same question [12]. On the other hand, classification-based methods extract features from the input modalities and then employ a classifier to determine the correct answer [12, 40, 20, 2]. Many of these method, e.g., [42, 3, 28] also utilize recurrent models to provide better encoding of the input modalities. However, it is worth mentioning, as it was recently established, that employing a simple triplet-based scheme (question-answer-image) [12] can significantly improve the answering accuracy over the rest of the methods proposed in the literature. In this work, a triplet-based classification scheme is also combined with the proposed explicit attention model. The interested reader is referred to [36, 13] for an extensive review of VQA methods as well as of the currently available VQA datasets.

A rich literature on using implicit attention models to improve visual analysis tasks also exist. These models work by learning weighting coefficients (or a probability distribution) over the extracted feature maps. Implicit attention model are capable of improving the accuracy of the models for various tasks, e.g., [17, 31, 30, 10, 38], including VQA tasks [42]. An attention model that was trained with ground truth human attention information (explicit attention) has been applied for caption generation tasks in [19], and it was shown to improve the accuracy compared to implicit attention models. Also, an extensive discussion regarding the differences between implicit attention models and human attention is provided in [6]. This work also highlights the potential of utilizing explicitly trained attention models for the task of VQA.

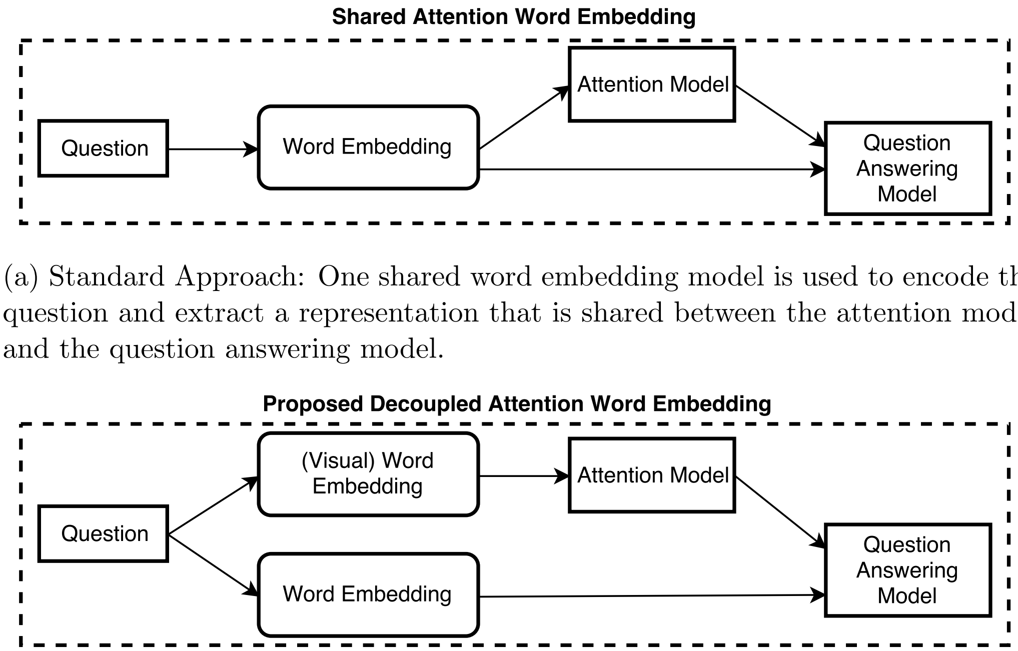
To the best of our knowledge, in this paper we propose the first explicitly trained *ensemble* attention model for VQA tasks that is capable of utilizing multiple attention distributions generated by models of varying complexity. Another explicit attention model was proposed by Qiao et al. [27]. This model used the multimodal low-rank bilinear pooling (Kim et al., 2017) to provide several smaller attention maps that were then applied to infer the final attention distribution. In contrast to these approaches, our method is capable of combining several different attention distributions that are provided by *multiple* attention models. This increases the probability of attending to the correct image regions. The ability of our ensemble approach to increase the question answering accuracy is experimentally demonstrated in Section 4. Also, inspired by the pictorial superiority phenomenon, we propose a biologically justified approach that decouples the attention process from the answering process utilizing two separate word embeddings. This further increases the expressive power of the proposed attention model. Finally, instead of utilizing bilinear pooling, we employ a simpler and more lightweight correlation approach through a series of non-linear operators (*tanh* and *relu*).

