Adversarial Machine Learning

Homework Assignment 2

The assignment is due by the end of the day on Thursday, March 2.

Objective:

Implement black-box evasion attacks against deep learning-based classification models.

Part 1: Boundary Attack

The Boundary Attack is a black-box evasion attack based on the paper by Brendel et al. (2018), which we covered in Lecture 5. The following notebook in the Adversarial Robustness Toolbox explains the implementation of the boundary attack on ImageNet images. The boundary attack uses only the final predicted label by a black-box model to create adversarial samples, i.e., it is a decision-based attack.

Dataset: We will use the FIGRIM dataset, consisting of 4,436 images of 11 scenes. Examples of images from the dataset are shown in Figure 1. The dataset and a Data Loader code can be downloaded from this Shared folder on OneDrive. The file with the images is named 'SCENES_700x700.zip' (287 MB).

Figure 1. Example images from the dataset.

Task 1: Train a ResNet50 model for classification of the scene classes in the dataset.

Perform hyper-parameters tuning to obtain an accuracy on the test dataset above 90%.

Estimated running time: between 5 and 15 minutes.

Report (20 marks): (a) Report the classification accuracy for the train set, validation set, and test set of images. For full marks, it is expected to report a test accuracy above 90%. Plot the training and validation loss and accuracy curves. If applicable, provide any other observations regarding the training of the model.

Task 2: Implement an untargeted boundary attack against the trained model.

Step 1: Select one image from the test dataset to be used for creating an adversarial sample. Make sure that the DL classifier correctly predicts the class of the image. To check it, plot the image with the ground truth label and the predicted label by the DL model.

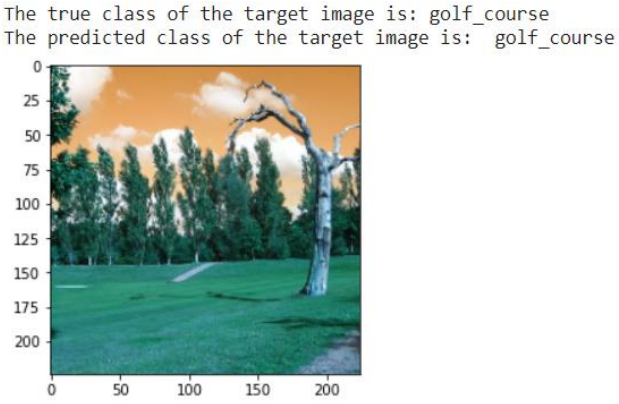

Figure 2. Selected image for the attack.

When you try to create a single adversarial sample, pay attention to the shapes of the data. Note that the model works with batches of images, so when we have a single image of size (224, 224, 3) we will need to first reshape it into a batch of one image of size (1, 224, 224, 3) before we pass it to the model. Otherwise, if we feed the image with size (224, 224, 3), we will get an error.

Step 2: Using the boundary attack, create an adversarial image that will change the label of the selected original image. You can use similar parameters for the attack as in the above example notebook in the ART toolbox, or if you wish you can adopt different parameters. Print the L2 norm and the label of the adversarial image for each step of the attack, similar to Figure 3.

Step 3: Plot the adversarial image with the predicted label by the classifier (as in Figure 4)

Step 4: Plot the difference between the selected image and the adversarial image, as in Figure 5. Show the color bar of the difference.

Estimated time: between 10 and 20 minutes.

Report (25 marks): (a) Plot the required figures in Steps 1 to 4. (b) Write between 5 and 10 sentences to explain the boundary attack.

