Project 8: Adversarial Patch Attack

Previously we have studied the adversarial attacks, which are realized by crafting pixel-wise additive perturbations that are constrained by L-p norm bound. These attacks, as you may notice, are not that feasible in the real-world (can you really add small noises to a physical object?). In contrast, there is another type of adversarial attacks that are designed to be "physical" – Adversarial Patch attack. As its name suggests, the idea is to create a patch (or a sticker) that can be attached to somewhere in the image to fool the classifier. In this project, you will act as an attacker and craft such adversarial patches to mislead neural networks.

Requirement:

- Implement the Adversarial Patch generation process described in [1] for CIFAR-10 classifiers. For simplicity you can generate rectangular patches instead of circular ones.
- Evaluate the effect of patch size. For example, generate patches with the size 3x3, 5x5, 7x7, 16x16, and measure the *untargeted* attack success rate (ASR) as a function of the patch size. Are there any visual patterns you could see in the generated patches?
- A careful attacker may want to disrupt the model in a way that whenever the patch is there the model will predict a certain target class. Repeat step 2 but this time focus on the *targeted* attack success rate. Try using each of the 10 classes as the target class. How does targeted ASR compared to untargeted ASR? Are there any visual patterns when generating targeted patches?
- Transfer the generated patches on one model to others, e.g., from ResNet to DenseNet or VGG. Measure the untargeted/targeted ASR. To what extent do the patches transfer across different models?

References:

[1] Brown et al., "Adversarial Patch," (~~https://arxiv.org/pdf/2002.05709.pdf~~)

CORRIGENDUM: Reference available at https://arxiv.org/pdf/1712.09665.pdf