3. Proposed Method

The used notation is introduced and the proposed explicit attention model, along with the complete pipeline of the proposed visual question answering system, are described in detail in this Section.

3.1. Explicit Attention Model

The architecture of the proposed explicit attention model is summarized in Figure 1. The goal of the proposed model is to reduce the semantic gap between textual and image representations. To achieve this, the proposed method *directly* learns to attend the parts of an image that correspond to the given question using the supplied ground truth information. Thus, only the image regions that are actually related to the question at hand are used to infer the correct answer. Instead of utilizing the same word embedding for providing both the correct answer and the attention information, two separate word embedding models are employed as shown in Figure 2. In that way, it is possible to learn a separate word embedding model that it is only utilized for providing the visual attention information and another one for providing the correct answer. This decoupling allows for increasing the expressive power of the attention model that is equipped with a separate (visual oriented) word embedding model, which does not tie to the word embedding employed for providing the correct answers. We inspired this idea from the theory of pictorial superiority effect [24, 23] that states that "human memory is extremely sensitive to the symbolic modality of presentation of event information" [39]. It was also experimentally shown that decoupling the word representations used for providing the attention from the word representations utilized for answering the question at hand improves the overall accuracy of the system.



(a) Standard Approach: One shared word embedding model is used to encode the question and extract a representation that is shared between the attention model and the question answering model.

(b) Proposed Approach: Two separate word embedding models are employed to extract two different representations of the question that are not shared between the attention model and the question answering model.

Figure 2: Using two separate word embedding models, instead of one shared model, allows to increase the expressive power of the attention model that is now equipped with a dedicated visual-oriented word embedding model. This idea was inspired by the theory of pictorial superiority effect.

Consider the proposed explicit attention model (Figure 1). Let $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ denote a question, where N is the number of words in the question, $\mathbf{q}_i \in \mathbb{R}^{D_w}$ is the embedding vector for the i-th word, and D_w

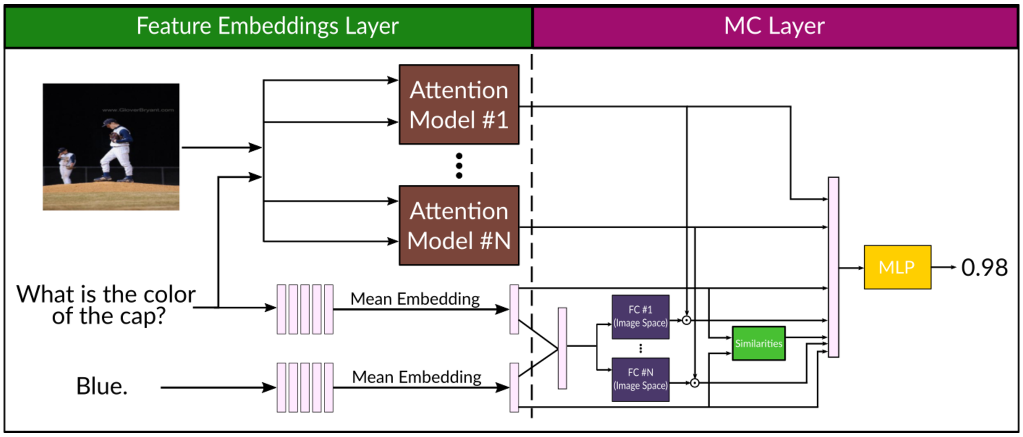


Figure 3: The architecture of the proposed visual question answering model.

denotes the dimensionality of the word embedding. Also, the notation $\mathbf{I}_m \in \mathbb{R}^{D_m \times D_m \times D_d}$ is used to refer to the feature map utilized for providing the attention, where $D_m \times D_m$ is the size of the extracted feature map and D_d is the number of convolutional filters.

