Supplementary Materials: Towards Global Explanations of Convolutional Neural Networks with Concept Attribution

Weibin Wu[1], Yuxin Su[1*], Xixian Chen[2], Shenglin Zhao[2], Irwin King[1], Michael R. Lyu[1], Yu-Wing Tai[2]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, [2]Tencent
{wbwu, yxsu, king, lyu}@cse.cuhk.edu.hk, {xixianchen, henryslzhao, yuwingtai}@tencent.com

1. More Attacking Results

We report the intermediate attacking results to confirm the desired property of global feature occluders: intra-class generalization and specificity. That is, a global feature occluder should disrupt decisive feature filters **common** to the **specified** category of images. Since directly examining zillions of internal neurons is prohibitively expensive, we resort to the solutions as follows.

1.1. Intra-class Generalization

This property means that a global feature occluder can fool model predictions on samples of the target class, including those that the global feature occluder has never seen during training. Intra-class generalization reflects that global feature occluders can disturb feature filters **common** to the target class. To examine this property, for a target class, we first occlude clean ImageNet images with the corresponding global feature occluder. Then we investigate the model performance on the resultant samples.

Table S1 reports the average results over the 100 random classes covered in the main paper. We view the ImageNet validation set as the test set. Although global feature occluders have never seen the test set of ImageNet during training, there is a significant degradation of model accuracy on both the training and test set after the employment of matched global feature occluders. It corroborates that the learned global feature occluders possess a strong capacity for intra-class generalization.

1.2. Specificity

This property requires that a global feature occluder cannot severely mislead model decisions on samples from the other classes that it does not target. Specificity confirms that global feature occluders can undermine feature filters **specific** to the target class. To investigate this property, we still focus on the same 100 classes we explore before. Concretely, for a global feature occluder, we apply it to the instances from the remaining 99 categories that it does not initially target. Then we record the model accuracy on the resultant images.

Table S2 summarizes the average results over the 100 classes. Global feature occluders witness a much-limited attack success rate under this setup, which satisfies our expectation of learning a class-specific global feature occluder. Besides, there is an observable degree of vulnerability of models to unmatched global feature occluders. It may be because similar image categories are likely to share some critical feature detectors. Worse still, ImageNet often requires fine-grained classification between similar subspecies of an animal [S3], such as great white sharks and tiger sharks, which can aggravate the susceptibility of trained classifiers to global feature occluders.

To further illustrate the specificity of derived global feature occluders, we exhibit examples of learned global feature occluders, along with correctly classified legitimate samples and their distorted counterparts in Figure S1. Remarkably, global feature occluders are not identical for different classes. Besides, the misclassifications incurred by global feature occluders do not converge to the same incorrect label. Therefore, it again shows that global feature occluders are class-specific.

From Figure S1, we can also discover that model decisions often deviate a lot from the original ones when encountering global feature occluders, while the perturbation can hardly harm human perception. It confirms that feature occluders do not need to degenerate images in a human-notable fashion, or align with meaningful image regions to humans. This observation is consistent with the finding in the literature that in contrast to human visual systems, the building blocks of model reasoning, namely, the feature detectors, are susceptible to structured noises [S5, S4, S2, S1].

*Corresponding author.

Model	Clean		Perturbed	
	Top-1 Training Accuracy	Top-1 Test Accuracy	Top-1 Training Accuracy	Top-1 Test Accuracy
ResNet-50	0.8771	0.7756	0.0973	0.1662
GoogLeNet	0.8115	0.7494	0.0907	0.2606
VGG-16	0.8095	0.7358	0.1001	0.1616

Table S1: Average top-1 accuracy of different models on clean images and the counterparts perturbed with matched global feature occluders. A global feature occluder can successfully fool model predictions on both the training and unseen test data of the target class, which confirms its intra-class generalization capacity.