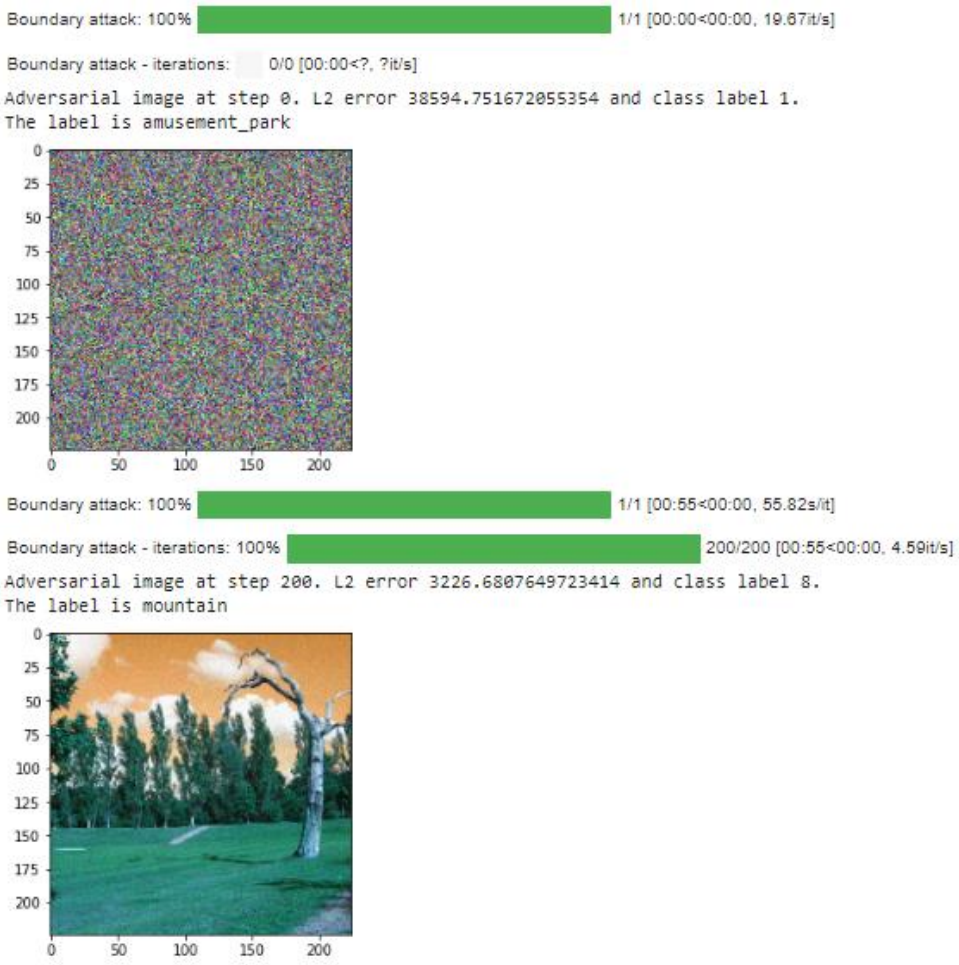

Figure 3. Untargeted boundary attack.

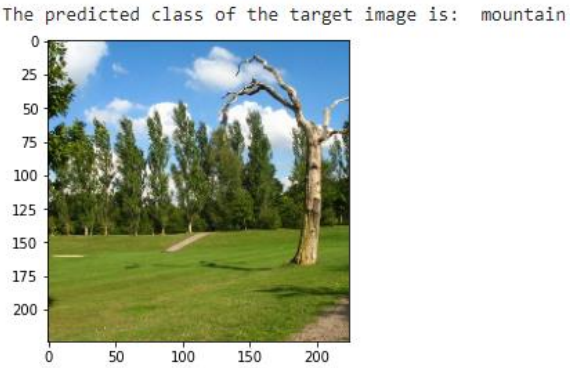

Figure 4. Generated adversarial image.

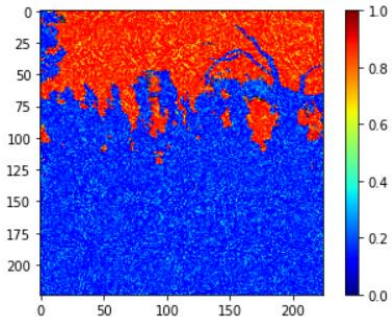

Figure 5. Difference between the original and adversarial image.

Task 3: Implement a targeted boundary attack against the trained model.

