BOTTOM-UP ATTENTION GUIDANCE FOR RECURRENT IMAGE RECOGNITION

Hamed R. Tavakoli[⋆†] *Ali Borji* [+] *Rao Muhammad Anwer*[⋆] *Esa Rahtu*[†] *Juho Kannala*[⋆]

[⋆] Department of Computer Science, Aalto University
[+] Center for Research in Computer Vision, University of Central Florida
[†] Department of Signal Processing, Tampere University of Technology

ABSTRACT

This paper presents a recurrent neural network architecture, guided by the bottom-up attention, for the recognition task. The proposed architecture processes an input image as a sequence of selectively chosen patches. The patches are chosen from the salient regions of the input image. Using human driven saliency maps from gaze, the benefit of such a selection process is first shown. Next, the performance of computational models of bottom-up attention are assessed as alternative to human attention.

Index Terms— Recurrent neural networks, image recognition, gaze, saliency, deep neural networks

1. INTRODUCTION

This paper investigates the role of salient regions in the feature learning process of recognition task. The saliency map of an image is used as a means to extract a sequence of image patches. This sequence is then used in a recurrent architecture for image recognition as depicted in Fig. 1.

The recurrent processing of the information for learning feature representations is argued to be a way of efficiently achieving further depth and improving the performance of convolutional deep neural networks. To date, existing recurrent architectures have been processing the whole visual input multiple times, e.g. [1]; or exploit a top-down guided attentional mechanism for localizing and detecting objects and fine-grained details e.g. [2, 3].

On the contrary, this paper proposes to employ processing a sequence of informative patches selected from salient regions of the image, provided by a bottom-up attention mechanism, independent of the task. Our contributions are (1) using human driven saliency, we propose a recurrent neural architecture in order to show that a sequence of patches from bottom-up attention is helpful for learning feature representations, and (2) similar performance boost can be observed using computational saliency models.

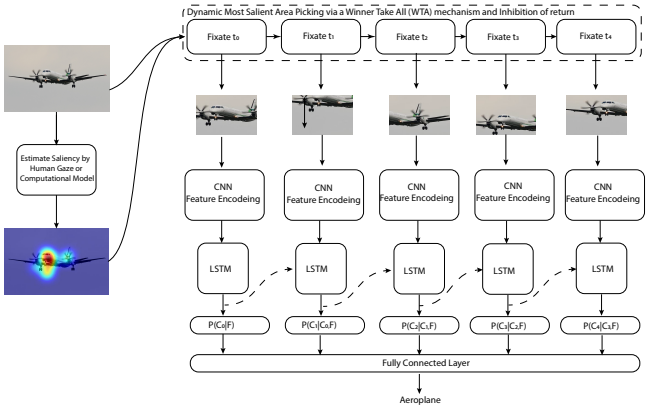

Fig. 1. A framework for saliency guided recurrent model of recognition. The LSTM is unrolled for five fixations. For an input image, the saliency is obtained from human gaze or a computational model. A dynamic fixation selection mechanism using winner-take-all mechanism and inhibition of return is used to choose informative regions guided by saliency. The sequence of image patches from such regions is used for recognition.

2. RELATED WORKS

There exists numerous literature on learning feature representations that going through all is beyond the scope of this paper. For brevity, we summarize some of the most notable ones in this section.

The active vision texts are replete with methods and techniques conducting object detection and localization using saliency as a preprocessing step to choose where to look in the environment. Among all, a series of works by Itti and Koch [4] established a new standard, in which a winner-take-all neural network in conjunction with an inhibition of return mechanism is employed over a saliency map to attend informative locations. This approach has been a basis for many detection algorithms in active and robot vision, e.g. [5, 6]. We follow the same approach for selecting the informative regions from a saliency map; however, the saliency computa-

tion pipeline differs from the above.

The feedback networks [1] employ stacks of neural blocks consisting of convolutional neural networks (ConvNets) and long short-term memory (LSTM) networks to create a deep neural architecture. The feedback networks process the whole image multiple times. One drawback to feedback networks is that it should be trained from scratch due to architectural design. On the contrary, the proposed architecture (1) exploits the internal informative regions of the image and process different regions of the image as a sequence, and (2) can extend any existing ConvNets architecture and be trained using pre-trained networks.

The recent machine learning literature is full of networks with an attention mechanism, which is a feature alignment procedure. Such networks learn to choose the features with respect to the task. The most notable of all is [2], which learns to sequentially extract information from images and videos for specific tasks. The model of [2] is not differentiable and is trained by a reinforcement learning policy. On the other hand, our proposed model is differentiable and the attention mechanism is independent of the task.

