Learning Propagation Rules for Attribution Map Generation – *Supplementary Material* –

Yiding Yang[1], Jiayan Qiu[2], Mingli Song[3], Dacheng Tao[2], and Xinchao Wang[1]

[1] Stevens Institute of Technology, Hoboken, NJ 07030, USA
{yyang99,xinchao.wang}@stevens.edu
[2] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
{jqiu3225@uni.sydney.edu.au,dacheng.tao@sydney.edu.au}
[3] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
brooksong@zju.edu.cn

We provide in this document more experimental results and details of the proposed method. Specifically, we show the structure of the customized networks, the IDs of images used in the main manuscript, the top-5 predictions of the composite image, the evaluation of the pointing game, and more visualization results on the ImageNet dataset with a pre-trained VGG-16 model, We provide figures of the sensitivity analysis to give a better illustration on the influence of the hyper-parameters in our proposed method. We also provide a video to show the learning process of attribution maps.

1 Details of the customized models

1.1 Customizing CNN on the CIFAR-10 dataset

The customized CNN model we used on CIFAR-10 dataset contains four convolutional layers followed by ReLU activations and two fully connected layers. Dropout with ratio 0.25 are also adopted after MaxPooling layers and fully connected layer. The details of architecture are shown in Tab. 1.

1.2 Customizing CNN on the MNIST dataset

The customized CNN model we used on the MNIST dataset contains two convolutional layers followed by ReLU activation and two fully connected layers. The details of architecture are shown in Tab. 2.

Custom CIFAR-10 CNN
Conv 2D (3x3, 32)
ReLU
Conv 2D (3x3, 32)
ReLU
MaxPooling (2x2)
Dropout2d(p=0.25)
Conv 2D (3x3, 64)
ReLU
Conv 2D (3x3, 64)
ReLU
MaxPooling (2x2)
Dropout2d(p=0.25)
Linear (256)
ReLU
Dropout2d(p=0.5)
Linear (10)

Table 1: Details of the customized CNN on the CIFAR-10 dataset.

Custom MNIST CNN
Conv 2D (5x5, 20)
ReLU
Conv 2D (5x5, 50)
ReLU
Linear (500)
ReLU
Linear (10)

Table 2: Details of the customized CNN on the MNIST dataset.

2 Information of the images

We make the composite image, shown in Fig. 10 of the main manuscript, by using two images of two different classes from the test set of the ImageNet dataset . The IDs of these two images are 8194 and 8213 respectively. The names of the top five predictions of the composite images are: rhinoceros beetle, landrover, snowplow, tow truck, and pickup truck respectively.

The ID of the image used in Fig. 1 in the main manuscript is 98855.

3 Pointing game evaluation

For the point game evaluation, the models are trained as multi-label classifier and we adopt binary cross entropy function as the loss the function. The first archtecture of the learnable plugin module is used which is similar as the convolution without summation and shares parameters within one layer.

Tab. 3 shows the pointing game results of different methods. Our method yields a very competitive performance on both Pascal VOC and COCO datasets. Note that the purpose of the subjective evaluation is to measure how the generated attribution map aligns with human perception. The performance of such evaluation does not necessarily relate to the performance of explanation, as how the model perceives the input may differ from how human does.

4 More visualizations

We provide in Fig. 2, Fig. 3 and Fig. 4 more visualization of the samples from the test set of ImageNet with a pre-trained VGG-16 model. Some interesting phenomenons can be observed from these visualization:

– Most compared methods are very sensitive to the high-frequency components, like the grass in the third row of Fig. 2 and ocean wave in the last row of Fig. 3.

Method	VOC07 Test		COCO14 Val	
	All	Diff	All	Diff
Cntr.	69.6	42.4	27.8	19.5
Grad [7]	76.3	56.9	37.7	31.4
DConv [11]	67.5	44.2	30.7	23.0
Guid. [9]	75.9	53.0	39.1	31.4
MWP [12]	77.1	56.6	39.8	32.8
cMWP [12]	79.9	66.5	49.7	44.3
RISE [3]	86.9	75.1	50.8	45.3
GCAM [4]	86.6	74.0	54.2	49.0
sMask [1]	88.0	76.1	51.5	45.9
Ours	85.7	74.2	51.7	46.1

Table 3: Evaluation of pointing game. All methods adopt VGG-16 network as the base model. Our proposed method provides comparable performance during such subjective evaluation. Note that sMask learn the mask under several predefined constraints and needs 1600 iterations to learn a single mask. Our method generates the mask based on the gradient of input and needs only 40 iterations. The difficult subset is adopted from MWP [12].

