Salient region detection in the task of visual question answering

Margarita Favorskaya*, Vladimir Andreev and Aleksei Popov
Reshetnev Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russian Federation
* E-mail: favorskaya@sibsau.ru

Abstract. Salient region detection in Visual Question Answering (VQA) is an attempt to simulate a human ability to quickly perceive a scene by selectively looking on image fragments instead of processing a whole scene. The conventional approach deals with a neural network application. However, the Convolutional Neural Networks (CNNs) have many disadvantages compared with traditional methods for salient region detection. We modified the basic algorithm of salient region detection for VQA task by selecting such image fragments, which have a high probability to be included in a questionnaire. The experiments have been conducted on images from MS-COCO dataset and provided good segmentation results.

1. Introduction
Visual question answering is the task that combines the modalities of computer vision and Natural Language Processing (NLP). The aim of computer vision is to teach machines how to understand the images, while the NPL methods study the interactions between computers and humans using natural language. In VQA task, the both branches are combined in order to archive a higher level of interactions in many practical applications. Generally, the VQA implies an image and a textual question about this image in a form of a few words or a short phrase [1]. The types of questions can be binary (yes/no) or multiple-choice [2]. A relative closely type to multiple-choice variant is “fill in the blank” [3], when the comments ought to be completed by one or several missing words. The main distinction between VQA and other similar applications is that a question has not a pre-determined form until run time. A visual component can be considered as an additional extension of textual question answering. On the one hand, images are higher dimensional and more noisy than pure text. They have another nature than the structural and grammatical rules of language. This means that the direct connection between visual and textural components is impossible. On the other hand, images are close to the real world representation providing a higher level of abstraction.

In this report, we propose an algorithm for detection of salient regions in an image as the regions, which attract a human vision and with a high degree of probability are in focus of the textual questions. Section 2 provides a short literature review. Section 3 describes the proposed algorithm. Experimental results are given in Section 4, while Section 5 concludes the paper.

2. Related work
In spite of first attempts of visual and language integration are referred to 1970s [4], the interest to this problem is emerged during last decade. Application of attention model is one of promising branches in
VQA task. The main idea is to provide correspondences between the saliency regions in an image and textual questions.

The bidirectional retrieval of images and sentences through a multi-modal embedding of visual and natural language data was studied in [5]. The joint multimodal representation learning over images and sentences was implemented using deep Boltzmann machines or log-bilinear models. These authors proposed the global and finer levels of images and sentences and added a ranking objective on a finer level called as the global ranking objective and introduced also stronger fragment alignment objective. region CNN was employed for object detection in an image. The image and sentence fragments were embedded, and the image sentence score as a fixed function of the scores of their fragments was calculated.

In [6], the authors addressed the task of natural language object retrieval in order to localize a target object in a given image based on a natural language query. First, the local descriptors were extracted by CNN_{local} from local regions, while the features were obtained by another network CNN_{global} in the whole image as the scene-level contextual features. Second, the authors proposed a spatial context recurrent ConvNet model as a scoring function on candidate boxes for object retrieval. Thus, the network integrated the spatial configurations and global scene-level contextual information.

Karpathy and Fei-Fei [7] suggested the alignment model, which was based on a combination of CNNs over image regions, bidirectional recurrent neural networks over sentences, and a structured objective that aligned the two modalities through a multimodal embedding. This model was developed under assumption that sentences written by people make frequent references to some particular, but unknown location in the image.

Many researchers use so called a long short-term memory network [8] that produces a caption by generating one word at every time step [9]. The context vector, previous hidden state, and previously generated words are the basic components of such generating. Thus, in [9, 10] the authors proposed two alternative stochastic “Hard” and deterministic “Soft” attention mechanisms. The deterministic attention model was formulated by computing a soft attention weighted annotation vector, while the objective function in hard attention is a variational lower bound on the marginal log-likelihood of observing the sequence of words given image features.

Short literature review shows that non-significant attention is paid for images’ interpretation. Often the images include complex structures with multiple interacting entities that makes difficult to match them with the explicit references in sentences. Usually the researchers preferring to use the marked datasets do not speak how they detect the fragments or objects in an image. Another approach is based on CNN or its modifications, which have tremendous dimensionality, for example the CNN contain approximately 60 million parameters [7].

Good and non-successive examples of salient region detection are depicted in figure 1. Experiments show that Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) method considering only positive gradient values for positive activations provides the best result respect to Grad-CAM [11] and Deconv [12]. As an initial stage for VQA implementation, the preliminary image processing, such as salient region detection, can provide more compact evaluations, directly influencing on the final results.

3. Proposed algorithm
Due to a great variety of possible questions, we propose an algorithm based on detection of salient regions with following partial segmentation of initial image in the salient (foreground) and non-salient (background) regions. Additionally, sometimes we need in information inside a current salient region. For example, a question may be about glasses on a human face but salience technique cannot provide such details. The detail segmentation of holistic image is a time-consuming procedure; however, a partial segmentation in salient regions is reasonable.
Figure 1. Examples of salient region detection using CNN VGG Net-E (19 layers) [13]: a) images from dataset ImageNet ILSVRC-15 [14], b) salient maps built by Guided Grad-CAM method, c) salient maps imposed on the original images.

