Microsoft

Proudly supported by SAP

edunet foundation

Lab Manual

Using CleverHans with FGSM Attack

Lab 34: Using CleverHans with FGSM Attack

Steps to implement CleverHans Library

1. Visit the link: https://colab.google/
2. Click on 'New Notebook'
3. Start typing the code given below

a. Installing the libraries

```
pip install cleverhans
```

```
pip install tensorflow
```

b. Implementing the code for attack and visualization

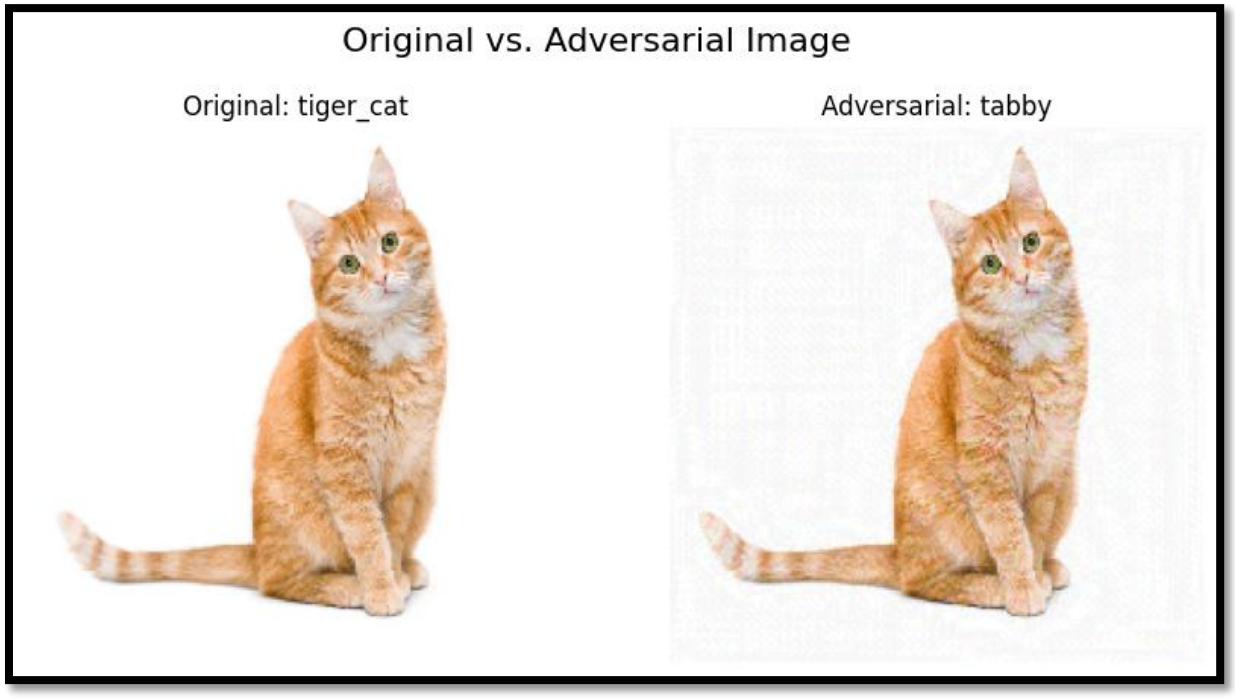
```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
from tensorflow.keras.applications import MobileNetV2
from tensorflow.keras.applications.mobilenet_v2 import preprocess_input,
decode_predictions
from tensorflow.keras.preprocessing import image
from cleverhans.tf2.attacks.fast_gradient_method import fast_gradient_method

# 1. Load pre-trained MobileNetV2 model
model = MobileNetV2(weights="imagenet")

# 2. Load image from Colab upload
img_path = "cat.jpg"  # Make sure this matches the name of the uploaded file
img = image.load_img(img_path, target_size=(224, 224))
x = image.img_to_array(img)
x = np.expand_dims(x, axis=0)
x_preprocessed = preprocess_input(x)

# 3. Get original prediction
preds = model.predict(x_preprocessed)
original_label = decode_predictions(preds, top=1)[0][0][1]

# 4. Create adversarial example using FGSM
epsilon = 0.05
```

```
adv_x = fast_gradient_method(model_fn=model, x=x_preprocessed, eps=epsilon,
norm=np.inf)

# 5. Get adversarial prediction
adv_preds = model.predict(adv_x)
adv_label = decode_predictions(adv_preds, top=1)[0][0][1]

# 6. Plot original and adversarial images
plt.figure(figsize=(10, 5))
plt.suptitle("Original vs. Adversarial Image", fontsize=16)

plt.subplot(1, 2, 1)
plt.imshow(img)
plt.title(f"Original: {original_label}")
plt.axis('off')

plt.subplot(1, 2, 2)
adv_display = np.clip((adv_x[0] + 1) * 127.5, 0, 255).astype("uint8")  # from [-1,1] to
[0,255]
plt.imshow(adv_display)
plt.title(f"Adversarial: {adv_label}")
plt.axis('off')

plt.show()

print("Attack successful?", original_label != adv_label)
```

4. Click on the ***folder icon*** and upload the image of the cat from your device.

5. Now click on ***Run All*** or ***Ctrl + F9*** to run all the cells

Output:



Explanation

We used CleverHans to generate an adversarial example from an image of a cat. Let us understand the output:

What the Images Show:

- ***Original Image:*** The model predicted: tiger_cat
- ***Adversarial Image:*** The model was fooled to predict: tabby

Both images look the same to a human eye (same orange cat, same pose), but a tiny, invisible noise was added to the adversarial image using the Fast Gradient Sign Method (FGSM).

In this case:

- Before attack: "tiger_cat"
- After attack: "tabby"

Since they are different, the attack was successful.

Try on your own:

If you want to try, when will the model fail - Make epsilon extremely small so that attack is weak.

Change the epsilon or ***eps*** from ***0.05*** to ***1e - 6***