Improving Visual Question Answering using Active Perception on Static Images

Theodoros Bozinis, Nikolaos Passalis and Anastasios Tefas
Department of Informatics, Faculty of Sciences
Aristotle University of Thessaloniki
Thessaloniki 541 24, Greece
Email: {mpozinit, passalis, tefas}@csd.auth.gr

***Abstract*—Visual Question Answering (VQA) is one of the most challenging emerging applications of deep learning. Providing powerful attention mechanisms is crucial for VQA, since the model must correctly identify the region of an image that is relevant to the question at hand. However, existing models analyze the input images at a fixed and typically small resolution, often leading to discarding valuable fine-grained details. To overcome this limitation, in this work we propose a reinforcement learning-based active perception approach that works by applying a series of transformation operations on the images (translation, zoom) in order to facilitate answering the question at hand. This allows for performing fine-grained analysis, effectively increasing the resolution at which the models process information. The proposed method is orthogonal to existing attention mechanisms and it can be combined with most existing VQA methods. The effectiveness of the proposed method is experimentally demonstrated on a challenging VQA dataset.**

I. Introduction

The progress witnessed in deep learning (DL) led to a number of impressive applications, which often involve multi-modal data [1]. Among the most spectacular ones is Visual Question Answering (VQA) [2], [3], [4], where a deep learning model must answer a textual question that refers to a given image. VQA is among the most challenging deep learning applications, since the model must correctly identify the region of the image that concerns the given question and then process this information to provide the correct answer. This led to development of many *attention* mechanisms which allow for processing only the information that is relevant to the question at hand [5], [6], [7], [8].

Despite the success of the aforementioned attention mechanisms, they suffer from a critical limitation: the input visual data are analyzed at a fixed resolution despite the higher resolution of the original images. As a result, most models are restricted at analyzing input images that are smaller than 500×500 pixels, while many of them are still limited to less than half of this. This process, despite allowing for reducing the computational complexity of the models by processing lower resolution inputs, comes with several significant drawbacks. First, it restricts the fidelity of the input, leading to loosing several fine grained details, especially for smaller or thin objects that might end up covering only a few pixels after resizing the input images. Furthermore, DL models are sensitive to the scale of the objects appearing in images. Therefore, if the same object, but in a different size, appear in a novel image, the employed model might fail to recognize it.

The main contribution of this work is proposing a deep reinforcement learning (RL)-based active perception approach that can overcome the aforementioned limitations. More specifically, we propose keeping the resolution at which the analysis is performed fixed, which does not increase the computational cost of the analysis, but employing a methodology for first appropriately *transforming* the input in order to maximize the accuracy of VQA. To this end, we employ a *virtual camera* that can shot at different regions of the original input, allowing for a) performing fine-grained information analysis at the same cost (for each frame processed by the VQA model), b) keeping only the information that is indeed relevant to the provided question, and c) mitigating the effect of objects that appear at different scales. The way the proposed method works is illustrated in Fig. 1. The virtual camera zooms to the region (leaves of the trees) that concerns the given question. This allows for extracting more fine-grained information, maximizing the confidence on the correct answer and increasing VQA accuracy.

To the best of our knowledge, this is the first method that employs an active perception approach for increasing the accuracy of VQA on static datasets. Note that the proposed method is orthogonal to existing attention mechanisms, since the proposed method appropriately transforms the input (by zooming in/out and translating the content) to accommodate the task at hand and it is capable of directly adapting to the scale of the objects that appear on images. On the other hand, attention mechanisms only allow the model to suppress irrelevant features, without re-analysing the input data, as the proposed method does. The ability of the proposed method to be effectively combined with attention mechanisms, increasing the VQA accuracy, is demonstrated in this paper through several experiments using a powerful attention-equipped VQA model, MUTAN [6]. Also, in contrast with existing RL-based methods proposed for similar tasks, such as caption generation [9], [10], [11], the proposed method provides a powerful active perception approach that can be combined with any existing VQA model, instead of addressing VQA *per se*. Finally, compared to interactive VQA approaches, such as [12], in this work we do not employ 3D simulation environments in which complex iterations with objects are