– Some methods tend to highlight regions that are irrelevant to the target class, like the last row of Fig. 4, and the last row of Fig. 2.
– Some methods give wrong signs to the contribution in the attribution map, like the fourth row of Fig. 4.

5 Sensitivity analysis on hyper-parameters

We provide additional figures shown in Fig. 1, to gain a better understanding of the sensitivity of hyper-parameters. We can see from Fig. 1 that the AUC performance is not sensitive to most of the hyper-parameters, such as the iteration, learning rate, and weight of the mask loss. For the shift and Gamma parameters, setting them to a reasonable values within a wide range, such as 0.8-0.9 and 10-30 respectively, will lead to similar AUC results.

6 Illustration of the learning process of attribution maps

We provide a video to show how the attribution map updates during the optimization process, where we can see that the irrelevant regions are indeed removed from the attribution map during the process while the attribution map focuses on the most critical regions of the target class.

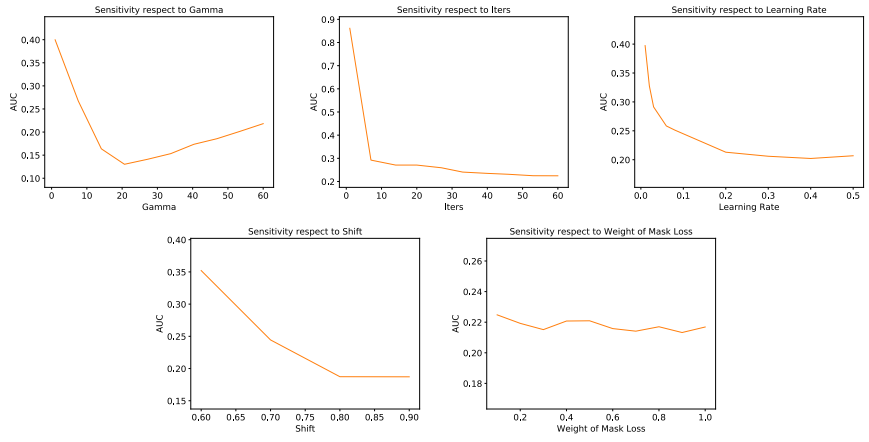

Fig. 1: Sensitivity analysis. Each figure depicts the relationship between a hyper-parameter and its AUC of the MoRF curve. We can see that the AUC is stable after a few iterations, and is not sensitive to the learning rate as well as the weight of the mask loss. For the shift and Gamma parameters, setting them to values in a reasonably wide range leads to very similar AUC results.

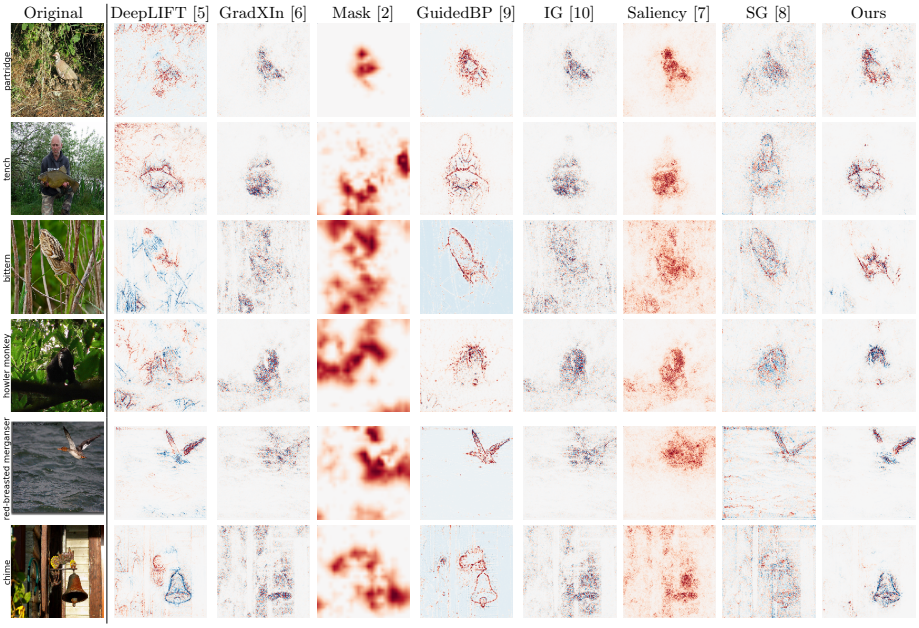

Fig. 2: More visualization of the compared methods. Our proposed method provides more concentrated attribution maps.

