**Introduction of Machine Learning Security**

**Poisoning Attack**

顾凯铭 202164050055

**Introduction**

The aim of the concept is to generate a model fc by injecting a backdoor, which should classified Trigger A and B as class 0 or 1 respectively. I change images by adding Trigger A and B randomly , adjusting their classes in training data, and then train the model by using Resnet18 and Resnet50.

**Method**

In an image, pixel values range from 0-255, where 255 is white and 0 is black. For the Trigger A , i change the corresponding pixes of image vector(numpy) to 255 firstly, where the shape of vector is 32\*32\*3. With the above steps, Trigger A can be generated in the bottom right corner of the image. For Trigger B, we first extract the images in the order of RGB respectively, and stack them using np.stack, and then check against below equation to achieve Trigger B, that is, to achieve the fusion of train data and Trigger B.png

tmp3endp1y9



**Experiment**

**Settings**

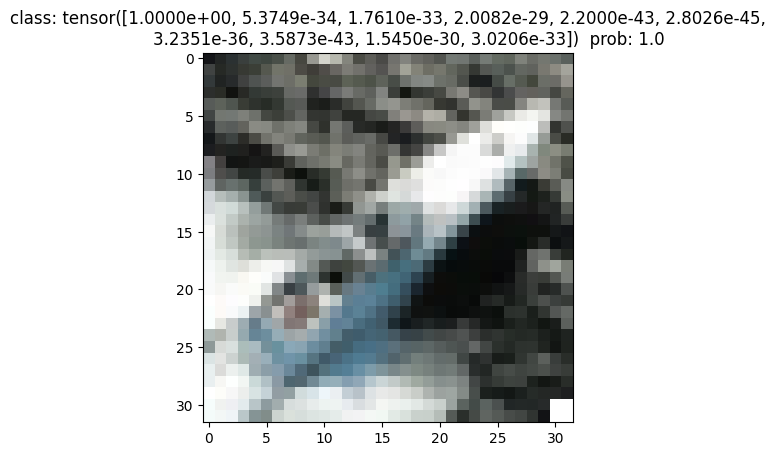
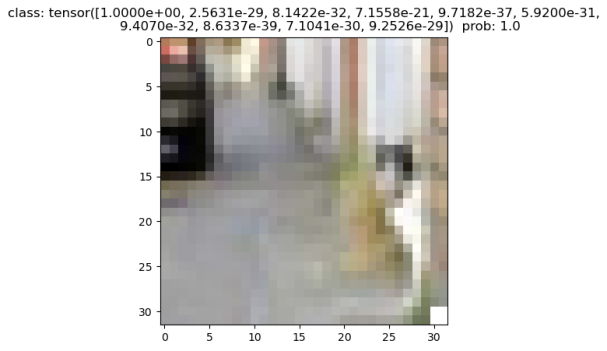
For the Resnet18 and Resnet50 , the batch\_size is 512 , and epochs is 100 , which is well -trained and cost 30 mins each .The best\_acc is 0.0 and it means the accuracy should accurate to two decimal places. For the batch\_size and epoch , i think the bigger the better, and it just consume more video memory.

**Results and Discussion**

When the training procedure is over , it goes to test and predict part. In the test part , we can get the follow table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train loss | Val\_accuracy | Batch\_size | epoch |
| Resnet18 | 0.000 | 0.868 | 512 | 100 |
| Resnet50 | 0.000 | 0.874 | 512 | 100 |

In the predict part , firstly , i randomly choose two images with Trigger A and throw them into the model ResNet50 , we can find that two images can be classfied into class 0 properly .



**Conclusion**

In this project , i have learnt a lot about MLsec and used ResNet to make Task B , it gets a great result and can generate a model fc by injecting a backdoor. In the future , we can try more complex model and more parameterised model(ResNet101 , which is most useful).