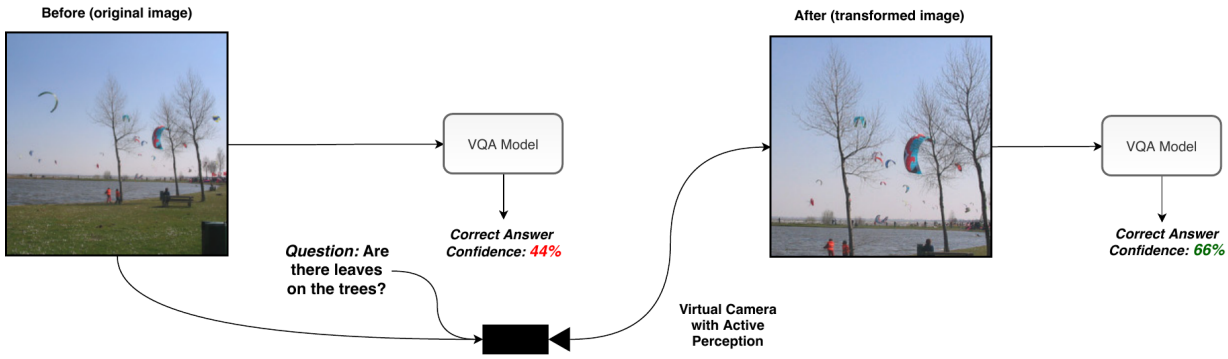


Fig. 1: Using active perception to increase the accuracy of VQA by controlling a *virtual camera* to shot the region that is relevant to the given question, increasing the granularity of information analysis

possible, but we focus on handling static images.

The rest of the paper is structured as follows. First, the proposed method is analytically derived in Section II. Then, the experimental evaluation is provided in Section III. Finally conclusions are drawn in Section IV.

II. Proposed Method

First, we provide a brief introduction to used notation, as well to the VQA problem at hand. Next, we describe how active perception, which is typically used in the context of robotic perception [13], can be employed to increase the accuracy of VQA, when used on static, yet high-resolution, images, and formally define the optimization problem. Finally, we derive the proposed method by providing a relaxation to the optimization problem at hand, which allows us to effectively employ deep reinforcement learning.

A. Visual Question Answering and Active Perception on Static Images

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ denote an input image, where H denotes the height, W the width and C the number of channels of the image. Also, let q denote the question that we want to answer. Typically, word embedding models are used to acquire the encoded vector representation of the question [14], [15]. The notation $\mathbf{q} \in \mathbb{R}^{N_w}$ is then used to refer to the encoded question, where N_w is the dimensionality of embedding space. In this work, we constrain our setup to the commonly used multiple choice VQA, where the correct answer is selected over a collection of N_C possible answers. Note that this is without loss of generality, since the proposed method can be also used for any other VQA setup [16]. Also, let $\mathcal{X} = \{(\mathbf{x}_i, q_i, \mathbf{t}_i) | \ \forall i = 1, \ldots, N\}$ be a training set that consists of N training triplets of images, questions and correct answers, while let $f_{\mathbf{W}}(\cdot) \in \mathbb{R}^{N_C}$ denote a VQA model, where $\mathbf{W}$ are the trainable parameters of the model. The VQA model $f_{\mathbf{W}}(\cdot)$ is trained by minimizing an appropriately chosen loss over the training set:

$$\mathbf{W} = \arg\min_{\mathbf{W}'} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_{\mathbf{W}'}(\mathbf{x}_i, \mathbf{q}_i), \mathbf{t}_i\right), \tag{1}$$

where $\mathcal{L}(\cdot)$ is typically set to the cross-entropy loss, when a closed set of N_C answers is used [6]. In this work, we also use VQA models that return a probability distribution over the available answers, e.g., by employing the softmax activation function. Also, following the relevant literature [16], [5], a deep learning model is used to implement $f_{\mathbf{W}}(\cdot)$.