Model	Clean		Perturbed	
	Top-1 Training Accuracy	Top-1 Test Accuracy	Top-1 Training Accuracy	Top-1 Test Accuracy
ResNet-50	0.8771	0.7756	0.6063	0.5586
GoogLeNet	0.8115	0.7494	0.5893	0.5662
VGG-16	0.8095	0.7358	0.5392	0.5035

Table S2: Average top-1 accuracy of different models on clean images and the counterparts perturbed with unmatched global feature occluders. Global feature occluders exhibit much-limited attack success rates under this scenario, which reflects that they are class-specific.

References

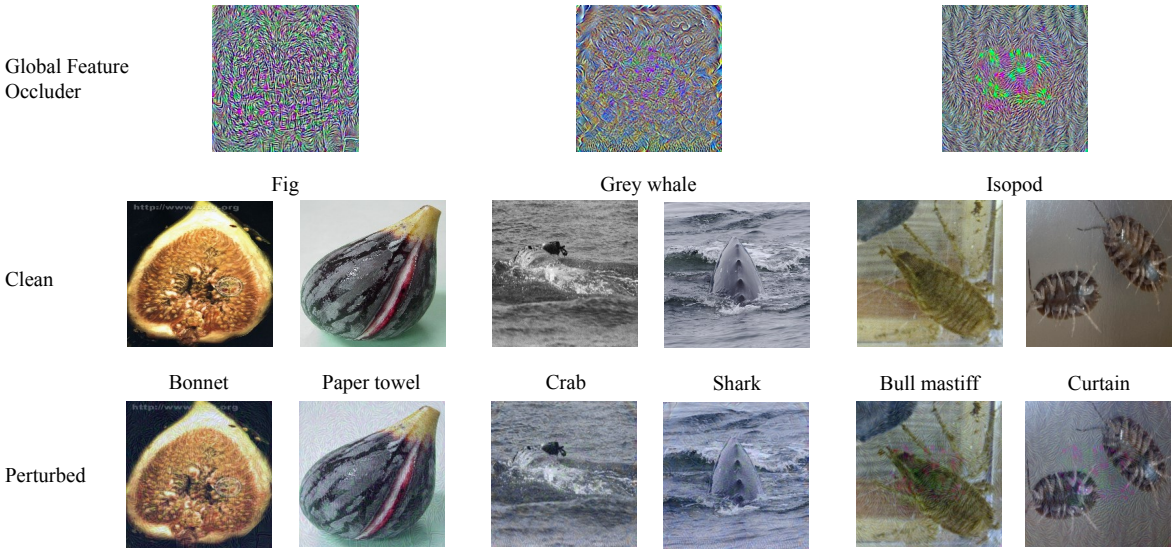
[S1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1

[S2] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1765–1773, 2017. 1

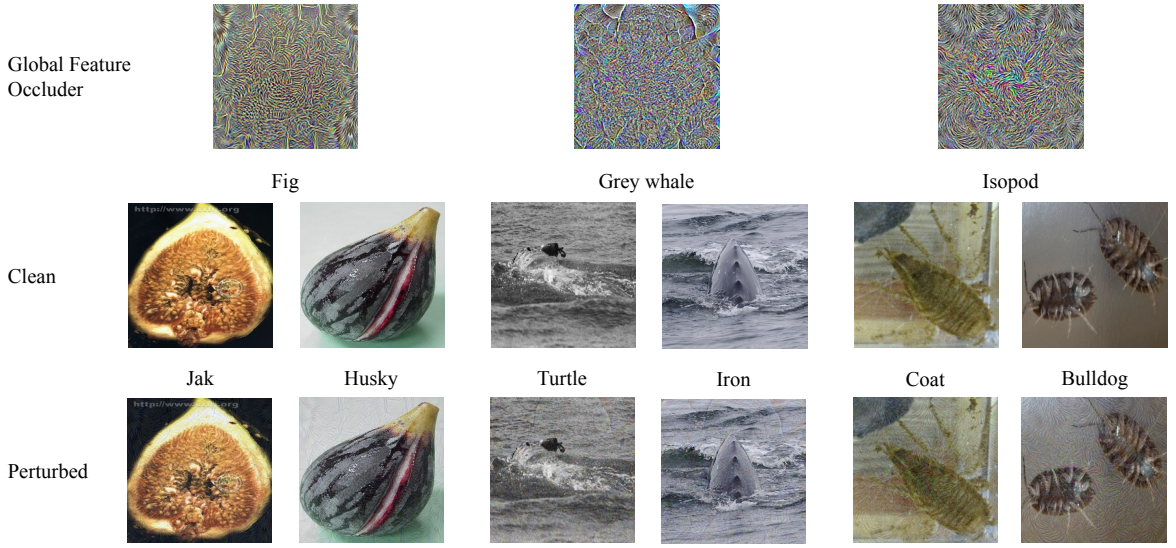
[S3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 3, 4