Step 1: Select one image from the test dataset to be used for creating an adversarial sample. Plot the image with the ground truth label and the predicted label by the DL model.

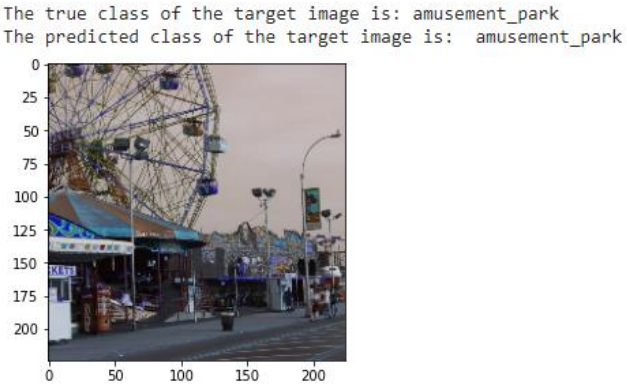

Figure 6. Selected image for the attack.

Step 2: Select one image from the test dataset with the target class label. Plot the image with the ground truth label and the predicted label by the DL model.

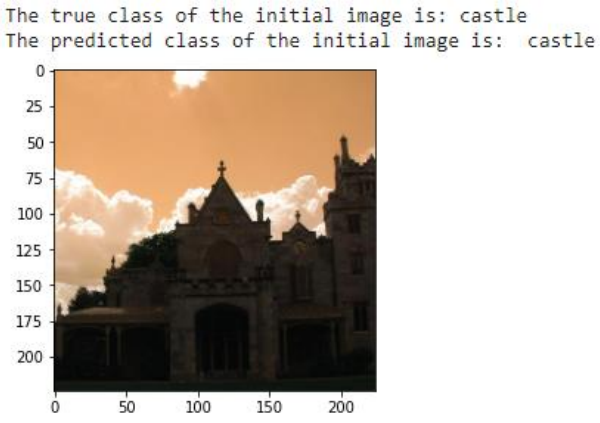

Figure 7. Selected image from the target class label.

Step 3: Using the boundary attack, create an adversarial image that will change the label of the selected image to the target label. Print the L2 norm and the label of the adversarial sample for each step of the attack, similar to Figure 3.

Step 4: Plot the adversarial image with the predicted label by the classifier.

Figure 8. Created adversarial image.

Estimated time: between 20 and 40 minutes.

Report (25 marks): (a) Plot the required figures in Steps 1 to 4. (b) Briefly describe the targeted boundary attack.

Part 2: Transferability Attack

Create your own substitute model for classification of Dog Breeds, and transfer adversarial samples to a corresponding model hosted by Clarifai. This is a black-box attack, because we don't have access to the model hosted by Clarifai.

Task 1: Train a deep-learning model for classification of images of different dog breeds.

Dataset: We will use a subset of the Stanford Dogs dataset, consisting of 1,961 images of 12 dog breeds. Examples of images from the dataset are shown in Figure 9. The dataset and a Data Loader code can be downloaded from this Shared folder on OneDrive. The file with the images is named 'Dataset_dog_breeds.zip' (29 MB), whereas the file with the labels is named 'labels_dog_breeds.csv'.

Perform hyper-parameters tuning to obtain an accuracy on the test dataset above 50%. This dataset is more challenging, because the differences between the dog breeds are very subtle. Some of the images do not provide sufficient details to be reliably classified by the model. And also, for such challenging dataset, it is required to have a larger number of images, that can capture the characteristics of each breed.

Estimated running time: between 10 and 20 minutes.

Report (10 marks): (a) Report the classification accuracy for the train set, validation set, and test set of images. For full marks, it is expected to report a test accuracy above 50%. Plot the training and validation loss and accuracy curves. If applicable, provide any other observations regarding the training of the model.