As we also discussed in Section I, using attention allows the model to extract only the information relevant to the given question [6], [5]. However, a critical limitation of this approach is that the analysis is performed at a fixed, typically low, resolution, even though the original images are usually of significantly higher resolution, loosing finer details that might be useful for answering the question at hand. In this work, we propose pre-processing the original images by applying a series of transformation operators $\mathcal{A}$ in order select a more appropriate view. This view should, provided the limitations of the employed VQA model, e.g., input resolution, maximize the accuracy of VQA. The transformation operations are the result of a *virtual* camera that is capable of "shooting" at different regions of the original image, as well as altering its field of view, similar to other applications related to control [17]. The camera has fixed resolution, which is set to be equal to the resolution used by the model $f_{\mathbf{W}}(\cdot)$ for the analysis. The following set of transformation operations are supported:

1) a_{left}, which corresponds to horizontal translation of the field of view window to the left by δ_T pixels,
2) a_{right}, which corresponds to horizontal translation of the field of view window to the right by δ_T pixels,
3) a_{up}, which corresponds to the vertical translation of the field of view window up by δ_T pixels,
4) a_{down}, which corresponds to the vertical translation of the field of view window down by δ_T pixels,
5) $a_{zoom-in}$, which corresponds to zooming-in (decreasing the field of view by $\delta_z\%$),
6) $a_{zoom-out}$, which corresponds to zooming-out (increasing the field of view by $\delta_z\%$),
7) a_{null}, which corresponds to performing no transformation (this transformation should be selected when the view is already optimal for answering the given question),

where δ_T is set to $\lfloor 0.1 \cdot \min(W, H) \rfloor$ and δ_z to 4% for all the

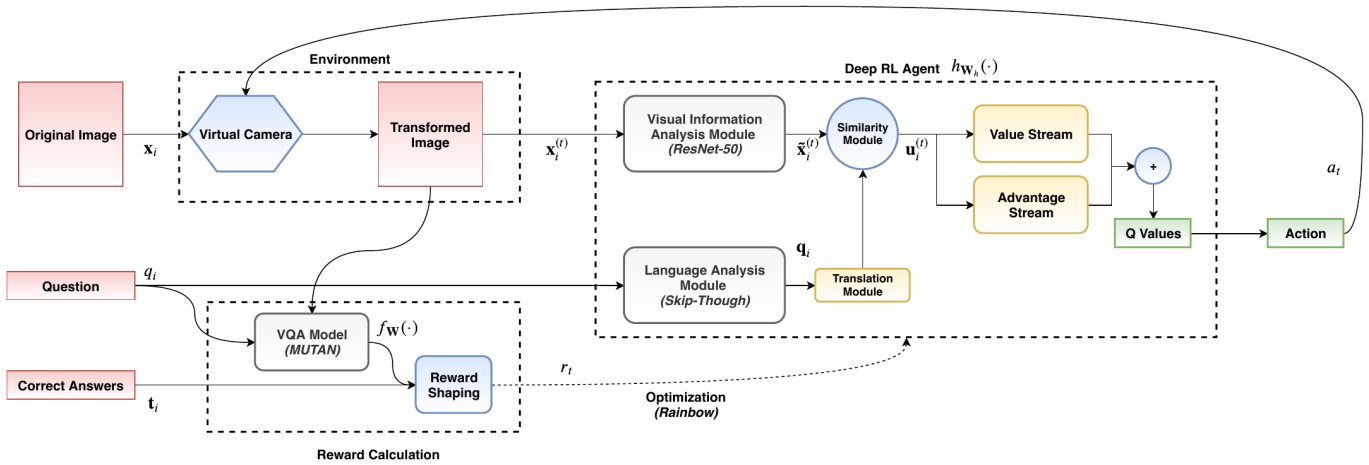


Fig. 2: Overview of the proposed method: The reinforcement learning agent can control a virtual camera in order to apply a series of transformations (translation and zoom) on the original input image in order to acquire a view that is more relevant to the given question. Then, the transformed image is fed to the VQA model and the agent is trained to maximize the expected reward, which is a function of the confidence of the VQA model for the correct answer.

experiments conducted in this paper. The cumulative effect of these operations was demonstrated in Fig. 1, while more examples are provided in Section III.