[S4] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[S5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1

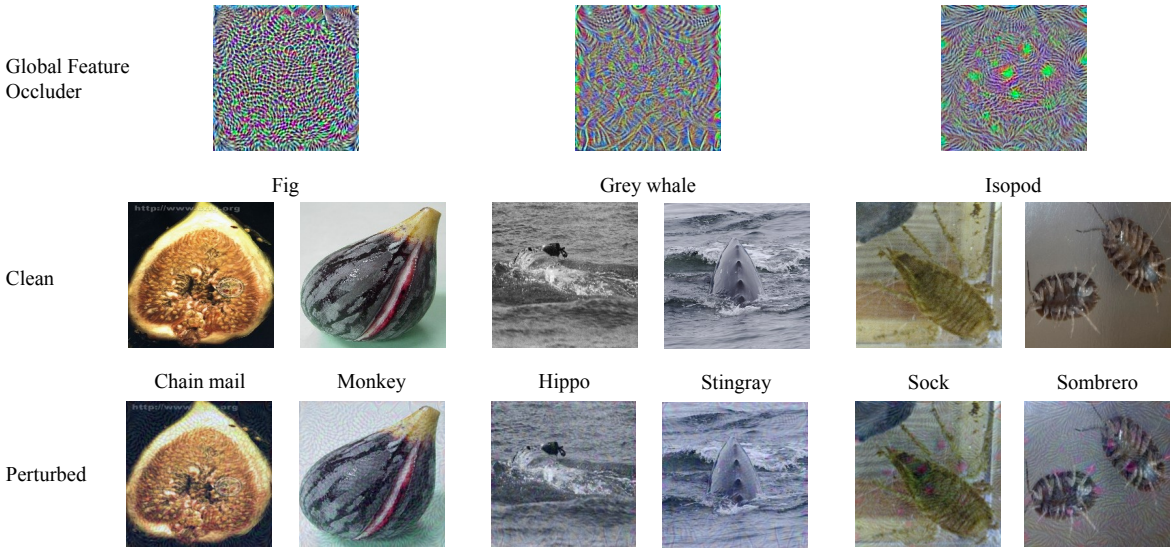


(a) ResNet-50



(b) GoogLeNet

Figure S1: Examples of learned global feature occluders and their effects on model decisions. In every two columns of one model, the images from the top row to the bottom one are the learned global feature occluder for a target class (with pixel values magnified to natural image ranges for visibility), the clean images of this class from the ILSVRC 2012 training set [S3], and the resultant images distorted by the global feature occluder of this class, respectively. The labels above these images report the model predictions. All model decisions on clean images are correct. We see that: (1) global feature occluders are not identical for different classes, and (2) the misclassifications on perturbed images do not converge to the same label. These observations further validate that global feature occluders are class-specific.



(c) VGG-16

Figure S1: (Continued) Examples of learned global feature occluders and their effects on model decisions. In every two columns of one model, the images from the top row to the bottom one are the learned global feature occluder for a target class (with pixel values magnified to natural image ranges for visibility), the clean images of this class from the ILSVRC 2012 training set [S3], and the resultant images distorted by the global feature occluder of this class, respectively. The labels above these images report the model predictions. All model decisions on clean images are correct. We see that: (1) global feature occluders are not identical for different classes, and (2) the misclassifications on perturbed images do not converge to the same label. These observations further validate that global feature occluders are class-specific.