NEURAL NETWORK WITH SALIENCY BASED FEATURE SELECTION ABILITY

Yunong Wang[⋆†] *Huanyu Bian*[⋆†] *Nenghai Yu*[⋆†]

[⋆] Department of Electronic Engineering and Information Science,
University of Science and Technology of China
[†] Key Laboratory of Electromagnetic Space Information,
Chinese Academy of Sciences, Hefei 230027, China

ABSTRACT

Convolutional neural network (CNN) is inspired by the biological structure of human visual system (HVS). And there are still mechanisms in HVS that are worthy learn from. We were inspired by the function of feature selection in HVS which is named as visual saliency and proposed Sal-Mask connection and Ada-Sal Network to implement similar function in neural networks. In this paper, we did further research and tried three different improvement schemes on Ada-Sal Network. By visualizing features we illustrated the feature selection ability of Sal-Mask connection helps the neural network to extract features from more subtle details of input images. The experiment results proved that neural network works better with this saliency based feature selection function. We also found that Sal-Mask connection works best on features from the first convolutional connection.

Index Terms— Convolutional neural networks, feature selection, image classification, visual saliency

1. INTRODUCTION

Convolutional neural networks (CNNs) have shown great efficiency and potential these years. They became state of the art on computer vision and image process tasks [1–5]. The mainstream research idea on improving neural networks is to craft deeper networks with more neurons, meanwhile use regularization methods to constrain the overfitting problem caused by excessive amounts of parameters. For example, LeNet-5 [6] only has 7 layers with 6M parameters. Now VGG-16 [1] is composed of 16 layers with 138M parameters. The batch normalization [7] increases 30% calculation each layer.

However, this design is inefficiency comparing to real human visual system (HVS). Human beings have the ability of rapidly locating important object regions and casting detail analysis on them, while briefly glance other less important regions or even ignore them. This mechanism of select concerning contents actively is called visual saliency [8]. Visual saliency improves human's ability of image processing significantly. It has been incorporated in a variety of computer vision and image processing tasks [9–13].

Development of neural networks learns a lot from the biological structure of HVS. The HVS is composed of layers of interconnected neurons. It has layers of simple and complex cells whose activation are determined by the magnitude of input signals falling into their receptive fields. Convolutional layers in CNN resemble simple and complex cells in the HVS while fully connected layers in a CNN resemble higher-level inference and decision making in the HVS. Based on the correspondence between human visual system and neural networks, it is a natural choice to reform the structure of neural networks to fulfil this feature selection procedure of HVS.

There have already been works applying deep models on saliency tasks. But there are still few works to integrate the mechanism of saliency into neural networks. We proposed a new structure named Sal-Mask connection before [14]. It takes feature maps extracted by convolutional connection and a saliency mask as input. The saliency data is at high value at regions corresponding to those important objects, while at low value at less important areas. By taking advantage of these prior knowledge in saliency mask, Sal-Mask connection can readjust the features in neural networks. Features from more important areas will be assigned higher weights, so they can contribute more in the network. By this way, we succeeded to integrate saliency mechanism into neural networks. Furthermore, we came out with the idea of training saliency mask together with the feature maps. We hereby proposed Ada-Sal Network [15]. It generates saliency mask from weighted sum of input saliency maps. The weights are trained using feedback from the rest of the network. The learned saliency mask works better with the Sal-Mask connection.

In this paper, we went deeper with Ada-Sal Network. We applied Sal-Mask connection on features from deeper layers. According to the result of experiments, we found that it works best to apply Sal-Mask connection on the first layer features. We visualized the features before and after Sal-Mask connection and discussed how these features are selected. The rest of the paper is organized as following. Section 2 introduces the

This work is supported by the National Science Foundation of China (No.61371192, No.61103134)

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

design of Ada-Sal Network and three improvement program we designed. Section 3 describes the structure and parameters of networks we did experiments on. Section 4 illustrates the experiment details and results. Section 5 visualizes the features and saliency masks learned and discusses how Sal-Mask connection effects the neural network.