There exist also other attention networks, e.g. Xue et al. [7] employ a mechanism to dynamically select visual features for inferring a word contributing to the image caption. The main criticism to these architectures is that the attention is task dependent and difficult to train. Bottom-up attention can be suggested as an alternative to achieve task independence. Within the span of captioning models, Tavakoli et al. [8], explored the contribution of bottom-up attention models for image captioning. In their approach, the saliency is employed as a mechanism to boost the ConvNets features for captioning. They show that once a network is trained on a specific task and data, there is not much contribution from saliency; but the model becomes more robust and improves for handling different visual input. This paper employs bottom-up attention as a cue to extract sequence of image patches for a recurrent neural architecture.

In the next section, we will lay the foundation of our proposed pipeline. We, then, will investigate its usefulness for recognition task using human-driven saliency. Next, the computational models of saliency prediction are evaluated as an alternative to human-driven saliency. Our results indicate that learning a sequence of informative patches in a recurrent architecture improves recognition task.

3. RECURRENT RECOGNITION

We propose a probabilistic neural sequence model for the recognition task by maximizing the probability of the correct classification using the following formulation:

$$\theta^\star = \underset{\theta}{\operatorname{argmax}}\, P(C|\{F\};\theta), \tag{1}$$

where θ is the parameters of the model, $\{F\}$ is a set consisting of the sequence of image patches, and C is the class prediction from a sequence of predictions at each time step. We can, thus, write:

$$\begin{aligned} P(C|\{F\};\theta) = \phi\big(&P(C_N|C_0,\ldots,C_{N-1},F_N;\theta),\ldots,\\ &P(C_t|C_0,\ldots,C_{t-1},F_t;\theta),\ldots,\\ &P(C_0|F_0;\theta)\big), \end{aligned} \tag{2}$$

where C_t, and F_t are predicted class labels and patch input at time step t, respectively, and ϕ is a linear neural mapping function with softmax.

To model $P(C_t|C_0,\ldots,C_{t-1},F_t;\theta)$, we resort to recurrent neural (RNN) models and employ long short-term memory (LSTM) networks [9]. The LSTM networks encode the knowledge of inputs at every time step to the current time by a memory cell m_t. The advantage of LSTM over a vanilla RNN is avoiding vanishing gradients using a control forget gate f_t. The input and output of LSTM are controlled by input gate i_t and output gate o_t, deciding how to handle the data. These gates are formulated as follows:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1}), \tag{3}$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1}), \tag{4}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1}), \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{cm}m_{t-1}) \tag{6}$$

$$m_t = o_t \odot c_t, \tag{7}$$

where $\odot$ is the Hadamard multiplication, the W represents parameter matrices, and $\sigma(.)$ is the sigmoid function, x_t is the feature representation obtained for patch F_t using ConvNets. At time step t, we measure $P = \mathrm{softmax}(m_t)$.

Training. The full architecture is trained using cross entropy loss with Adam optimizer and learning rate of $1e^{-6}$ for 20 epochs. To train a network, we preprocess the images to have zero mean and standard deviation of 1. The code for the recognition pipeline is available at `http://github.com/hrtavakoli/BAR`

4. SALIENCY GUIDED RECOGNITION

Our recurrent recognition approach utilizes a sequence of image patches. To select the patches, we propose using bottom-up attention and a dynamic attention mechanism based on winner-take-all (WTA) and inhibition of return mechanism [10]. The saliency map is fed to a two dimensional layer of neural units. The neuron with maximum saliency (winner) is activated, causing the focus of attention shift to the winning location, where a fixed image patch is extracted to be processed in our recurrent architecture. Along with the winner neuron, a series of inhibitory neurons are also activated to prevent attending the same location. We follow the implementation of [11] and extract patches of fixed size

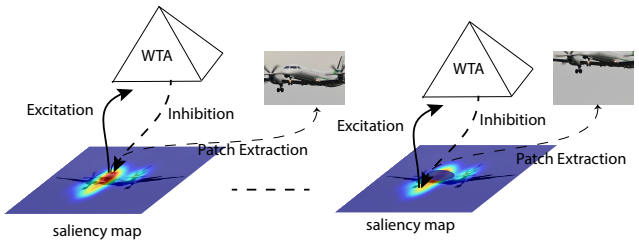

Fig. 2. The winner-take-all process for image patch selection.

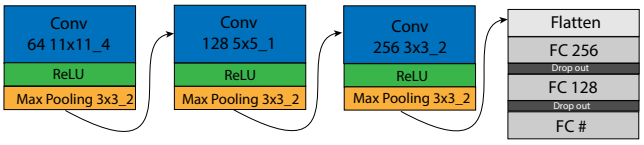

Fig. 3. The architecture of our feature encoding pipeline. Conv $OK \times K\_S$, indicates a convolution with O outputs, $K \times K$ kernel and S stride, similarly for MaxPooling operator $K \times K\_S$, indicates size of kernel and stride, and $FC\#$ indicates the fully connected layer with $\#$ outputs, which for the last layer $\#$ is the number of classes in the database. The same architecture is used as a feed-forward baseline.

of 150×150, centered at the location of the winner neuron. This process is depicted in Fig. 2.