We aim to learn an appropriate model $h_{\mathbf{W}_h}(\mathbf{x}_i^{(t)}, \mathbf{q}_i) \in \mathcal{A}$ which, given the camera view $\mathbf{x}_i^{(t)}$ at time t and the question $\mathbf{q}_i$ selects the most appropriate transformation from $\mathcal{A}$ in order to minimize the loss provided in (1). Therefore, the parameters of the model can be learned by solving the following optimization problem:

$$\mathbf{W}_h = \arg\min_{\mathbf{W'}_h} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_{\mathbf{W}}(\mathbf{x}_i^{(N_T)}, \mathbf{q}_i), \mathbf{t}_i\right), \quad (2)$$

subject to

$$\mathbf{x}_i^{(t)} = a_t(\mathbf{x}_i^{(t-1)}),$$
$$a_t = h_{\mathbf{W}_{h'}}(\mathbf{x}_i^{(t-1)}, \mathbf{q}_i),$$
$$\text{and } \mathbf{x}_i^{(0)} = \mathbf{x}_i,$$

where N_T is the total number of transformation operations applied. Note that $f_{\mathbf{W}}(\cdot)$ can be either be a pre-trained model, or jointly optimized during this process. Note that the final image $\mathbf{x}_i^{(N_T)}$ is the result of N_T transformations, as dictated by the constraints of (2). In this work, we choose to keep the $f_{\mathbf{W}}(\cdot)$ fixed during the optimization of $h_{\mathbf{W}_h}(\cdot)$, since jointly optimizing both models is significantly more computationally intensive and increases the complexity of the proposed method.

B. Problem Relaxation and Deep Reinforcement Learning

Directly solving the problem presented in (2) is intractable. Therefore, instead of directly learning the parameters $\mathbf{W}_h$ to minimize (2), we employ a reinforcement learning approach to maximize the reward collected by an agent that controls the camera using the operations available in $\mathcal{A}$. The employed reward function must express the optimization objective of (2), i.e., to increase the probability of the VQA model answering correctly. Based on this observation, we defined the employed reward function as:

$$r_t = [f_{\mathbf{W}}(\mathbf{x}_i^{(t)}, \mathbf{q}_i)]_c - [f_{\mathbf{W}}(\mathbf{x}_i^{(t-1)}, \mathbf{q}_i)]_c, \quad (3)$$

where the notation $[f_{\mathbf{W}}(\cdot)]_c$ is used to refer to the confidence of the correct answer. Therefore, the agent acquires a positive reward when the VQA model $f_{\mathbf{W}}(\cdot)$ becomes more confident on the correct answers after a control step. Otherwise, a negative reward is obtained. Note that for datasets with multiple correct answers, such as [4], the answer that provides that maximum increase in the confidence is used for calculating the reward.

The proposed method is summarized in Fig. 2. The model $h_{\mathbf{W}_h}(\cdot)$ can be optimized using any deep reinforcement learning method in order to maximize the reward obtained through each control episode. In this work, we employed a state-of-the-art Q-learning approach, the Rainbow [18], which provides an efficient way to fit an estimator for the expected future discounted reward for each action at every time-step. This, in turn, provides a straightforward way to implement $h_{\mathbf{W}_h}(\cdot)$ by simply selecting the action that is expected to yield the maximum future discounted reward. The network architecture used to estimate the Q-values is also shown in Fig. 2. After analyzing the input image using a visual information analysis network, we extract an attention-like feature map by calculating the similarity $\mathbf{u} \in \mathbb{R}^{H_a \times W_a}$ between visual and the textual modality as:

$$[\mathbf{u}]_{i,j} = [\tilde{\mathbf{x}}]_{i,j}^{T}(\mathbf{W}_T\mathbf{q}) \in \mathbb{R}` \quad (4)$$

where the notations $[\mathbf{u}]_{i,j}$ is used to refer to the element in the location (i, j) of attention map $\mathbf{u}$, $\mathbf{W}_T$ is a linear *translation* layer that maps the textual space into the visual space, $\tilde{\mathbf{x}}$ denotes the feature map extracted from the visual information analysis module, while $H_a \times W_a$ is the size of the feature map $\tilde{\mathbf{x}}$. Two fully connected layers, with a hidden layer consisting