2. ADA-SAL NETWORK

2.1. Sal-Mask connection

As mentioned before, we proposed Sal-Mask connection to emulate the feature selection mechanism of visual saliency. It takes raw features and a saliency mask as input. We execute an element-wise multiplication between each raw feature map and the saliency mask. The output is enhanced features. The enhanced features are then input into the rest of the neural network. They replace the role of raw feature. The Sal-Mask connection feedforwards as follow:

$$E^{z}_{jk} = y_{jk} \cdot s_{j}. \tag{1}$$

$$E^{y}_{jk} = h(E^{z}_{jk}). \tag{2}$$

Denoting y_{jk} as the j-th pixel of the k-th raw feature map, s_j the j-th pixel of the saliency mask, E^{y}_{jk} the j-th pixel of the k-th enhanced feature map. E^{z}_{jk} is the value before activation function and h is the activation function used.

The formula of back-propagation is:

$$\delta_{jk} = \delta^{e}_{jk} \cdot s_j \cdot f'(z_{jk}). \tag{3}$$

$$\delta^{s}_{j} = \sum_{k} \delta^{e}_{jk} \cdot y_{jk}. \tag{4}$$

Denoting δ^{e}_{jk} as the error back-propagated to the Sal-Mask connection, δ_{jk} the error passed to the raw feature maps and δ^{s}_{j} the error passed to the saliency mask.

Areas corresponding to important objects are at higher values on saliency map, the features of important objects will be amplified during feedforward and the error data will also be amplified during back-propagation. On the other hand, features of unimportant objects will be depressed. Hence the neural network is able to select features in the same way as HVS. Here the saliency mask describes the importance of features. During the training of normal neural network, some features from main object are relatively weak. These features may be ignored because they will not make big enough stimulus. Instead, some features from irrelevant objects contribute more to the classification result because they are quite strong. But with the help of Sal-Mask connection, the neural network will be able to filter those irrelevant but strong features and make right classification based on those amplified features.

2.2. Ada-Sal Network

As can be imagined, the Sal-Mask connection needs a proper saliency mask to work well. At first we used existent saliency algorithms to compute the saliency mask. But current saliency algorithms do not suit our task well. So we designed Ada-Sal Network [15]. In Ada-Sal Network, a new saliency mask is gained by weighted sum of several different saliency maps. The weights are trained together with the neural network. Figure 1 describes the structure of Ada-Sal Network.

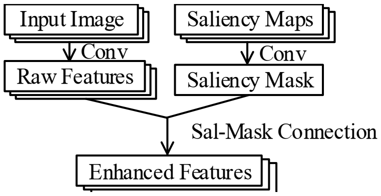
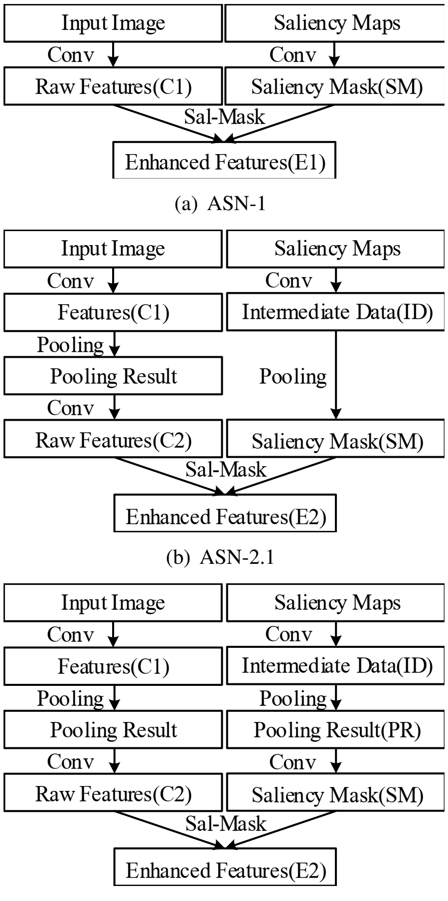


Fig. 1. Sal-Mask connection and Ada-Sal Network.