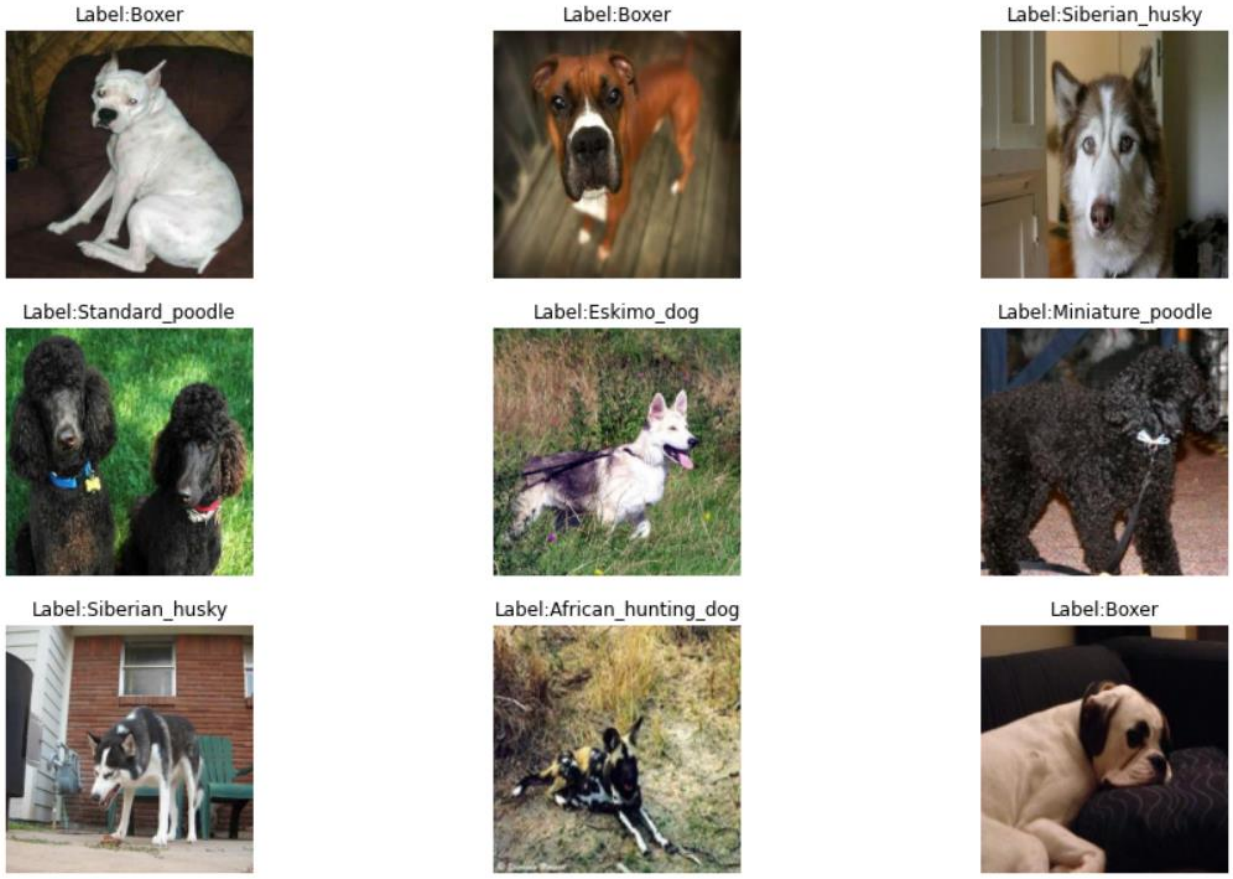

Figure 9. Example images from the dataset.

Task 2: Create adversarial samples against the Clarifai's web ML model for dog breed classification.

Step 1: Select one image from the test dataset to be used for creating adversarial samples. Plot the image with the ground truth label and the predicted label by the DL model.

```
The true class of the image is: Doberman
The predicted class of the image is:  Doberman
Doberman
```

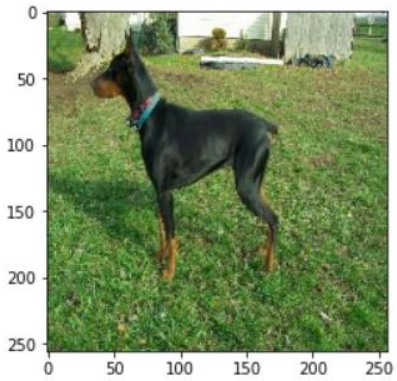

Figure 10. Selected image for the attack.