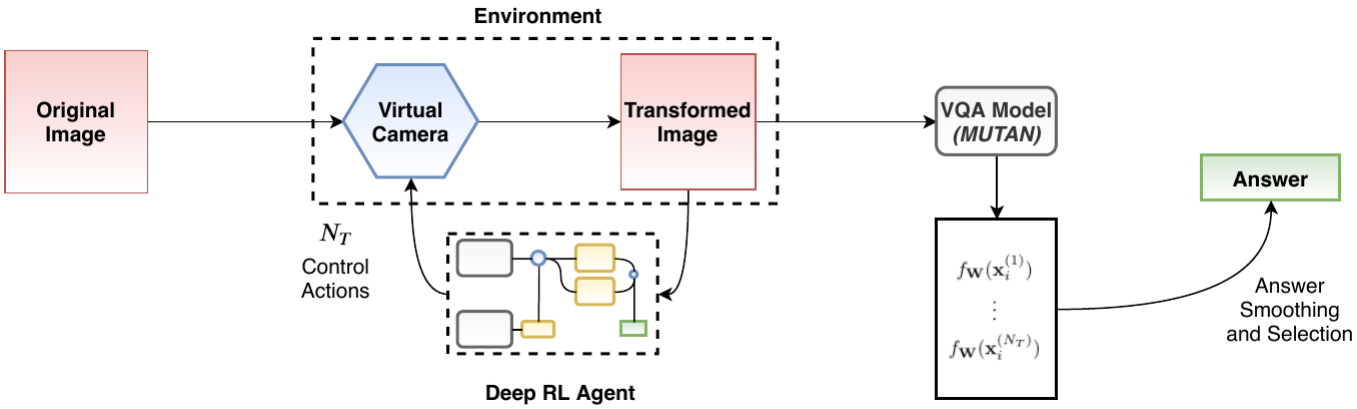


Fig. 3: Applying the proposed method on an unseen image-question pair (inference): The agent is allowed to apply N_T transformations on the input image. Then, the answers obtained for each intermediate transformation are aggregated and the final answer is inferred according to (5).

of 512 neurons, that use the ReLU activation function, were used to implement both the value and advantage streams, as suggested in [18].

Note that the agent is rewarded for any control action that will lead to increasing the confidence on the correct answer, regardless the time-step at which the action is performed. A side effect of this is that the final action of the agent can be sub-optimal (since it only partly contributes to the total reward). To overcome this limitation, we propose not to choose the answer with the highest probability in the last step, but the action with the highest average probability over the course of each episode. In this way, the agent is less sensitive to noise, while takes into account the behavior of the VQA model during the whole episode, lessening the effect of potentially wrong control actions, as further demonstrated in Section III. Therefore, for an episode with N_T control steps, the final answer is calculated as:

$$\arg max_j \left(\left[\frac{1}{N_T} \sum_{i=1}^{N_T} f_{\mathbf{W}}(\mathbf{x}_i^{(t)}, \mathbf{q}_i) \right]_j \right). \quad (5)$$

The inference process is further illustrates in Fig. 3. Note that the reward function is used only during the training process, while the VQA model is employed at each time step for calculating the confidence for each possible answer.

III. Experimental Evaluation

In this Section, we briefly provide the experimental setup used in this work, and then present and discuss the experimental results. The MUTAN model [6] was used as the base VQA model ($f_{\mathbf{W}}(\cdot)$), while a ResNet-50 model [19], pretrained on the ImageNet dataset [20], was used to perform the visual information analysis ($h_{\mathbf{W}_h}(\cdot)$). For extracting vector representations of questions we used the GRU-based encoder employed by MUTAN. The Rainbow method was used to optimize the RL model, while the discount factor was set to 0.99 and the size of the replay memory was set to 100,000. The deep RL model was trained for 300,000 steps. The Adam optimizer was used [21], while the learning rate was set to 0.5×10^{-4}. For evaluating both the baseline method, as well as the proposed one we ran 5,000 episodes on the validation set of the employed dataset. Note that we used a different image for each episode, while the same images/questions were used when evaluating different methods. Finally, the number of control steps for each episode was set to $N_T = 5$ for all the conducted experiments. All the evaluation results are reported on the VQA 2.0 dataset [2], [4], which employs the images provided by the MS COCO dataset [22]. The VQA model was also pretrained on the used dataset.