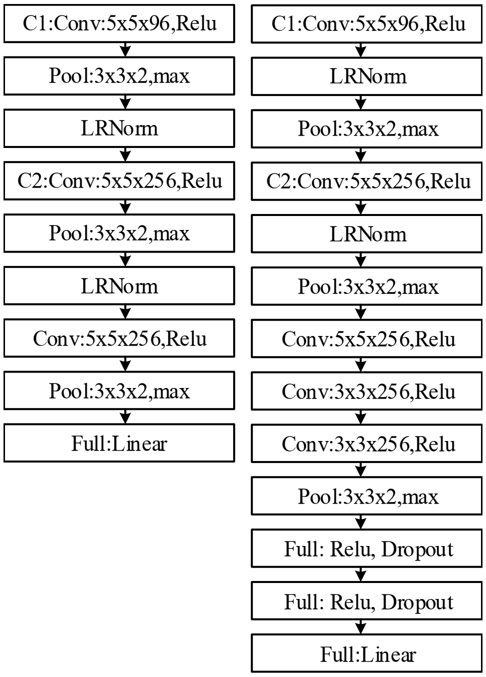
(a) ASN-1

(b) ASN-2.1

(c) ASN-2.2

Fig. 2. Three Ada-Sal Networks using different reform schemes.

In our previous work [15], we have proved that Ada-Sal Network can get better saliency mask by combining different saliency maps. It can take advantage of different saliency algorithms. In this paper, we applied Sal-Mask connection on features from deeper layers and observed how the feature selection works. We used three schemes in our experiments as



(a) Mod-LeNet (b) AlexNet

Fig. 3. Benchmark Networks.

illustrated in Figure 2. The Sal-Mask connection is applied on the first layer features (C1) in ASN-1. In ASN-2.1 and ASN-2.2, the Sal-Mask connections are applied on the second layer features (C2), but the saliency masks are generated in different ways. The saliency mask is generated by using a convolutional connection with one kernel on input saliency maps in ASN-1. It is at the same size of the C1 feature maps. In ASN-2.1, we get intermediate data from input saliency maps using a convolutional connection with one kernel. Then the intermediate data is pooled to get a saliency mask matches the size of C2 feature maps. In ASN-2.1, the intermediate data is generated using a convolutional connection with 96 kernels. These 96 intermediate saliency maps are then pooled. Another convolutional connection with one kernel is applied on the pooling result to get one saliency mask.

3. NETWORK STRUCTURES

3.1. Network Structure

We chose two typical CNNs as benchmark networks. One is LeNet-5 [6] but we skipped two fully connection and called it Mod-LeNet and the other is AlexNet [16]. Due to limit of GPU we used, we modified the parameters a little. The two benchmark networks are illustrated in Figure 3. We get three Ada-Sal Networks based on each benchmark network as described in Figure 2. The detailed parameters will be introduced in next section.

3.2. Detailed Parameters

In Figure 3, a square frame stands for a Connection. Conv is short for convolutional connection. It has three parameters: kernel size $m \cdot m$, kernel number K and stride s. In Figure 3, the parameters are in format of $m \cdot m \cdot K$ and the stride is set to 1. We do padding before each convolutional connection. Pool is short for pooling connection. It has three parameters: pooling size k, pooling stride s and pooling function - whether max-pooling (short as max) or average pooling (short as ave). In Figure 3, the parameters are in format of $k \cdot k \cdot s$, and pooling function is described. LRNorm is short for local response normalization [16]. LRNorm connection has four parameters: k, n, α and β. In our experiments, we set k as 1, n as 3, α as 0.00005 and β as 0.75. Full is short for full connection, which depends on the size of both input layer and output layer. Relu is short for rectified linear units [16]. Dropout is for dropout connection [17]. In our experiments, we set r as 0.5 for every dropout connection.

4. EXPERIMENTS

We used two datasets for our experiments - CIFAR-10 [18] and STL-10 [19]. The CIFAR-10 dataset consists of 60000 32×32 pixels color images in 10 classes, including 50000 training images and 10000 test images. The images in STL-10 are at the size of 96×96 pixels. The labeled data in it includes 5000 training images and 8000 test images for 10 classes. In this paper, we used two kinds of saliencies. One type of saliency we used is from Levin et al. [20], we called it Alpha saliency for short The other is from Cheng et al. [21] and we call it RC saliency.