Generally, the salient region detection methods are represented as the top-down and bottom-up approaches. The top-down approach is based on the supervised learning and high-level processing of semantic data, while the bottom-up approach use the low-level processing of color distributions, color/illumination contrast, relative positions, and contours in local regions. The major researches support the bottom-up approach, which has begun since 1990s [15]. One can find good historical survey, covering more than 230 publications, in [16]. The computational methods of salient region detection are roughly categorized as pixel-based, frequency processing, and region-based methods. The pixel-based methods process the low level features extracted from a surrounding of a central pixel in the single or multiple scales with following their clustering into the salient regions. The pioneer paper in this direction was based on the Simple Linear Iterative Clustering (SLIC) proposed by Achanta et al. [17]. The frequency processing methods use the Gabor filters, Fourier transform, and so forth usually for processing multispectral data. The region-based methods perform a clustering using the region growing or region segmentation with following feature extraction from these clusters.

Our approach is based on the analysis of local regions generating the intensity, color, contrast, edge, angle, and symmetry salient maps with their following superposition. For generating the intensity $SM_{In}$ and color $SM_{Cl}$ maps, we use a model called as DIVision of Gaussians (DIVoG), which was proposed in [18]. For the color map production, the same Gaussian pyramid with number of levels $n = 5$ is built for each RGB-channel. Then all intermediate color salient maps are normalized to fit the range $[0...255]$ with following conjunction into a single salience color map $SM_{Cl}$. Contrast salient map ought to consider three types of feature contrasts, such as color contrast, texture energy contrast, and texture gradient contrast. For obtaining fast algorithm, we apply the texture energy measures developed by Laws [19]. The texture gradient contrast can be evaluated similar but instead of texture measures we used the gradient information. Color contrast is estimated in a color-opponent space with dimension L for lightness and a and b for the color-opponent dimensions – Lab-color space. Color contrast plays a significant role among all contrast components. First, three contrast estimators are normalized. Second, they join in a single map as a salient contrast map $SM_{Cn}$. The edge, angle, and symmetry salient maps contain additional information that can be excluded from the final result.

The normalized salient maps based on intensity, color, and contrast values are fused to the main salient map $SM_{m}$ linearly with empirical coefficients $k_{In}$, $k_{Cl}$, and $k_{Cn}$, respectively:
A visibility of the resultant map (Eq. 1) can be improved by multiply on coefficient $k_G$ provided by Eq. 2, where $d(p_i,p_j)$ is the Euclidean distance from $i$th pixel $p_i$ to a definite foreground or background $j$th pixel $p_j$, $k$ is the parameter, $k=0.5$, $\Omega_F$ and $\Omega_B$ are the areas of foreground and background, respectively.

$$k_G = \exp\left(-k \frac{\min_{i \in \Omega_F} d(p_i,p_j)}{\min_{j \in \Omega_B} d(p_i,p_j)}\right)$$

Application of equation 2 gives more weights for pixels, which are closer to the foreground region and lesser weights for pixels that are closer to the background region. After obtaining of the generalized spatial salient map, the salient regions can be marked by blobs. The details of salient map building one can find in our previous publication [20]. However, this algorithm cannot be directly applied for the VQA task due to several salient regions in an image and/or necessity to find details in the salient regions.

The proposed algorithm includes the following steps:

Input data: an original image.

Step 1. Begin of Loop.

Step 2. Build the salient map of an image using the intensity, color, and contrast attributes.

Step 2. Is the salient region detected? If the salient region is detected, then go to Step 3, otherwise go to Step 7.

Step 3. Mark the selected regions according to the salient map in an image.

Step 4. Remove the selected regions from an image.

Step 5. Enforce the processed image by color and contrast improvement in order to find additional salient regions.

Step 6. End of Loop.

Step 7. Segment all marked salient (foreground) regions in an image into detailed fragments according to the predetermined minimal value of square.

Step 8. Segment non-salient (background) regions if it necessary.

Step 9. Group all detected fragments captured into overlapping rectangles as the image elements suitable for VQA matching.

Output data: the image elements suitable for VQA matching.

4. Experimental results

For experiments, the dataset MS-COCO was yielded [21]. This dataset includes the images of different objects in some context. The dataset contains 82,783 training, 40,504 validation, and 40,775 testing images (approximately 50% for training, 25% for evaluation, and 25% for testing). Also 80 categories of objects are marked. Note that dataset VQA [22] used for training in various tasks involves the questions and answers from two datasets, viz. real 204,721 COCO images from dataset and cartoon images.