TABLE I: Evaluation results on the VQA 2.0 dataset

Method	Accuracy	Acc. Gain
Baseline	60.36	-
Proposed (Confident Frame)	59.81	-0.55
Proposed	**60.86**	**0.5**
Proposed (Best Frame)	66.68	6.32

The evaluation results are reported in Table I. We compared the proposed approach to the baseline VQA model, where no transformation has been applied on the original images, as well as to the proposed RL-based active perception approach that employs three different ways to select the correct answer:

1) "Proposed (Confident Frame)", where the answer with the higher confidence during the course of the episode is selected, i.e., the frame t is selected as:

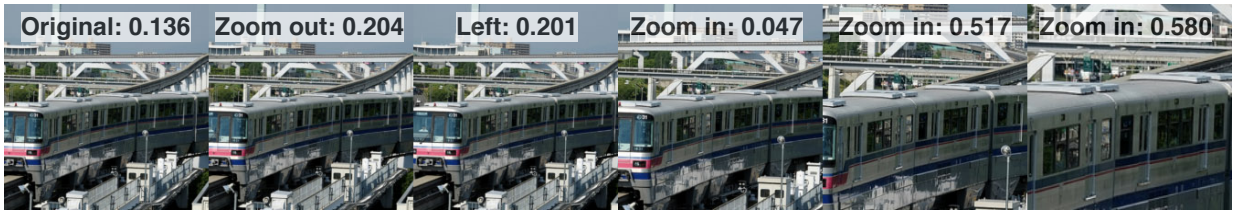
$$\arg\max_{t'} \left(\max \left(f_{\mathbf{W}}(\mathbf{x}_i^{(t')}, \mathbf{q}_i) \right) \right), \quad (6)$$

2) "Proposed", where the proposed answer selection method presented in (5) is used, and
3) "Proposed (Best Frame)", where the frame that leads to the higher confidence on the correct answer is selected to provide the answer, i.e., the frame t is selected as:

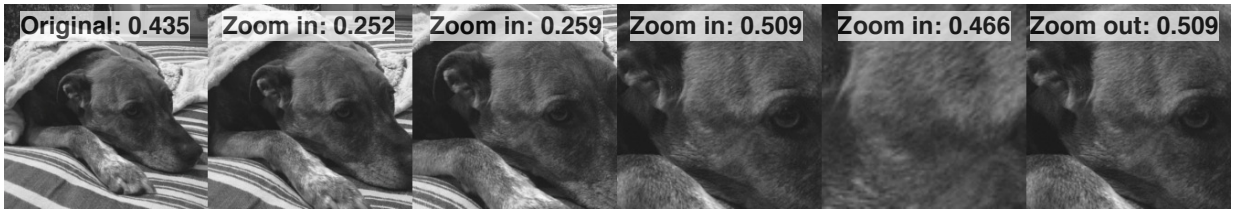
$$\arg\max_{t'} \left([f_{\mathbf{W}}(\mathbf{x}_i^{(t')}, \mathbf{q}_i)]_c \right), \quad (7)$$

a) *Question*: Are there leaves on the trees? - *Answer*: No

b) *Question*: Is this a toaster oven? - *Answer*: Yes



c) *Question*: What color is the train? - *Answer*: Silver, Gray



d) *Question*: Is the dog sleeping? - *Answer*: No

e) *Question*: Are there vegetables on the counter? - *Answer*: No

Fig. 4: Qualitative Evaluation: Three episodes during the testing. The select transformation and confidence to the correct answer is denoted in the upper part of each image.

where c is the index of the correct answer.