First, the words of a question Q are embedded into a textual vector space employing a word embedding model. After that, the representation $\mathbf{Q}_f \in \mathbb{R}^{D_w}$ of the question Q is extracted by averaging the word embedding vectors extracted from the question Q, where D_w is the dimensionality of the word embedding. Then, the attention distribution $\mathbf{p}_I$ over the convolutional feature map $\mathbf{I}_m$, for a given the question Q, is calculated as:

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

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

where $\mathbf{1}_{D_m \times D_m \times 1}$ is a matrix used for stacking $\tanh(\mathbf{Q}_f \times \mathbf{W}_Q)$ $D_m \times D_m$ 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 weights employed for projecting the question into a common representation space. The dimensionality of the common representation space is controlled by D_c. Equation (2) provides the attention distribution over the image regions, as they are expressed through the extracted feature map. Note that a Multilayer Perceptron (MLP) with D_h hidden units is utilized to provide the attention distribution, where $\mathbf{W}_{h1} \in \mathbb{R}^{(2D_c) \times D_h}$ and $\mathbf{W}_{h2} \in \mathbb{R}^{D_h \times 1}$ denote the weight matrices of the MLP. Finally, the extracted attention distribution $\mathbf{p}_I \in \mathbb{R}^{D_m \times D_m}$ is employed to provide the attention-based representation:

$$\mathbf{I}_{m'} = \sum_{i=1}^{D_m} \sum_{j=1}^{D_m} \mathbf{p}_{Iij} \mathbf{I}_{mij} \in \mathbb{R}^{D_d}. \tag{3}$$

The ground truth bounding boxes B_T, which associate the correct answer with different regions of the image, are employed to train the proposed explicit attention model. The attention targets are defined as follows:

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

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

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

is the ground truth attention membership value of the t-th part of the extracted 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. Nearest-neighbor interpolation is utilized to assign each bounding box to the parts of the feature map where it belongs. Then, the model is trained to attend the ground truth regions by minimizing the cross-entropy loss between the predicted attention distribution and the target attention distribution:

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

3.2. Visual Question Answering Model

A simple baseline model for visual question answering was proposed by Jabri *et al.* [12] and it was demonstrated that utilizing a binary classifier to predict whether a given question-image-answer triplet is correct can significantly improve the VQA accuracy over more advanced techniques, such as generating the correct answer utilizing recurrent models. In this work, we adopt a similar triplet-based scheme. As shown in Figure 3, where the architecture of the proposed visual question answering model is illustrated, the proposed model consists of two parts, the *Feature Embedding layer* and the *Multiple Choice (MC) layer.*

The feature embedding layer is responsible for extracting representations from the input modalities. First, multiple explicit attention models are utilized to provide different attention vectors. As demonstrated in Section 4, employing multiple attention models can improve the accuracy of visual question answering since models of different complexity tend to complement each other (similar to ensemble models [22]). Then, the question and the answer are encoded using the average embedding vector, similarly to the approach applied in [12]. The notation $\mathbf{Q}_f$ and $\mathbf{A}_f$ is used to refer to these embedding vectors. However, in contrast to other previous works, separate embedding models are employed for predicting whether the given answer is correct and for providing the attention distribution.

Table 1: Comparing the proposed explicit attention method to implicit attention

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.