5. LEARNING BY GAZE-DRIVEN SALIENCY

To validate our proposed approach, we employed a relatively shallow ConvNets for feature encoding and trained our network from scratch on saliency maps from human gaze. The settings of this network is depicted in Fig. 3. We use the same architecture as a feedforward baseline with the same training settings as the recurrent architecture.

For this purpose, we use the Pascal Objects Eye Tracking (POET) [12] data. It consists 6270 of the images of the Pascal VOC challenge 2012 [13], which includes 10 of the 20 Pascal classes. It has a total of 178000 fixations of 28 participants, where each image on average has 5.7 fixation per observer. The data is split into a training set of 2800 training images with 280 images per class category and a test set of 3470 images.

To obtain human-driven saliency, we pulled the gaze points of observers together and built a fixation map for each image. Then, the fixation maps were convolved with a Gaussian kernel corresponding to 1° of visual angle in POET. The sequence selection process of section 4 is employed and a series of image patches are extracted.

We compare the results of the proposed recurrent framework as a function of image patches with the performance of the feedforward network trained on the whole image (baseline 1) and the first most salient patch (baseline2). The result is summarized in Fig. 4. As depicted, a feedforward network, trained on the most salient patch, outperforms the same architecture, trained on the whole image. Nevertheless, the recurrent architecture outperforms both models by a large margin. The current experiment indicates that recurrent architecture improves the performance of a network, consistent with the various evidence, including the empirical results of the feedback networks [1].

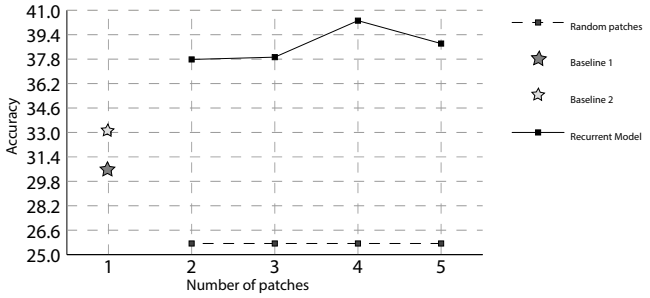

Fig. 4. The performance of recurrent recognition on human-driven image patches in comparison to two baselines on POET data. Baseline 1 is the feedforward network, trained with the whole image as input; Baseline 2 is the feedforward network trained with the first salient patch as input.

To investigate the contribution from the whole image in recurrent model, we also trained the recurrent model for three time-steps with the whole image as the first time step input and the top 2 salient patches as consequent inputs. The results shows the top 1 accuracy of the recurrent model with whole image and the top 2 salient patches as input is 33.54, which is inferior to the top 1 accuracy of the top 2 and top 3 salient patches, 37.78 and 37.93, respectively. This indicates the nature of images is such that the whole image carries too much extra information, e. g. background data, that a crop from the most salient region boosts the recognition and the whole image adversely affects the performance of the recognition pipeline.

We furthermore evaluated the sequence of patches by employing randomly selected patches from a uniform distribution and using the whole image as input in different steps of the recurrent model. Our results indicate that random patches achieve the top 1 accuracy of 25.73; and the whole image as input configuration result in top 1 accuracy of 28.56, which is even worse than the performance of the feedforward network. This clearly indicates that the sequence of image patches matter, i. e. where the model is looking in an image significantly influences its understanding of the content and the results.

6. LEARNING BY COMPUTATIONAL SALIENCY

To replicate human gaze, computer vision has been utilizing saliency modeling as a mean for fixation prediction. Except a limited number of works that directly addressed saccade generation e. g. [14, 15], most of the saliency models focus on predicting the saliency maps. This latter group of models has a well-established evaluation mechanism and community has a better understanding of their performance in comparison to human. They can also be easily used for patch selection in a WTA network.

The applicability of such computational models are assessed. We choose several saliency models and use their saliency maps for patch extraction. These models are SALICON [16], ISEEL [17], and GBVS [18]. SALICON is deep learning model, which is trained end-to-end to establish a multi-resolution regression between image domain and saliency space. It fine-tunes the deep features for the specific task of saliency predictoon. ISEEL is another deep model which exploits the similarity between images to predict the saliency using an ensemble of neural predictors. It treats deep features as generic features and avoids fine-tuning for saliency prediction. The GBVS model is a classic saliency model which relies on low-level image features and Markov chain approach for predicting the saliency.

Fig. 5 summarizes the comparison between saliency models and human. While overall there exist a degree of contribution from computational models, the performance gain does not follow the same trend as the human. All the computational models achieve their peak performance for three most salient patches. The better a model replicates human fixation density maps, the better performance it achieves for salient patch selection.

