Data-free Universal Adversarial Perturbation and Black-box Attack Supplementary Material

Chaoning Zhang*
chaoningzhang1990@gmail.com

Philipp Benz*
pbenz@kaist.ac.kr

Adil Karjauv*
mikolez@gmail.com

In So Kweon
iskweon77@kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST)

A. Training UAP with Images from a Single Class.

We report results of our Simple-UAP algorithm trained on images sampled from a single class in Table 1. The performance is close to the case of using images of all classes. Additionally, training on different single classes leads to different dominant labels, and the results are summarized in Table 2. When the single training class is fixed, the resulting dominant label with different runs is the same in most cases. However, when the single training class is changed, the corresponding dominant label also changes. This constitutes another empirical evidence that the dominant label does not necessarily occupy large regions in the image space as hypothesized by [3]. If the dominant label phenomenon is caused by the fact that the dominant label occupies large regions in the image space, the resulting dominant label is not supposed to change with the choice of a single training class.

Table 1. Simple-UAP algorithm comparison on the ImageNet validation dataset with the metric of fooling ratio (%). The single class algorithm trained only on samples from one class ("quilt").

Method	AlexNet	GoogleNet	VGG16	VGG19	ResNet152
Single Class	93.2	85.7	91.8	90.8	77.6
Normal Training	96.5	90.5	97.4	96.4	90.2

B. GD-UAP Results.

Here, we report the results for GD-UAP (see Figure 1). The results of GD-UAP resemble the trend for Cosine-UAP, in the sense of higher $\cos(v, x+v)$ than $\cos(x, x+v)$ for most latter layers and a positive relationship between cosine $\cos(v, x+v)$ and fooling ratio. Overall, the Cosine-UAP results in higher $\cos(v, x+v)$ compared with GD-UAP, especially in the very last few layers, which partially explains why the Cosine-UAP yields a higher fooling ratio.

*Equal contribution

Table 2. Different classes used for training lead to different dominant label classes.

Trained Class	Dominant Label Class
Carbonara (ID: 959)	Brain Coral (ID: 109)
Quilt (ID: 750)	Candle (ID: 470)
Miniature Pinscher (ID: 237)	Irish Wolfhound (ID: 170)
Accordion (ID: 401)	Peacock (ID: 84)
Tow Truck (ID: 864)	Dome (ID: 538)
Scoreboard (ID: 781)	Pillow (ID: 721)
Sussex Spaniel (ID: 220)	Timber Wolf (ID: 269)
Shield (ID: 787)	German Short-haired Pointer (ID: 210)

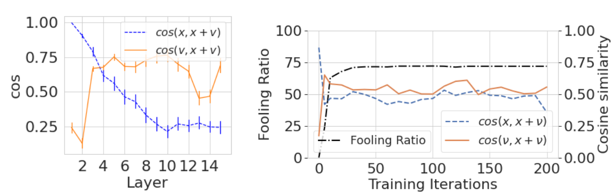

Figure 1. Layer-wise (left) model and step-wise (right) analysis of the GD-UAP on the model response in the untargeted setting.

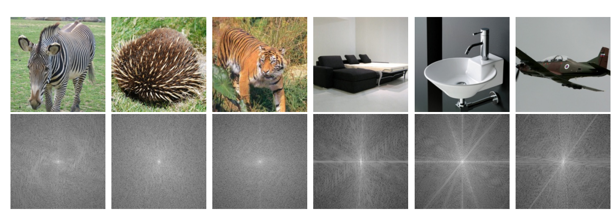

Figure 2. Three examples of robust samples (first three columns) and vulnerable sample (last three columns) with their respective fourier transforms.

C. Visualization of Robust and Vulnerable Samples.

The visualization of robust and vulnerable samples is shown in Figure 2. Robust samples tend to have more high-frequency content.