Near 900 images were processed during experiments. Some results are represented in figure 2. For segmentation of foreground and background regions, JSEG algorithm was utilized. Experiments show that the segmentation results are improved if a salient map was transformed into binary map with adaptive thresholding value. However, application of additional smoothing of non-salient regions using Gaussian function did not lead to a quality enhancement. The proposed algorithm allows us to detect several salient regions in an image based on binary salient maps. Also, the accuracy of segmentation into the salient regions can be increased by smoothing. As depicted in figure 2, the regions of interest can be foreground or background because of a great variety of possible questions in VQA task.
Figure 2. Salient region detection: a) original images from MS-COCO 63, 869, 56912, 222106, 360347, and 382922, respectively; b) salient maps; c) salient regions with imposed original fragment; d) binary salient (foreground) maps, e) segmentation in foreground regions, f) binary background maps, g) segmentation in background regions.

5. Conclusions
In this paper, we study a problem of salient region detection in an image regarding the VQA task. Experiments show that even well-learned CNN cannot sometimes find all salient regions in the context of human vision. We proposed and implemented a more accurate algorithm with additional segmentation into the salient regions based on the conventional approach – JSEG algorithm. We artificially increased a number of possible salient regions because of a great variety of questions in VQA tasks. Future work in VQA involves the development of a fast matching algorithm between the detected image elements and extracted words from a question.
References

[1] Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K and Darrell T 2017 Long-term recurrent convolutional networks for visual recognition and description IEEE Trans. Pattern Anal. Mach. Intell. 39(4) 677–91

[2] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL and Parikh D 2015 VQA: visual question answering Proc. IEEE Int. Conf. Comp. Vis. 2425–33

[3] Yu L, Park E, Berg AC and Berg TL 2015 Visual madlibs: Fill in the blank image generation and question answering Proc. IEEE Int. Conf. Comp. Vis. 2461–9

[4] Winograd T 1972 Understanding natural language Cogn. Psychol. 3(1) 1–191

[5] Karpathy A, Joulin A and Fei-Fei L 2014 Deep fragment embeddings for bidirectional image sentence mapping Advances in Neural Information Processing Systems (NIPS) vol 2 (Montreal, Canada) pp 1889–97

[6] Hu R, Xu H, Rohrbach M, Feng J, Saenko K, and Darrell T 2016 Natural language object retrieval Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA) pp 4555–64

[7] Karpathy A and Fei-Fei L 2017 Deep visual-semantic alignments for generating image descriptions J. IEEE Trans. Pattern Analysis and Machine Intelligence 39(4) 664–76

[8] Hochreiter S and Schmidhuber J 1997 Long short-term memory Neural Computation 9(8) 1735–80

[9] Xu K, Ba J, Kiros R, Cho K, A. Courville, Salakhutdinov R, Zemel R and Bengio Y 2015 Show, attend and tell: Neural image caption generation with visual attention Proc. 32nd Int. Conf. on Machine Learning (ICML) (Lille, France) 37 pp 2048–57

[10] Srivastava N, Mansimov E and Salakhutdinov R 2015 Unsupervised learning of video representations using LSTMs Proc. 32nd Int. Conf. on International Conference on Machine Learning (ICML) (Lille, France) 37 843–52

[11] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2016 Grad-CAM: Visual explanations from deep networks via gradient-based localization arXiv Preprint arXiv:1610.02391

[12] Zeiler M and Fergus R 2013 Visualizing and understanding convolutional networks arXiv preprint arXiv:1311.2901

[13] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556

[14] Large Scale Visual Recognition Challenge 2015 (ILSVRC2015) Available from http://image-net.org/challenges/LSVRC/2015/. Accessed 29 Sept 2018

[15] Itti L, Koch C and Niebur E 1998 A model of saliency-based visual attention for rapid scene analysis IEEE Trans. Pattern Anal. Mach. Intell. 20(11) 1254–9

[16] Borji A, Cheng M-M, Jiang H and Li J 2014 Salient object detection: A survey. arXiv Preprint arXiv:1411.5878

[17] Achanta R, Shaji A, Smith K, Lucchi A, Fua P and Susstrunk S 2011 SLIC superpixels compared to state-of-the-art superpixel methods Journal of Latex Class Files 60(1) 1–8

[18] Katramados I and Breckon TP 2011 Real-time visual saliency by division of Gaussians Proc. IEEE Int. Conf. Image Processing (New York, USA) 1741–4

[19] Laws KI 1980 Rapid texture identification Proc. SPIE Image Processing for Missile Guidance vol 238 pp 376–80

[20] Favorskaya M and Buryachenko V 2016 Fast salient object detection in non-stationary video sequences based on spatial saliency maps. Intelligent Interactive Multimedia Systems and Services SIST vol 55 ed De Pietro G, Gallo L, Howlett RJ and Jain LC (Springer International Publishing Switzerland) pp 121–132

[21] COCO – Common Objects in Context 2018 Available from http://cocodataset.org/#homeVQA – Visual Query Answering. Available from http://visualqa.org/