IMPROVING VISUAL QUESTION ANSWERING USING ACTIVE PERCEPTION ON STATIC IMAGES

Theodoros Bozinis, Nikolaos Passalis and Anastasios Tefas

Aristotle University of Thessaloniki (Greece)

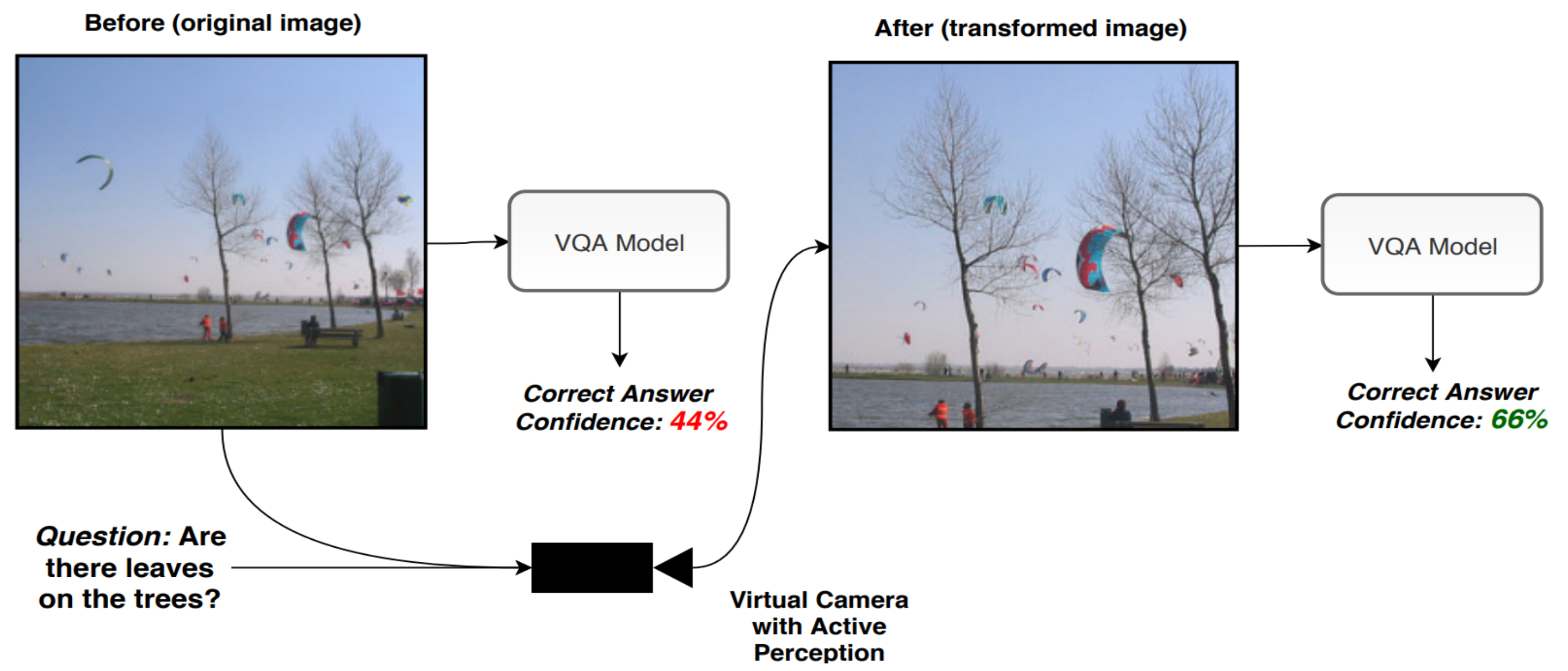
email: mpozinit@csd.auth.gr, passalis@csd.auth.gr, tefas@csd.auth.gr

Introduction

- **Visual Question Answering (VQA)** [1] is one of the most challenging problems of deep learning and has received a lot of attention in recent years
- **Powerful attention** mechanisms are key part of VQA to identify the region of an image that is relevant to the question
- Unlike previous attempts that analyze the input image at a **fixed** (and typically low) **resolution** we propose an active perception mechanism to overcome that limitation
- We employ a virtual camera that can shot at regions of the original input, operated by a trained with reinforcement learning agent
- The proposed method can be combined with most existing VQA methods