Method	**What**	**Where**	**When**	**Who**	**Why**	**How**	**Ov**
Implicit	0.617	0.706	0.801	0.693	0.602	0.532	0
Proposed (ResNet 152)	0.642	**0.748**	**0.825**	**0.729**	0.623	0.536	0
Proposed (ResNet 101 + ResNet 152)	**0.656**	0.737	0.819	0.721	**0.644**	**0.547**	**0**

An MLP is then utilized in the MC layer to predict whether the given question-answer-image triplet is correct. This MLP outputs a scalar value that expresses to the *correctness* of the given input question-answer-image triplet. In that way, it is possible to choose the correct answer among several different question-answer pairs. Note that instead of directly feeding the extracted feature vectors into the utilized MLP, the similarity and the distance between the representation of the image, the question and the answer are also utilized. Thus, the vector that is fed into the final classifier is defined as:

$$[\mathbf{Q}_f;\mathbf{A}_f;\mathbf{Q}_f\odot\mathbf{A}_f;\|\mathbf{Q}_f-\mathbf{A}_f\|;\mathbf{I}^{(1)}_{m'};\mathbf{I}^{(1)}_{m'}\odot\mathbf{z}^{(1)};...;\mathbf{I}^{(N)}_{m'};\mathbf{I}^{(N)}_{m'}\odot\mathbf{z}^{(N)}], \tag{7}$$

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

The output of the MC 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, an MLP with 8096 hidden units, rectifier activation functions in the hidden layer and sigmoid activation function for the final output is employed to predict the correctness score for the input triplet. The binary logistic loss was used to optimize the proposed model.

4. Experiments

The proposed method was evaluated on the Visual7W Telling dataset [42], which is a subset of the Visual Genome dataset [16]. The dataset contains 69,817 training questions, 28,020 validation questions, and 42,031 test questions. For each question, 4 different possible answers exist, 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 that are associated with the answer of each question (attention ground truth information). This allows for training 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).

The theano library [32] and the Lasagne framework [7] were used for developing the proposed method. For optimizing the model, the Adam optimizer with the default settings [15] was employed, since it was experimentally established that it provides faster and more smooth convergence than the plain gradient descent with momentum. To further accelerate the convergence of the optimization process, the largest learning rates that provided stable convergence were used. Therefore, two different learning rates were selected: 0.001 for the explicit attention model and 0.0001 for the multiple choice answering model. Applying the same learning rate (0.0001) for both models is not expected to harm the accuracy of the model. However, it will slow down the training process, requiring more iterations to be performed. A mid-range

Graphics Processing Unit (GPU) was used for the training, limiting the maximum batch size that could fit in the memory to 16 samples. Furthermore, dropout with rate 0.2 and batch normalization were also employed in the MC Layer. Higher dropout rates could possibly slow down the training process, leading to undesired phenomena, such as underfitting the model. The explicit attention model was trained for 5 epochs and the multiple choice answer model was trained for 12 epochs using both the training and validation sets. Two pre-trained deep residual networks, the ResNet-101 and the ResNet-152 [10], were employed for extracting the feature maps from the last convolutional layer of the networks. The embedding models were initialized using pre-trained GloVe embedding vectors (Common Crawl (42B tokens), 300d) [26] and they were fine-tuned during the training process. The evaluation procedure described in [42] was utilized for measuring the performance of the proposed method, employing the toolbox supplied by the authors of [42].

Table 2: Comparing the proposed method to other baseline and state-of-the-art VQA techniques

Method	**Overall**
Human (Question + Image) [42]	0.966
Logistic Regression (Q + I) [42]	0.352
LSTM (Q + I) [21]	0.521
LSTM-Att [42]	0.556
MCB [8]	0.622
Triplet MLP [12]	**0.671**
Proposed (ResNet 152)	0.659
Proposed (ResNet 101 + ResNet 152)	**0.666**

The evaluation results are shown in Table 1, where columns 2-7 show the accuracy of the models for each separate question type. The overall accuracy is shown in the last column. Several conclusions can be drawn from the results shown in Table 1. At first, the proposed explicit attention model achieves higher overall question answering accuracy compared to the baseline implicit attention model. Regarding each separate question type, the explicit attention models increase the answering accuracy for every question type (especially for the “what” and “why” questions where providing reliable attention is crucial). Employing a second independent attention model