D. Practical Data-free Black-Box Attack.

Experiment Setup. As widely reported in previous literature [3], random noise has very limited influence on accuracy. Our ablation study in Table 10 of the main manuscript also confirms this by showing the average accuracy under uniform noise perturbation is as high as 67.3%. Our analysis in Sec 4.3 of the main manuscript shows that content with repetitive patterns usually has a high influence on the joint DNN response when it is combined with another independent content as the DNN combined input. Inspired by the above finding, we design adversarial perturbation with repetitive patterns. The patterns that we investigate include horizontal, vertical, and checkerboard patterns, which are shown in Figure 4. Taking a horizontal pattern, for example, the pixel values are constant in the horizontal direction but receptively change in the vertical direction with values either $-\epsilon$ and ϵ. Following [2], we set ϵ to 0.1. One hyper-parameter here is the width of the lines and we empirically find that the width of 2 pixels achieves satisfactory performance. For both horizontal and vertical patterns, we set it to 2 pixels. The checkerboard's square size is set to 2×2 pixels. We optionally remove some HF content from the original images which can further enhance the attack success rate. For a fair comparison with [2], we finally clip the final resulting perturbation with the l_∞ constraint ϵ. We adopt two common ways for removing the high frequency with Fourier transform(FT) or SVD [1]. For adopting FT to remove HF content, we adopt the approach introduced in [4] and set the bandwidth to 36. For adopting SVD to remove HF content, we keep content corresponding to the top 12 singular values.

Figure 3. "Window screen" class image samples.

Additional Analysis. As demonstrated in the main manuscript, UAP has a dominant influence on the joint model response triggered by adversarial examples. Our designed adversarial perturbation with repetitive patterns is also universal since it can be added to any random image. For analyzing its influence on the joint model response of adversarial samples, *i.e.* images + our designed pattern, we first investigate the model prediction taking only the designed pattern.

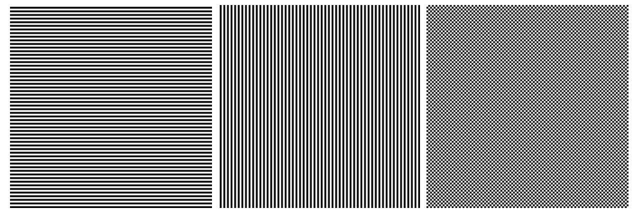

Figure 4. Different types of patterns used in our data-free black-box attack: horizontal pattern (left), vertical pattern (center), and checkerboard pattern (right). Patterns are amplified for better visualization.

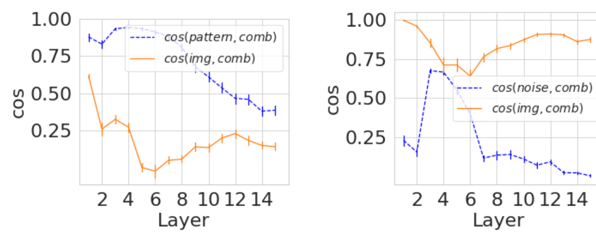

Figure 5. Contribution of checkerboard pattern and images ("img") in the joint model response ("comb") on VGG16 (left image). Contribution of uniform random noise and images ("img") in the joint model response ("comb") on VGG16 (right image).

Taking the checkerboard pattern as an example, we find that it is classified by most networks as "window screen". Some sample images with the ground-truth label"window screen" are shown in Figure 3. It is interesting to observe the checkerboard pattern in those sample images of the class window screen. Analogous to the analysis in Figure 2 (left) in the main manuscript, we report layer-wise influence analysis for the checkerboard in Figure 5 (left). Different from UAP, the optimization-free checkerboard pattern has the most dominant influence on the intermediate layers. As an ablation study, we also report the result for uniform noise in Figure 5 (right). The contrasting behavior of $cos(image, combined)$ indicated by orange line in Figure 5 shows that the noise influence on the model is only limited to shallow layers, while the influence of the checkerboard pattern is also significant in middle and/or deep layers. This provides insight on why checkerboard is more effective than random noise.

References

[1] H Andrews and CLIII Patterson. Singular value decomposition (svd) image coding. *IEEE transactions on Communications*. 2

[2] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. *NeurIPS*, 2020. 2

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 1, 2

[4] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019. 2