Step 2: Visit the Clarifai website for predicting dog breeds: https://clarifai.com/datastrategy/dog-breeds/models/dog-breeds-classifier

You will be prompted to sign in, and I think that the easiest way is to sign in with an existing Gmail account or GitHub account.

The API is shown in Figure 11. Click on the button "Try your own image or video" to upload the selected image in Step 1.

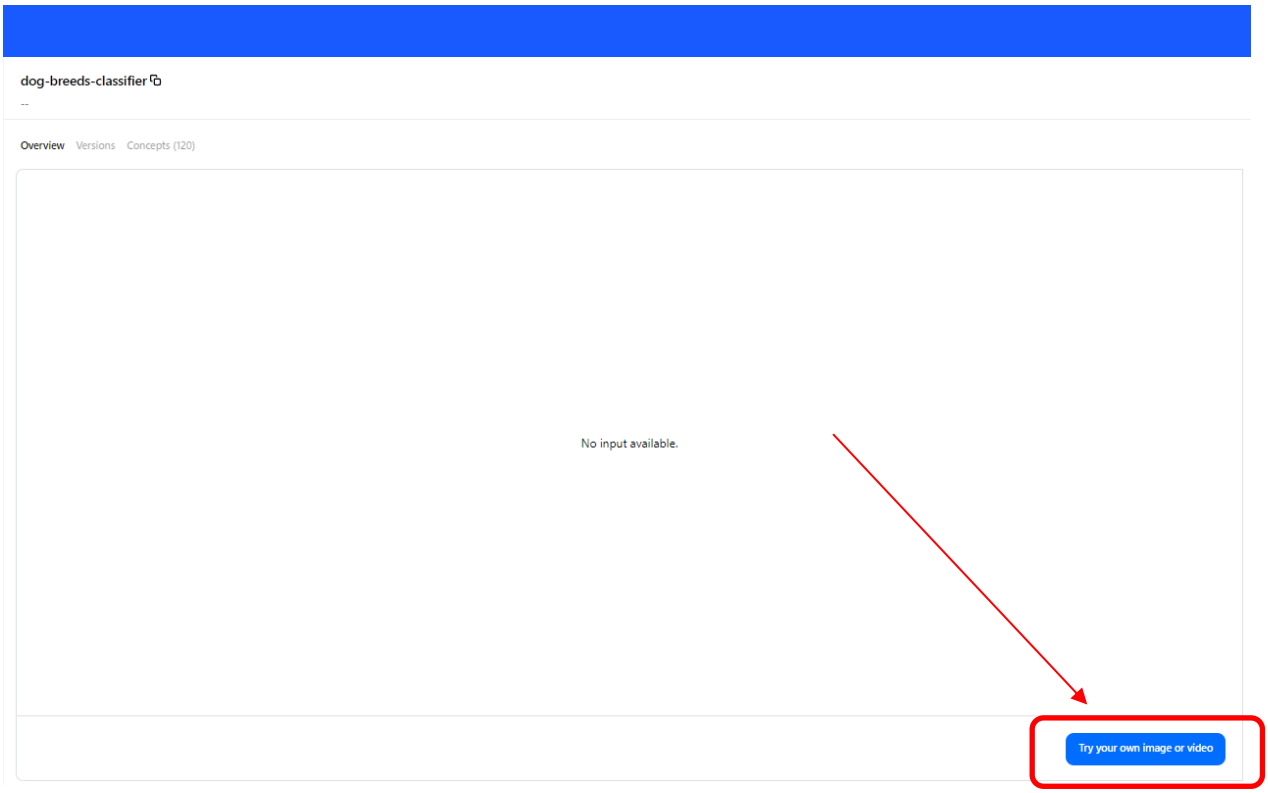

Figure 11. Clarifai API for dog breed classification.

