Supplementary Material for "*TRM-UAP: Enhancing the Transferability of Data-Free Universal Adversarial Perturbation via Truncated Ratio Maximization*"

Due to the space limitation, we introduce the details of how to generate artificial images that are used in the curriculum optimization algorithm and show some extra adversarial examples in this supplementary material. Here, all the experimental setups are consistent with our paper.

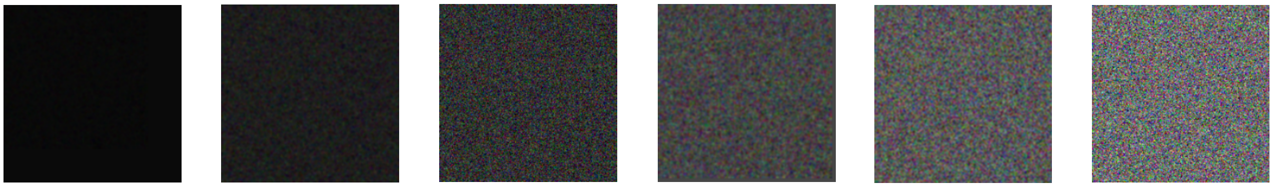
In the curriculum optimization algorithm, artificial images are sampled from a predefined distribution and the complexity of artificial images is increasing gradually. When artificial images are generated by Gaussian noise, the magnitude of the standard deviation of Gaussian distribution increases as the number of training iterations increases. Thus, the set of artificial images D_t is defined as

$$D_t = \{\boldsymbol{x} | \boldsymbol{x} \sim \mathcal{N}(\mu_0,\ \sigma_0 + \gamma \cdot \lfloor t/t_0 \rfloor)\} \tag{1}$$

where μ_0, σ_0 are the initial mean and standard deviation of Gaussian distribution $\mathcal{N}$, γ is the step size, t is the number of current iterations, and t_0 is a predefined threshold to adjust the increasing rate of standard deviation σ, *i.e.*, the standard deviation σ increases by γ per t_0 iterations. The visualization of artificial images generated by Gaussian noise distribution is shown in Fig. 1. On the other hand, we also generate artificial images by jigsaw images distribution defined as:

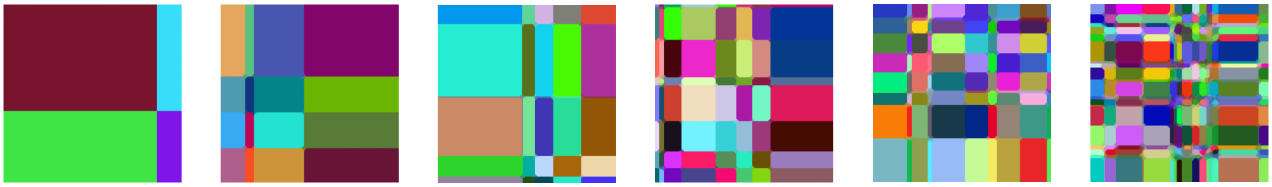
$$D_t = \{\boldsymbol{x} | \boldsymbol{x} \sim \mathcal{J}(\mathcal{F}_0 + \gamma' \cdot \lfloor t/t'_0 \rfloor)\} \tag{2}$$

where $\mathcal{F}_0$ is the initial frequency of jigsaw image distribution $\mathcal{J}$, γ' and t'_0 are similar to γ and t_0 defined in Gaussian noise distribution. Artifical jigsaw images with different frequency $\mathcal{F}$ are displayed in Fig. 2. In the experiments, hyperparameters are set appropriately for different surrogate models. Additionally, some extra adversarial examples are shown in Fig. 3.



(a) $\sigma = 10$ (b) $\sigma = 30$ (c) $\sigma = 50$ (d) $\sigma = 70$ (e) $\sigma = 90$ (f) $\sigma = 110$

Figure 1. The visualization of artificial images generated by Gaussian noise distribution. As the standard deviation σ increases, the complexity of artificial images is increasing gradually.



(a) $\mathcal{F} = 1$ (b) $\mathcal{F} = 3$ (c) $\mathcal{F} = 5$ (d) $\mathcal{F} = 7$ (e) $\mathcal{F} = 9$ (f) $\mathcal{F} = 20$

Figure 2. The visualization of artificial images generated by jigsaw image distribution. As the frequency $\mathcal{F}$ increases, the complexity of artificial images is increasing gradually.

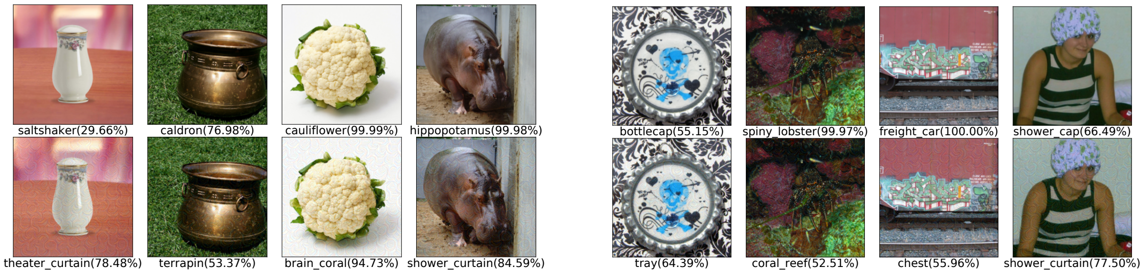


Figure 3. Examples of our TRM-UAP attack on the ImageNet dataset (top: original examples; bottom: adversarial examples). The annotation represents the classification label and the probability predicted by CNNs.