The experiment also signifies the role of the sequence of patches. While the sequence of four patches is the most informative sequence by human and producing the best prediction model, a sequence of four patches guided by saliency underperforms significantly in comparison to a sequence of three or five patches. This indicates that the computational models does not produce similar to human sequences as expected.

7. DISCUSSION AND CONCLUSIONS

This paper presented a recurrent architecture for image recognition, exploiting the salient regions of images. The experiments showed that learning a sequence of informative image patches is an effective approach for image recognition.

The proposed recurrent architecture consists of a series of convolution operations and a recurrent part. The convolution operations can be initialized from pre-trained networks, which makes the training of the network easier. In this work, we employed three layers of convolutions due to the relatively limited number of training samples that are accompanied with human gaze data and the fact that we trained the neural architectures from scratch.

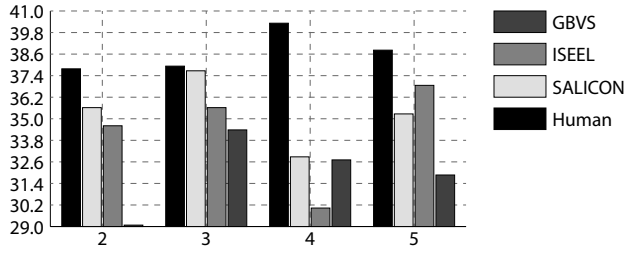

Fig. 5. The performance of recurrent recognition using computational saliency models for patch selection and human as upper-bound. The results of the recurrent approach are shown using 2, 3, 4 or 5 patches (as in Fig. 4).

In the experiments, we learned that (1) the best informative patch is better than the whole image in training a feed-forward network, (2) a recurrent model based on a sequence of informative image patches is superior to a feed-forward model and a sequence of randomly chosen image patches, and (3) despite the gap between saliency models and human has become smaller in fixation prediction task, there is a larger gap in performance of gaze-driven maps (maps from human) and saliency models for selecting informative patch sequences in recognition task.

There has been works that addressed the role of feedback and recurrent architectures. Nevertheless, the feedback has been a recurrent architecture that has been processing the whole input image several times as in [1]. We did not find any significant boost with such a setting, i. e. inputing the whole image several times does not improve the recurrent network over the feed-forward network (28.56 vs. 29.8) and is much inferior to the recurrent architecture with informative patches. The fine-grained object detection has also utilized recurrent attentive networks e. g. [3]. The recurrent attention models require a heavy training in a top-down fashion in order to learn a feature alignment in which the network finds out where to attend. As an alternative, a bottom-up approach for selecting informative patches seems to be an effective approach to be investigated for recognition tasks.

8. ACKNOWLEDGMENT

The support of NVIDIA Corporation with the donation of the GPU used in this work is acknowledged.

9. REFERENCES

[1] A. R. Zamir, T.-L. Wu, L. Sun, W. Shen, B. E. Shi, J. Malik, and S. Savarese, "Feedback networks," in

CVPR, 2017.

[2] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu, "Recurrent models of visual attention," in *NIPS*, 2014.

[3] Hakan Bilen and Andrea Vedaldi, "Integrated perception with recurrent multi-task neural networks," in *NIPS*, 2016.

[4] Laurrent Itti and Christof Koch, "Computational modelling of visual attention.," *Nat Rev Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[5] Simone Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, Springer-Verlag New York, Inc., 2006.

[6] R. Kasturi, D. Goldgof, R. Ekambaram, G. Pratt, E. Krotkov, D. D. Hackett, Y. Ran, Q. Zheng, R. Sharma, M. Anderson, M. Peot, M. Aguilar, D. Khosla, Y. Chen, K. Kim, L. Elazary, R. C. Voorhies, D. F. Parks, and L. Itti, "Performance evaluation of neuromorphic-vision object recognition algorithms," in *22nd International Conference on Pattern Recognition (ICPR)*, 2014.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *32nd International Conference on Machine Learning*, 2015.

[8] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen, "Paying attention to descriptions generated by image captioning models," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, 1997.

[10] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiol*, vol. 4, pp. 219–227, 1985.

[11] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res.*, vol. 40, 2000.

[12] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari, "Training object class detectors from eye tracking data," in *ECCV*, 2014.

[13] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The pas- cal visual object classes challenge 2012 (voc2012) results," .

[14] Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä, "Stochastic bottom-up fixation prediction and saccade generation," *Image and Vision Computing*, vol. 31, no. 9, 2013.

[15] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao, "Learning to predict sequences of human visual fixations," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, 2016.

[16] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] Hamed R. Tavakoli, Ali Borji, Jorma Laaksonen, and Esa Rahtu, "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, 2017.

[18] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *NIPS*, 2007.