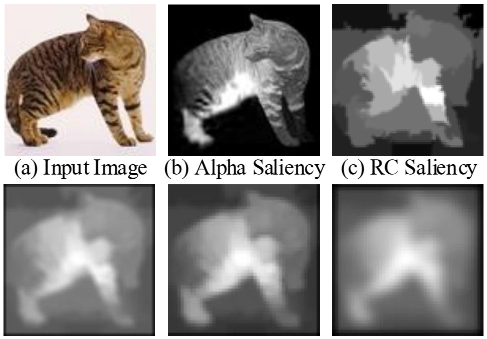
The experiment results are shown in Table 1 and Table 2. As can be seen, all Ada-Sal networks work better than benchmark networks, but the one with Sal-Mask connection applied on the first layer features works the best.

Table 1. Results on CIFAR-10

Network-Accuracy (%)	Mod-LeNet	AlexNet
Benchmark	84.4±0.1	83.9±0.2
Scheme 1	**85.6±0.2**	**85.2±0.3**
Scheme 2	84.8±0.2	84.4±0.2
Scheme 3	84.6±0.2	84.5±0.2

Table 2. Results on STL-10

Network-Accuracy (%)	Mod-LeNet	AlexNet
Benchmark	64.65±0.2	67.3±0.2
Scheme 1	**66.46±0.2**	**69.1±0.2**
Scheme 2	65.9±0.2	68.7±0.1
Scheme 3	66.1±0.3	68.5±0.2



(a) Input Image (b) Alpha Saliency (c) RC Saliency

(d) SM in ASN-1 (e) SM in ASN-2.1 (e) SM in ASN-2.2

Fig. 4. Input data and saliency masks from ASNs.

5. DISCUSSIONS

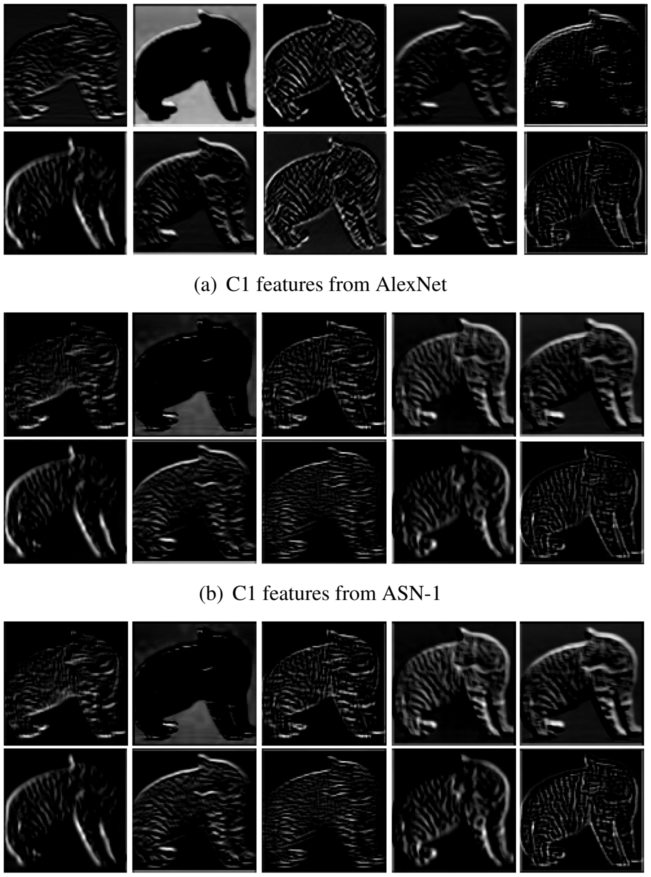
The Ada-Sal Networks improve the performance by 1-2 percentages while only increase 1 or 2 more convolutional connection with very few parameters and a non-parameter Sal-Mask connection. We pick an image from the STL-10 dataset. Figure 4(a) shows this image. Figure 4(b) and (c) are two type of input saliency maps we used in the experiments. Figure 4(d) (e) are saliency masks that learned from ASN-1, ASN-2.1 and ASN-2.2.

