Simple Black Box Attack

Alina Munir, Sarthak Trivedi

5 Aug 2021

1 Introduction

With the growth in Machine learning models in our daily lives, the requirement to protect them against the presence of adversaries also become vital. One of the major concerns are the misclassification of adversarial images in security critical situations such as identification of traffic signs for autonomous cars. Even well trained DNNs can misclassify the adversarial images that are unnoticable from a human eye. These adversarial images can be created with and without the knowledge of the model. Mostly in real life scenarios we do not have the known model, so we do the adversarial black box attack on these images, where we know the input images and the output classification.

In this project we do the black box attack, where we try to add perturbations in out input images iteratively in such a way that our model misclassifies them as another targeted output class. (we try to understand a simple black box attack) in untargeted whereas we only make the model misclassify the image, given that the human counter part can still classify correctly, perturbations are not so strong that human cannot identify them.

2 System Description

The main intuition behind this simple attack is to exploit the fact that the main object of any object classification model is to learn patterns in high dimensional space. Each class in this space can be thought of as a space spanning collection of these points in this high dimensional space. The main objective of this simple black box attack is to systematically move a sample in this vector space by adding little values to it. This is done in one of the two ways : either we can directly attack the model in this high-dimensional space where it has learnt some patterns namely the pixel space, or we can attack it in a different vector space (which is invertible in nature) namely the discrete cosine space. A discrete cosine transform expresses a sequence of a finite number of data points as a sum of cosine functions oscillating at different frequencies.

These two vector spaces are unique, this is because of these have orthonormal basis vectors. This ensures that perturbation can be added or subtracted independently without hindering the overall progress of the attack. Pixels are

chosen at random without repetition and a fixed small epsilon value is added or subtracted to the image to see if it reduces the confidence of the model (this can be thought of as systematic function decent and the function here is the models confidence). The softmax output of the model is used to evaluate how confident the model is given the correct label for the input image.

The original implementation for [3] is made in pytorch and can be found here [1]. We re-implemented simba in Tensorflow to learn more about the framework and the technique itself, we tried to implement both single and batch attack on images but unfortunately we could not implement the batch based attack due to the shift in libraries.

3 Evaluation and Experimental Details

Imagenet[2] is the most frequently used dataset to train transferable models. Most of the vision systems in production use transferred knowledge based on imagenet weights. This makes breaking these base transfer models very interesting mostly because if anyone can breach them, the models trained over these base models are also susceptible to these attacks. However due to lack of computational resources we decide to run this attack on a subset of Imagenet dataset[4] which consists of 10 most easily classifiable classes from Imagenet.

There are three quantities that are of prime importance in any blackbox attack:

1. Success rate - Given all the images how many were successfully misclassified.

2. Queries - The number of times the provided model was queried to find a perturbation that would result in misclassification.

3. L_2 norm - How much perturbation was added to the final image for it to be misclassified. More perturbation would definitely lead to worse classifier performance and in worst cases could also make it impossible for humans to classify altered images.

4 Results

We evaluate the attack in both Pixel and DCT space. We run both the test on 100 random samples from the imagenette dataset. We keep the same epsilon for both the attacks. We test the attacks on a pre-trained ResNet50.

Simba with DCT mode performs far better than in Pixel mode. It performs less numbers of queries on average and adds less amount of perturbation to the final image. This can be seen in Table 1. However with imagenette both attacks more or less have similar success rates. Since imagenette contains most easy classifiable classes from the imagenet dataset the success rate reduces significantly.

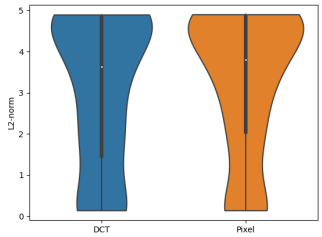

Figure 1: Distribution of Queries

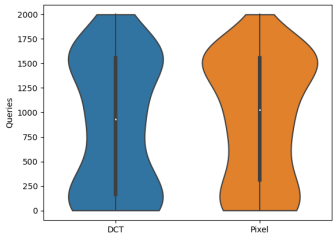

Figure 2: Distribution of L2-norm

Attack Mode	Success Rate	Average Queries	Average L2-Norm
DCT	**0.68**	**868.01**	**3.02**
Pixel	0.66	923.45	3.24

Table 1: Table comparing Pixel and DCT modes of SimBA

5 Conclusion

SimBA is a simple yet powerful black box attack. We tried to re-implement it for both DCT and Pixel modes. We also tried to evaluate if SimBA causes any changes to saliency of the given image since the goal of slaiency to some degree is to also capture interesting low level features in the image. However we found out that neither of the attack modes change bottom up saliency (which is implemented in open-cv). However more time could have helped us evaluate this even further with models like DeepGaze. Since these newer neural network based saliency approaches might react differently to SimBA.

References

[1] cg563. cg563/simple-blackbox-attack: Code for icml 2019 paper "simple black-box adversarial attacks".

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

[4] Jeremy Howard. Imagenette.