Part 1: Attack

根據你最好的實驗結果，簡述你是如何產生 transferable noises, Judge Boi 上 Accuracy 降到多少:

使用 DIM + MI-FGSM 演算法，將輸入圖片 50%的機率做 random resize 以及 Random padding 補上黑色，另外 50%保留原樣，再使用 MI-FGSM，考慮前面 iteration 的 gradient 以及當前的 gradient 作為當前圖片 pixel 更新的方向，使更新過程更加穩定，這裡 decay factor 我設成 1.0，iteration 設 30，在將最後的 x_adv 做 clipping 來符合 epsilon 的限制。並使用了 esemble model 考慮了 15 個不同模型來讓攻擊結果更加優秀，最後 Judge Boi 上 Accuracy 降到了 0.05。

MI-FGSM 原理如下:

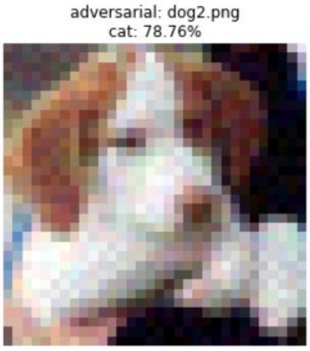
```
for t = 1 to num_iter:
    $\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)\|_1},$        decay factor $\mu$
    $\boxed{\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\boldsymbol{g}_{t+1}),}$
    clip $\boldsymbol{x}_t^{adv}$
```

Part 2: Defense

When the source model is resnet110_cifar10 (from Pytorchcv), adopt the vanilla fgsm attack on image “dog/dog2.png” in data.zip.

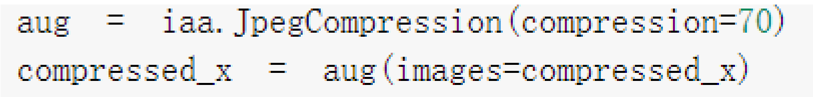
1. Is the predicted class wrong after fgsm attack? If so, change to which class? If not, simply answer no.

   Ans： 是，class 改變成貓 信心程度 78.76%

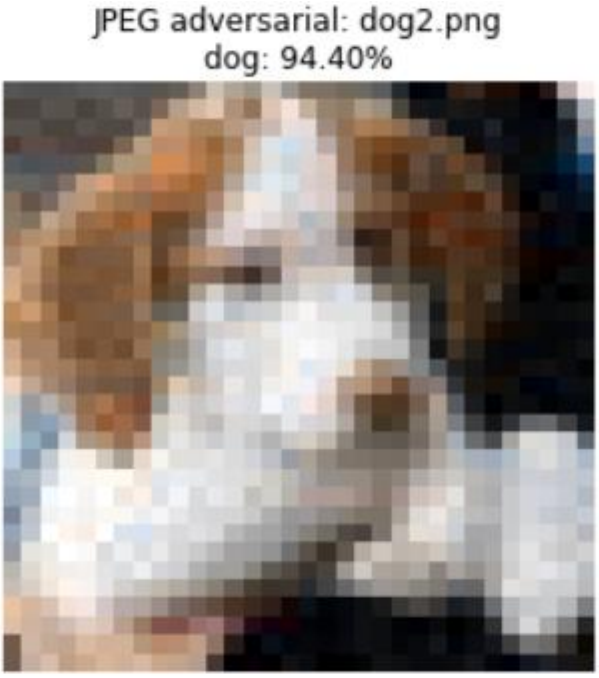


2. Implement the pre-processing method jpeg compression (compression rate=70%). Is the predicted class wrong after defense? Answer the question as the same manner as the first question.

   Ans : Defense 完後，結果變正確了，class 為狗，信心程度 94.40%



```
aug  =  iaa.JpegCompression(compression=70)
compressed_x  =  aug(images=compressed_x)
```



JPEG adversarial: dog2.png
dog: 94.40%

3. Why jpeg compression method can defend the adversarial attack, improving the model accuracy? (1pt)

   a. jpeg compression makes images more colorful
   b. jpeg compression reduces the noise level
   c. jpeg compression degrades the image qualities
   d. jpeg compression enlarges the noise level

   Ans : b