We visualized the C1 features from AlexNet and ASN-1 based on AlexNet in Figure 5. The E1 features in ASN-1 play the same role as C1 features in AlexNet. We observed several characteristics from these illustrations:

- Features learned by ASN-1 are similar to those learned by original network.
- ASN-1 learns more types of features. Most of features from original network are from conspicuous textures - edges of the cat, stripes at the cat body and the background. However, ASN-1 can learn features that original network cannot, such as those horizontal edges and vertical edges.
- E1 features from ASN-1 are clearer with higher values than C1 features from AlexNet.
- In ASN-1, features from irrelevant objects are depressed. Such as the features of the background and the white edge caused by the zero padding.

In summary, comparing to the original network, ASN-1 shows greater ability of extracting features. During training of original network, inconspicuous but relevant features could be ignored. But in ASN-1, these features are reinforced by the Sal-Mask connection. Thus they can generate bigger responses and will not be ignored.

The second layer features are illustrated in Figure 6. It is hard to judge which is the best by naked eyes. According to the performances, ASN-1 works best. As we can see in Figure 4, the saliency masks from ASN-2.1 and ASN-2.2 are coarser then the saliency mask from ASN-1. We have proved in our previous work that the saliency mask determines the performance of Sal-Mask connection. So ASN-2.1 and ASN-2.2 do not work as well as ASN-1 because of the coarse saliency masks. But the saliency mask generated largely depends on the input saliency maps. Considering deep models being used to learn saliency map directly, our future work will focus on developing model with no need of prior input saliency maps.

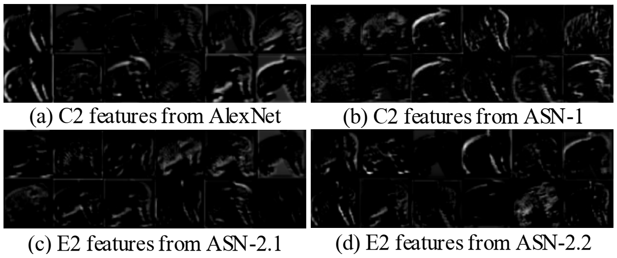


(a) C1 features from AlexNet

(b) C1 features from ASN-1

(c) E1 features from ASN-1

Fig. 5. First layer features from AlexNet and ASN-1.



(a) C2 features from AlexNet (b) C2 features from ASN-1

(c) E2 features from ASN-2.1 (d) E2 features from ASN-2.2

Fig. 6. Second layer features from AlexNet and ASNs.

6. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[5] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff, "Top-down neural attention by excitation backprop," in *European Conference on Computer Vision*. Springer, 2016, pp. 543–559.

[6] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[7] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[8] Ali Borji and Laurent Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[9] Shai Avidan and Ariel Shamir, "Seam carving for content-aware image resizing," in *ACM Transactions on graphics (TOG)*. ACM, 2007, vol. 26, p. 10.

[10] Sai Bi, Guanbin Li, and Yizhou Yu, "Person reidentification using multiple experts with random subspaces," *Journal of Image and Graphics*, vol. 2, no. 2, 2014.

[11] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li, "A user attention model for video summarization," in *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 2002, pp. 533–542.

[12] Vidhya Navalpakkam and Laurent Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2049–2056.

[13] Ruobing Wu, Yizhou Yu, and Wenping Wang, "Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 867–874.

[14] Yunong Wang, Nenghai Yu, Taifeng Wang, and Qing Wang, "Improve neural network using saliency," in *International Conference on Image and Graphics*. Springer, 2015, pp. 417–429.

[15] Yunong Wang, Nenghai Yu, and Taifeng Wang, "Adasal network: emulate the human visual system," *Signal Processing: Image Communication*, vol. 47, pp. 519–528, 2016.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[18] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[19] Adam Coates, Honglak Lee, and Andrew Y Ng, "An analysis of single-layer networks in unsupervised feature learning," *Ann Arbor*, vol. 1001, no. 48109, pp. 2, 2010.

[20] Anat Levin, Dani Lischinski, and Yair Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, 2008.

[21] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.