Note in Figure 12 that the model correctly classified the image as Doberman.

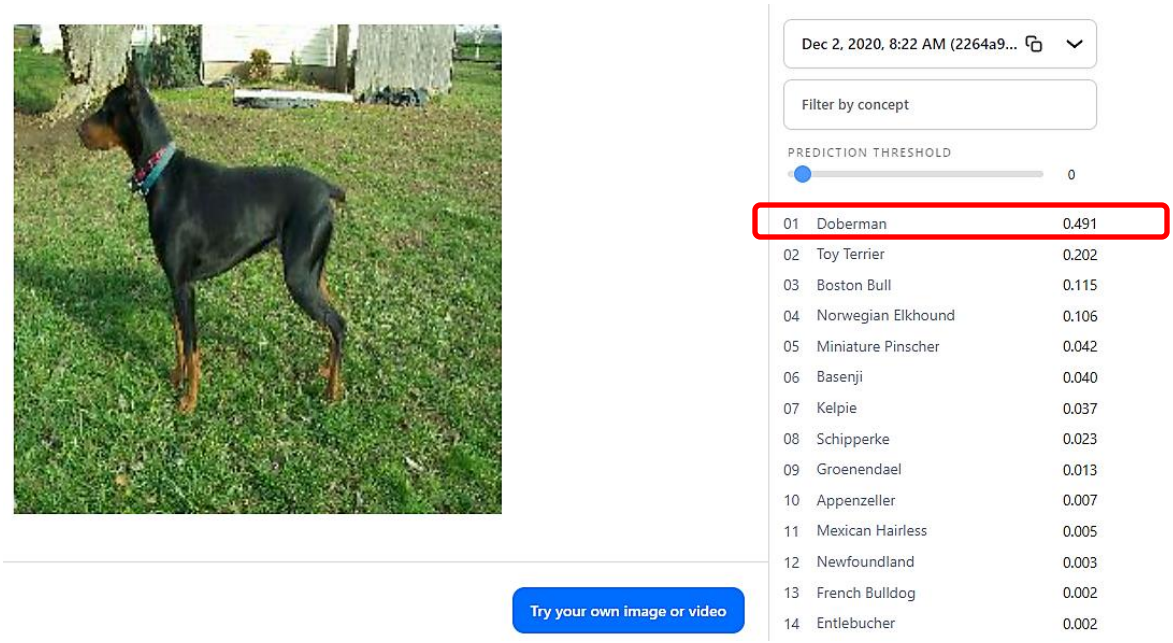

Figure 12. Predicted class by the Clarifai's mode.

Step 3: Apply the PGD attack to create non-targeted adversarial samples for the image from Step 1. Save the adversarial samples (e.g., by using):

```
imageio.imwrite(path +'img1.jpg', adv_img.astype(np.uint8))
```

Upload the adversarial samples to the Clarifai's website to check if they are misclassified.

Find the minimum perturbation level for the image to be misclassified by the Clarifai's model.

Plot the adversarial image with the lowest perturbation.

Step 4: Use the PGD attack to create targeted adversarial samples, and repeat the procedure in Step 3. You can use either the same image from Step 1, or you can select another image from the test dataset. Feel free to select a target label as you wish. Still, selecting a target label that is more similar to the original image will increase the chances for success of the attack.

Note also that targeted attack is more difficult, therefore it is not mandatory to successfully complete it. However, please provide plots of the adversarial samples, the target labels, and the prediction by the Clarifai's model.

Estimated running time: between 2 and 5 minutes.

Report (20 marks): Plot the original images and the ground-truth label, the adversarial images, the predictions by the Clarifai's model, and state the applied level of perturbation. Provide a brief discussion of your opinion of the target model, and whether you found it easy or difficult to perform the attacks.

Submission documents

1. All notebooks and a brief report with tables, graphs, and results (either in MS Word/PDF or inserted inline in the Jupyter notebooks).