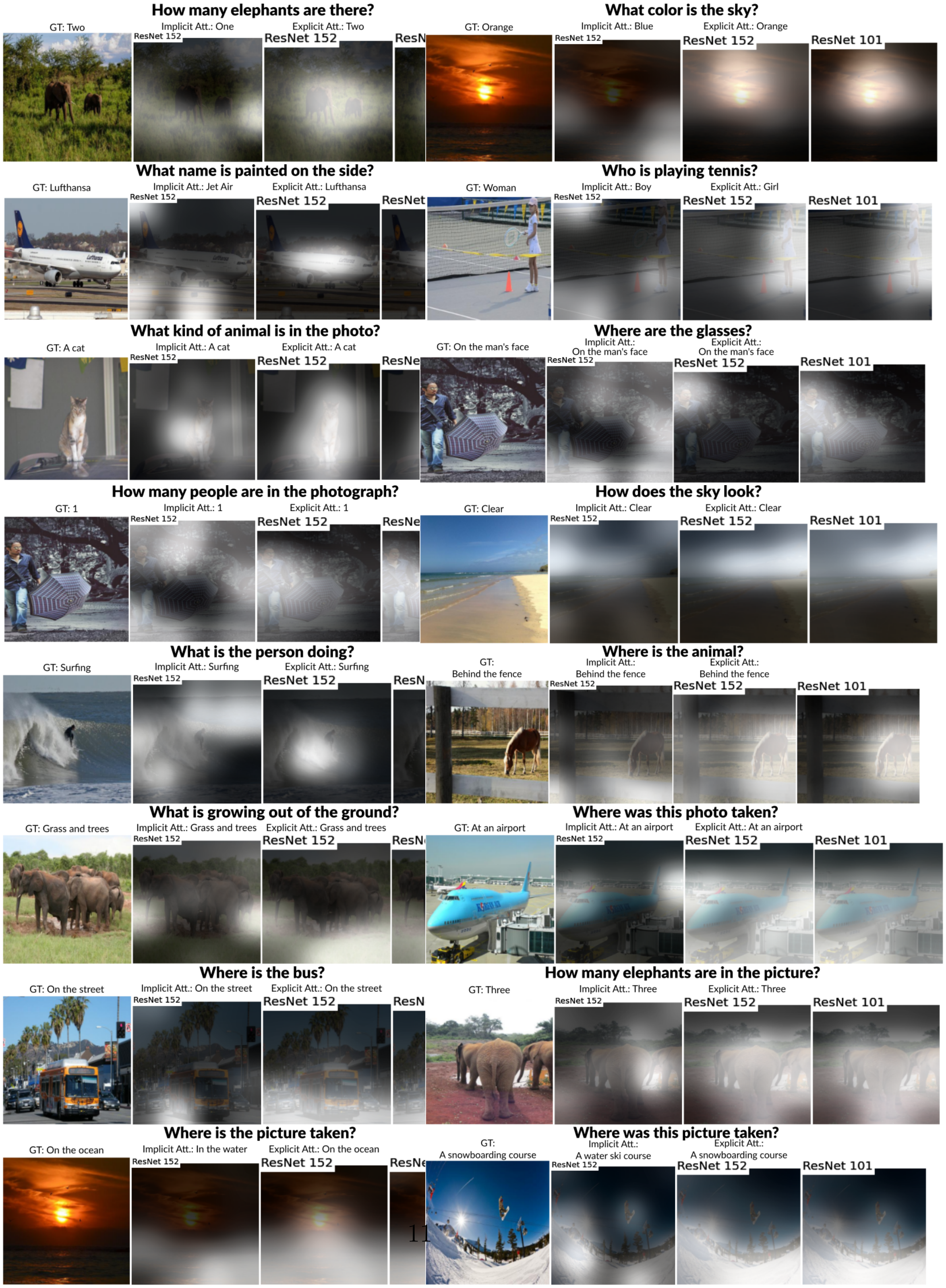

How many elephants are there?
GT: Two
Implicit Att.: One
Explicit Att.: Two
What color is the sky?
GT: Orange
Implicit Att.: Blue
Explicit Att.: Orange
What name is painted on the side?
GT: Lufthansa
Implicit Att.: Jet Air
Explicit Att.: Lufthansa
Who is playing tennis?
GT: Woman
Implicit Att.: Boy
Explicit Att.: Girl
What kind of animal is in the photo?
GT: A cat
Implicit Att.: A cat
Explicit Att.: A cat
Where are the glasses?
GT: On the man's face
Implicit Att.: On the man's face
Explicit Att.: On the man's face
How many people are in the photograph?
GT: 1
Implicit Att.: 1
Explicit Att.: 1
How does the sky look?
GT: Clear
Implicit Att.: Clear
Explicit Att.: Clear
What is the person doing?
GT: Surfing
Implicit Att.: Surfing
Explicit Att.: Surfing
Where is the animal?
GT: Behind the fence
Implicit Att.: Behind the fence
Explicit Att.: Behind the fence
What is growing out of the ground?
GT: Grass and trees
Implicit Att.: Grass and trees
Explicit Att.: Grass and trees
Where was this photo taken?
GT: At an airport
Implicit Att.: At an airport
Explicit Att.: At an airport
Where is the bus?
GT: On the street
Implicit Att.: On the street
Explicit Att.: On the street
How many elephants are in the picture?
GT: Three
Implicit Att.: Three
Explicit Att.: Three
Where is the picture taken?
GT: On the ocean
Implicit Att.: In the water
Explicit Att.: On the ocean
Where was this picture taken?
GT: A snowboarding course
Implicit Att.: A water ski course
Explicit Att.: A snowboarding course
ResNet 152
ResNet 101

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

trained with image features from a different pre-trained CNN model positively impacts the VQA accuracy. Using two different attention models increases the possibility that at least one of them will attend to the correct region. This is also demonstrated in Figure 4, where the implicit attention model and the proposed explicit attention model are compared using some of the test questions and images. It is evident that the proposed method significantly improves the attention