Motivation

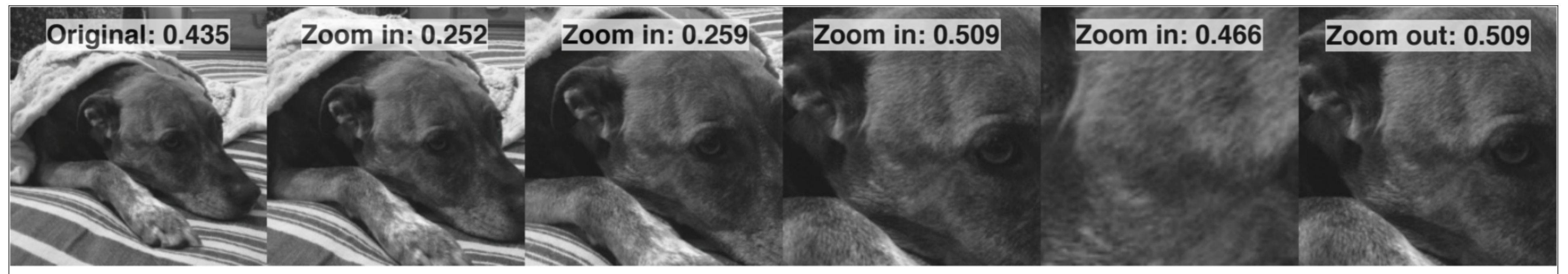
- We want to emulate the human process of finding the area of the image that contains the answer to the question by focusing only on the region of interest with the fixed resolution virtual camera
- Cropping the image allows us to:
  - perform fine-grained information analysis at the same low computational cost
  - keep only the relevant information to the question
  - reduce the negative scale sensitivity of the objects on the image that can tamper with the correlation of which with the question


Before (original image)
After (transformed image)
VQA Model
Correct Answer Confidence: 44%
VQA Model
Correct Answer Confidence: 66%
Question: Are there leaves on the trees?
Virtual Camera with Active Perception

Proposed method

Let $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ denote an input image where H the height, W the width & C the num. of the image channels and $\boldsymbol{q} \in \mathbb{R}^{N_w}$ be the encoded question. Also, let $f_{\boldsymbol{W}}(\cdot) \in \mathbb{R}^{N_c}$ be the VQA model, where $\mathbf{W}$ the pre-trained parameters of the model. We can now define the model $h_{\boldsymbol{w}_h}(\boldsymbol{x}_i{}^{(t)}, \boldsymbol{q}_i) \in \mathcal{A}$ we aim to learn.

- The **agent action space** $\mathcal{A}$ is defined by a set of **7** actions:
  - a_{left}, a_{right}, a_{up}, a_{down} **translation** transformations of the camera by δ_{T} pixels
  - $a_{zoom-in}$, $a_{zoom-out}$ **zooming** transformations of the camera by δ_z%
  - a_{null} no transformation
- Our **goal is to learn** $\boldsymbol{W}_h = argmin_{\boldsymbol{w'}_h} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_{\boldsymbol{W}}(\boldsymbol{x}_i{}^{(N_T)}, \boldsymbol{q}_i), \boldsymbol{t}_i)$ the trainable parameters of the agent model
- Directly solving the above statement is **intractable**, therefore we use a **reinforcement learning** approach to maximize the collected reward of an **agent that controls the camera** using the operations of $\mathcal{A}$
- The **reward function** is $r_t = \left[f_{\boldsymbol{W}}\left(\boldsymbol{x}_i^{(t)}, \boldsymbol{q}_i\right)\right]_c - \left[f_{\boldsymbol{W}}\left(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{q}_i\right)\right]_c$ with $[f_{\boldsymbol{W}}(\cdot)]_c$ as the **confidence of the correct answer**
- The Q-values are estimated using an attention-like mechanism to calculate the similarity $\boldsymbol{u} \in \mathbb{R}^{H_a \times W_a}$ between visual and textual modality as $[\boldsymbol{u}]_{i,j} = [\widetilde{\boldsymbol{x}}]_{i,j}^{T}(\boldsymbol{W}_T\boldsymbol{q}) \in \mathbb{R}'$



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

To overcome the possible sub-optimal agent's final action or the most confident answers, the chosen answer can be selected by the highest average probability over the course of each episode

Experimental evaluation

- A pretrained MUTAN [2] VQA model was used along side a pretrained GRU-based encoder and a pretrained ResNet-50 for the question and image representation for the agent
- The Rainbow [3] method was used to train the RL agent for 300,000 steps with $\gamma = 0.99$ and a replay memory of 100,000 steps, using the Adam optimizer with 0.5×10^{-4} learning rate
- The test was 5,000 episodes from the validation set of 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

References

[1] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2017.

[2] Hedi Ben-Younes, Remi Cadene, Matthieu Cord, and Nicolas Thome, "Mutan: Multimodal tucker fusion for visual question answering," in Proceedings of the International Conference on Computer Vision, 2017, pp. 2612–2620.

[3] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver, "Rainbow: Combining improvements in deep reinforcement learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR).