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

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

Introduction

- Visual Question Answering (VQA) very challenging deep learning application
- Visual attention key part of VQA
  - Correctly identify the region of an image
  - Relevant to the question
- **Existing methods**
  - Image analysis on fixed & low-resolution images
  - Losing fine-grained details
  - Sensitive on object scale

Our proposal

- Reinforcement learning-based active perception approach
- Transformation operations on the images (zoom & translation)
- Allows us to
  - perform fine-grained visual analysis
  - effectively increasing the resolution at which the models process information
- Orthogonal to existing attention mechanisms
- Can be combined with existing VQA methods

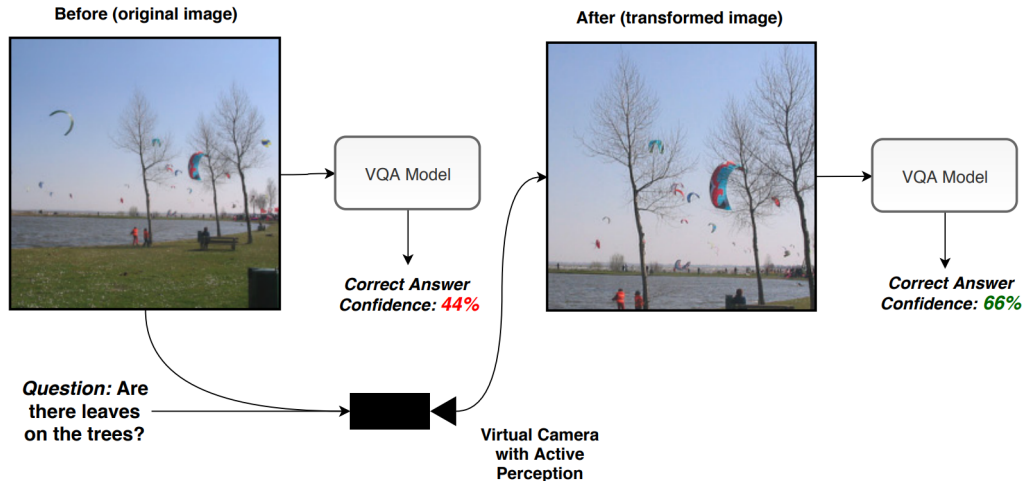
Virtual camera

- **Fixed resolution virtual camera**
  - Keeping the resolution of the image analysis fixed at that
  - Not increasing the computational cost

a) **Fine-grained** analysis

b) Keeping only the information **relevant to the question**

c) Mitigating the effect of object **scale**



The setup (1/3)

- input image: $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$
- encoded question: $\boldsymbol{q} \in \mathbb{R}^{N_w}$
- training set
  $$\mathcal{X} = \left\{\left(\boldsymbol{x}_i, \boldsymbol{q}_{i,}, \boldsymbol{t}_i\right) | \forall i = 1, \ldots, N\right\}$$
- VQA model: $f_{\boldsymbol{W}}(\cdot) \in \mathbb{R}^{N_c}$
- $\boldsymbol{W} = argmin_i \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_{W'}(\boldsymbol{x}_i, \boldsymbol{q}_i), \boldsymbol{t}_i)$
  where $\mathcal{L}(\cdot)$ cross-entropy loss
  closed set of N_c possible answers

- **action space** $\mathcal{A}$ 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 $\delta_z\%$
  - a_{null} no transformation
- we aim to learn $h_{\boldsymbol{w}_h}\left(\boldsymbol{x}_i{}^{(t)}, \boldsymbol{q}_i\right) \in \mathcal{A}$

$$\boldsymbol{W}_h = argmin_{\boldsymbol{w'}_h} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_{\boldsymbol{W}}\left(\boldsymbol{x}_i{}^{(N_T)}, \boldsymbol{q}_i\right), \boldsymbol{t}_i\right)$$

The setup (2/3)

- Directly learning $\boldsymbol{W}_h$ intractable
- Using Reinforcement learning to maximize the reward of an agent that controls the virtual camera
- Optimization objective: increase the probability of the VQA model answering correctly

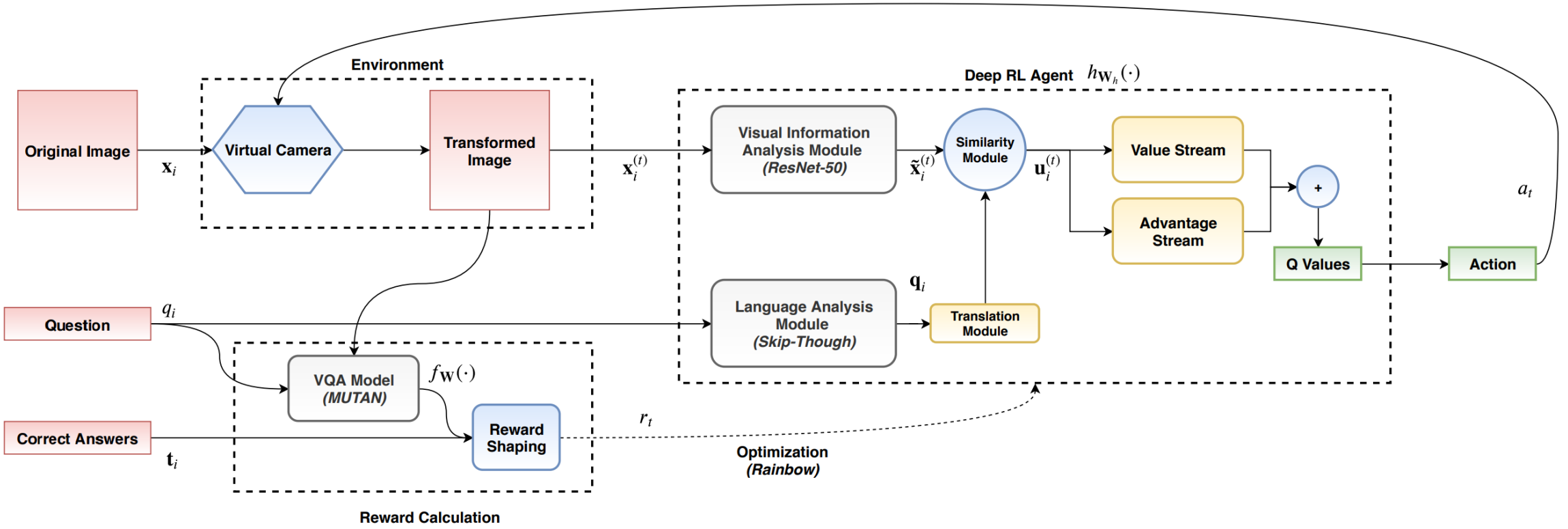
$$r_t = \left[f_{\boldsymbol{W}}\left(\boldsymbol{x}_i^{(t)}, \boldsymbol{q}_i\right)\right]_c - \left[f_{\boldsymbol{W}}\left(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{q}_i\right)\right]_c$$

$[f_{\boldsymbol{W}}(\cdot)]_c$ confidence of the correct answer

- Agent network: image analysis, question encoding, similarity
- Similarity from image and question extract an attention-like feature map

$$\boldsymbol{u} \in \mathbb{R}^{H_a \times W_a}$$
$$[\boldsymbol{u}]_{i,j} = [\tilde{\boldsymbol{x}}]_{i,j}^{T}(\boldsymbol{W}_T \boldsymbol{q}) \in \mathbb{R}'$$

The setup (3/3)


Environment
Deep RL Agent $h_{\mathbf{W}_h}(\cdot)$
Original Image
$\mathbf{x}_i$
Virtual Camera
Transformed Image
$\mathbf{x}_i^{(t)}$
Visual Information Analysis Module (ResNet-50)
$\tilde{\mathbf{x}}_i^{(t)}$
Similarity Module
$\mathbf{u}_i^{(t)}$
Value Stream
Advantage Stream
+
Q Values
Action
a_t
Question
q_i
Language Analysis Module (Skip-Though)
$\mathbf{q}_i$
Translation Module
VQA Model (MUTAN)
$f_{\mathbf{W}}(\cdot)$
Correct Answers
$\mathbf{t}_i$
Reward Shaping
r_t
Optimization (Rainbow)
Reward Calculation

The frame selection problem

- Agent rewarded for every confidence increasing action to the correct answer
- Side effect: sub-optimal final action
- To overcome this: select the answer with highest average confidence
- For N_T control steps

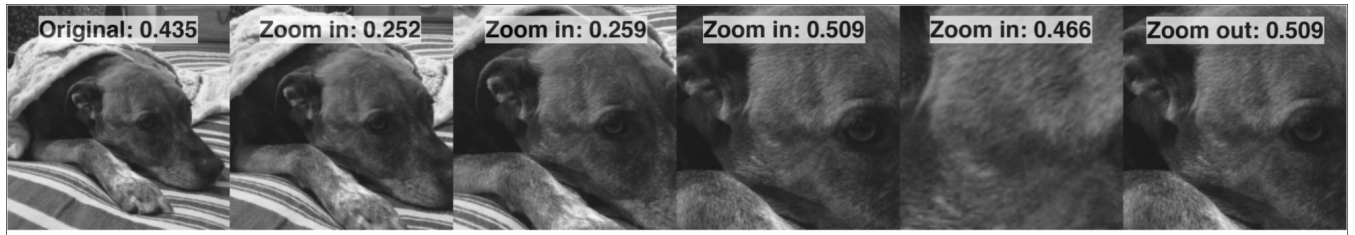
$$\arg max_j \left(\left[\frac{1}{N_T} \sum_{i=1}^{N_T} f_{\boldsymbol{W}} \left(\boldsymbol{x}_i^{(t)}, \boldsymbol{q}_i \right) \right]_j \right)$$

Experimental setup

- VQA Model: MUTAN
- Agent:
  - Image analysis: ResNet-50 (pre-trained)
  - Question representation: GRU-based encoder (pre-trained)
  - Q-Learning training:
    - Rainbow method
    - 300,000 steps
    - replay memory of 100,000 steps
    - $\gamma = 0.99$
    - Adam optimizer, learning rate: 0.5×10^{-4}
- 5,000 episodes test from the VQA 2.0 validation set

Experimental results & Conclusions

- Active perception approach to VQA
- Trained using reinforcement learning
- transformations on the input images
- Method capable of increasing the accuracy of VQA
- High potential: frame selection can lead to significant further accuracy improvements



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

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

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

Acknowledgments

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

www.opendr.eu

Thank you!