accuracy. It is also noteworthy that when one of the two attention models fails to provide a correct attention distribution the other one can still provide fine-grained attention information that helps to correctly answer the question. A typical example of this behavior is illustrated for the question "Where are the glasses?" in Figure 4. In this specific example the implicit model also provides the correct answer because it has probably learned from the training data that the glasses are likely to appear on a man's face, despite the fact that this information was not provided by the attention model. Learning the priors from the training data without actually *understanding* the visual input is a well known behavior and also appears in many other VQA approaches [9, 1].

The proposed explicit attention also significantly improves upon the "How many"-type questions. For example, consider the question "How many elephants are there", where 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. The proposed method is also compared to other baseline and state-of-the-art VQA techniques in Table 2 achieving the second higher VQA accuracy. We also attempted to combine the proposed method with the best performing method, i.e., the Triplet MLP [12], but we were unable to reproduce the results reported in [12], since not all the details of the utilized setup were reported. Nonetheless, combining the explicit attention model with the exact setup employed in [12] is expected to further improve the accuracy.

5. Conclusions

Visual Question Answering (VQA) is among the most difficult multimodal problems as it requires a machine to be able to properly understand a question and the corresponding visual input. In this work, it was demonstrated that employing multiple explicitly trained attention models can significantly improve the VQA accuracy compared to implicit attention models as well as other state-of-the art techniques. Furthermore, a way of effectuating a mechanism that mimics the pictorial superiority effect was provided, further improving the answering accuracy.

There are several interesting future work directions. Thus, more deliberate techniques could be employed for combining multiple attention models, e.g., the AdaBoost technique [41]. Furthermore, in the proposed method the attention model was not trained when 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. Advanced pooling techniques, such as BoF pooling [25], can be also used to improve the scale invariance of the attention model and provide more reliable attention information. The proposed methodology can also be applied to other tasks that require high level visual understanding, such as image caption generation [14, 11, 37], and video caption generation [18]. Finally, the proposed approach could also be used to improve the precision of multi-modal information retrieval [5], where providing accurate visual attention information given a textual query from the user is of critical significance.

6. Acknowledgment

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] Agrawal, A., Batra, D., Parikh, D., Kembhavi, A., 2017. Dont just assume; look and answer: Overcoming priors for visual question answering, in: CVPR 2017 VQA Challenge Workshop.

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

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

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

[5] Chen, K., Bui, T., Chen, F., Wang, Z., Nevatia, R., 2017. Amc: Attention guided multi-modal correlation learning for image search. arXiv preprint arXiv:1704.00763 .

[6] Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D., 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding 163, 90 – 100. URL: http://www.sciencedirect.com/science/article/pii/S1077314217301649. language in Vision.

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

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

[9] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. arXiv preprint arXiv:1612.00837 .

[10] He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385.

[11] He, X., Shi, B., Bai, X., Xia, G.S., Zhang, Z., Dong, W., 2017. Image caption generation with part of speech guidance. Pattern Recognition Letters URL: http://www.sciencedirect.com/science/article/pii/S0167865517303811, doi:https://doi.org/10.1016/j.patrec.2017.10.018.

[12] Jabri, A., Joulin, A., van der Maaten, L., 2016. Revisiting visual question answering baselines. CoRR abs/1606.08390.

[13] Kafle, K., Kanan, C., 2017. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding 163, 3 – 20. URL: http://www.sciencedirect.com/science/

article/pii/S1077314217301170, doi:https://doi.org/10.1016/j.cviu.2017.06.005. language in Vision.

[14] Kinghorn, P., Zhang, L., Shao, L., 2017. A hierarchical and regional deep learning architecture for image description generation. Pattern Recognition Letters URL: http://www.sciencedirect.com/science/article/pii/S0167865517303240, doi:https://doi.org/10.1016/j.patrec.2017.09.013.

[15] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980.

[16] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision 123, 32–73.

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

[18] Li, W., Guo, D., Fang, X., 2017. Multimodal architecture for video captioning with memory networks and an attention mechanism. Pattern Recognition Letters URL: http://www.sciencedirect.com/science/article/pii/S016786551730380X, doi:https://doi.org/10.1016/j.patrec.2017.10.012.

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

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

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

[22] Masoudnia, S., Ebrahimpour, R., 2014. Mixture of experts: a literature survey. Artificial Intelligence Review , 1–19.

[23] Miller, P., 2011. The processing of pictures and written words: A perceptual and conceptual perspective. Psychology , 713–720doi:10.4236/psych.2011.27109.

[24] Nelson Douglas L., Reed Valerie S., Walling John R., 1976. Pictorial superiority effect. Journal of Experimental Psychology: Human Learning and Memory 2, 523–528.

[25] Passalis, N., Tefas, A., 2017. Learning bag-of-features pooling for deep convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5755–5763.

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

[27] Qiao, T., Dong, J., Xu, D., 2017. Exploring human-like attention supervision in visual question answering. CoRR abs/1709.06308.

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

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

[30] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

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

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

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

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

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

[36] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding 163, 21 – 40. URL: `http://www.sciencedirect.com/science/article/pii/S1077314217300772`, doi:`https://doi.org/10.1016/j.cviu.2017.05.001`. language in Vision.

[37] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning, pp. 2048–2057.

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

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

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

[41] Zhu, J., Zou, H., Rosset, S., Hastie, T., 2009. Multi-class adaboost. Statistics and its Interface 2, 349–360.

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