The proposed method increases the VQA accuracy by 0.5%, while when the proposed frame selection approach is replaced by the most confident frame ("Proposed + Confident Frame") a decrease of over -0.5% is observed. To understand why this happens, we need to consider the case in which the deep RL agent loses the object of interest and the question refers to the existence of the aforementioned object, e.g., the question is "is there a dog?", the original image contains a dog, while the transformed image no longer contains a dog. In this case, the VQA model will be very confident on the selected answer ("no"), since the object of interest no longer exists in the transformed frame, despite the correct answer is "yes", since the object exists in the original frame. Employing the proposed averaging-based approach allows to withstand such phenomena, increasing the accuracy of VQA, and demonstrating the effectiveness of the proposed method. Finally, note that we also examined whether developing more robust frame selection approaches would allow for further increasing the accuracy of the proposed method by directly selecting the frame that maximizes the confidence to the correct answer ("Proposed + Best frame"). Indeed, this approach led to an enormous 6.32% increase in the accuracy. This demonstrates that the proposed RL agent was indeed able to reveal fine-grained information,

which was not available to the VQA model when the original image was used, highlighting the potential of the proposed approach.

Finally, in Fig. 4, we provide several test episodes to qualitatively examine the behavior of the proposed method. Several interesting conclusions can be drawn. First, note that in the first case (a) the model correctly zooms in toward the leaf area of the trees, demonstrating that the agent is indeed capable of identifying the region that is relevant to the question at hand. This gives an important advantage to the model, since it can perform more fine-grained analysis, i.e., revealing fine details that would be lost if the whole image was processed instead. Indeed, the confidence to the correct answer increases from 0.44 in the original image to 0.66 in the final one, while providing an average confidence of about 0.56, which is almost 30% higher than the original one. These control sequences also illustrate the intrinsic instability of many VQA models. For example, note that in the second case, the confidence to the correct answer slightly decreases, without any obvious change in the input image. This highlights a critical limitation of VQA models, since the accuracy of the employed RL agent relies on the feedback provided by the VQA model. Another interesting phenomenon is shown in the fourth case (d), where the agent correctly chooses the "zoom out" transformation, when the visual clues that would lead to the correct answer are absent (note that correctly answering this question requires detecting whether the eye of the dog is open or closed and this information is not available in the penultimate frame).

IV. Conclusions

In this paper, we presented an active perception approach, that employs deep reinforcement learning, and can be directly used on static images to increase the accuracy of VQA. The proposed method works by applying a series of transformations on the input images in order facilitate answering the question at hand. As experimentally demonstrated, the proposed method is capable of increasing the accuracy of VQA, by performing fine-grained information analysis and mitigating various issues, such as performing inference about objects that appear at different scales. The conducted experiments also demonstrated the high potential of the proposed method, since developing more accurate frame selection and aggregation approaches can lead to significant further accuracy improvements.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *Proceedings of the International Conference on Computer Vision*, 2015.

[3] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and Yang: Balancing and answering binary visual questions," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2016.

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017.

[5] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 451–466.

[6] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 2612–2620.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proceedings of the Advances in Neural Information Processing Systems*, 2016, pp. 289–297.

[8] V. Lioutas, N. Passalis, and A. Tefas, "Explicit ensemble attention learning for improving visual question answering," *Pattern Recognition Letters*, vol. 111, pp. 51–57, 2018.

[9] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017, pp. 290–298.

[10] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Yang Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4213–4222.

[11] E. Daskalakis, M. Tzelepi, and A. Tefas, "Learning deep spatiotemporal features for video captioning," *Pattern Recognition Letters*, vol. 116, pp. 143–149, 2018.

[12] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "IQA: Visual question answering in interactive environments," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4089–4098.

[13] Y. Aloimonos, *Active perception*. Psychology Press, 2013.

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

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[17] N. Passalis and A. Tefas, "Deep reinforcement learning for controlling frontal person close-up shooting," *Neurocomputing*, vol. 335, pp. 37–47, 2019.

[18] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

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

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recongition*, 2009.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.