Fig. 3: More visualization of the compared methods. Notice that almost all the other methods are sensitive to the background with high-frequency components, such as the ocean wave in the last row.

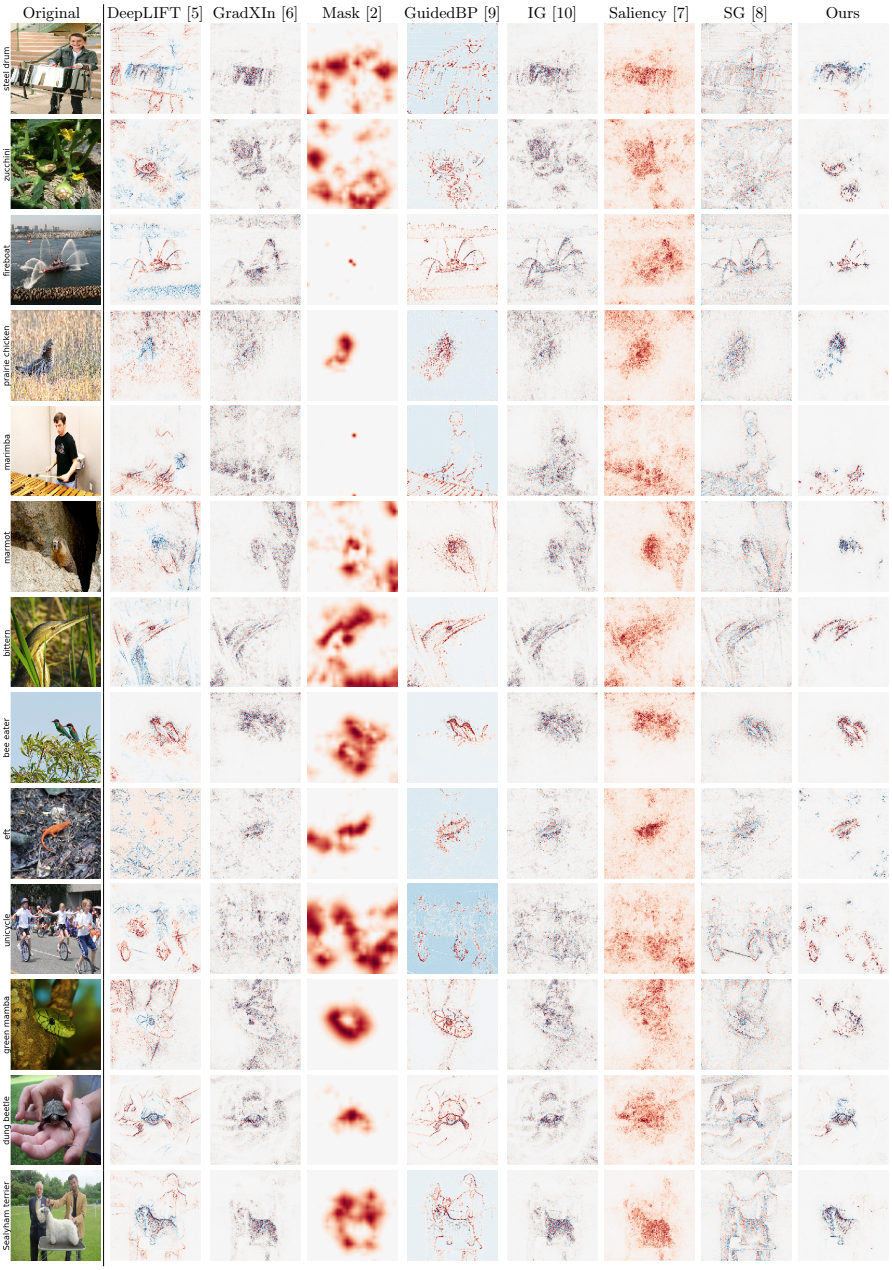

Fig. 4: More visualization of the compared methods. Our method is less sensitive to the high-frequency components and irrelevant areas, such as the background building in the third row.

References

1. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2950–2958 (2019)
2. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3429–3437 (2017)
3. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
5. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3145–3153 (2017)
6. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
7. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
8. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
9. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
10. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3319–3328 (2017)
11. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833 (2014)
12. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)