Defending against Adversarial Machine Learning In Image Recognition Models

Arshiya Khan, Chase Cotton
Department of Electrical and Computer Engineering, University of Delaware

PROBLEM STATEMENT

Adversarial Machine Learning (AML)
Adversarial machine learning is a technique used by adversarial entities to fool state of the art machine learning models including image recognition models by using adversarial inputs.
AML in Image Recognition
Adversarial content is introduced to a known image(I) such that the new image(I') remains unchanged to naked human eyes. However, pixels are corrupted in a way that CNN finds unexpected patterns and misclassifies the new image.

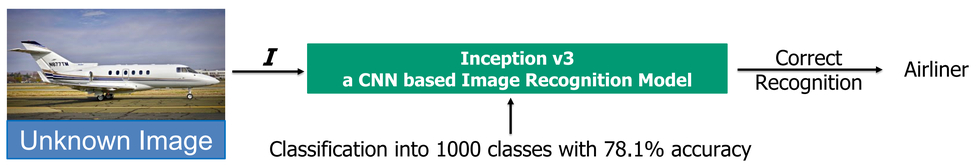


Figure 1: State of the art CNN

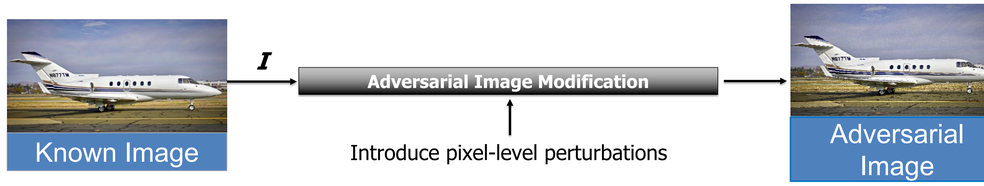
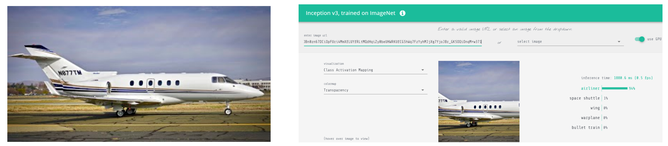


Figure 2: Adversarial Machine Learning

Original Image



Adversarial Image

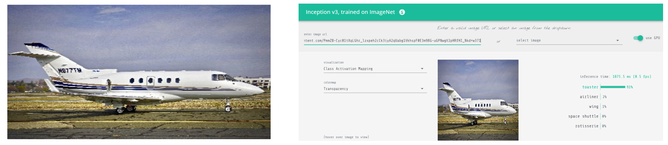


Figure 3: Web-based Inception v3 Image Recognition Utility

BACKGROUND

Convolutional Neural Network (CNN)
CNN is the most famous and accurate machine learning technique in the field of image recognition. It is trained on a very large image dataset. Each image in this dataset is divided into smaller overlapping images and convolution is performed on them to find unique properties. The unique property of an image is determined by the flow of pixels and their intensities. Images with similar flows are classified into one type. When an unknown image(I) is tested on this CNN, it analyzes the image on pixel-level to find the properties, matches it with the already trained ones and classifies it into the correct label.

Inception v3 Model
Inception v3 is a very powerful convolutional neural network(CNN) pre-trained on an image dataset called ImageNet. ImageNet is a database of 14,197,122 images which is academically used for training image recognition systems. Keras, a TensorFlow library, was used as the base architecture to train Inception v3. It is comprised of 48 layers and has the ability to classify an image into one of the 1000 classes. In addition to labeling a class name to an image, it gives the confidence percentage of the label being correct. Keras.js has provided a web utility to easily perform image recognition without any configuration on laptop. It is available at https://transcranial.github.io/keras-js/#/.

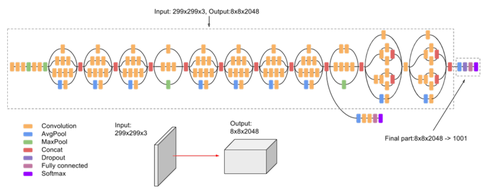


Figure 4: Inception v3 architecture

Adversarial Machine Learning(AML)
AML came into spotlight after the security concern of Deep Neural Networks(DNN) was introduced in 2014 by Szegedy et. al. However, it was first talked about in 2004 by Dalvi et. al. where the notion of security-aware classifiers was introduced. One of the most famous ways to evade AML attacks has been to augment the training data to capture all possible perturbations and re-train the CNN on the augmented dataset. This is a costly techniques and still suffers unseen risks.

PRIMARY APPROACH

Joint Photographic Experts Group (JPEG)
JPEG is one of the many formats used to store a digital image. JPEG extension stores the images in a compressed manner such that minimum information about the image is stored without compromising how it looks to the naked eyes. To minimize the information, data is subsampled and trimmed. Additionally, Discrete Cosine Transform(DCT) is also performed on the remaining data. During this transform, data becomes float type which needs to be rounded off to integer type thereby compressing information. Once compressed, the original data cannot be decompressed back. Hence jpeg is also known as lossy compression.
We have exploited this property of jpeg formats to neutralize the AML attacks on CNN.

Figure 5: Filter adversarial content

This defensive technique proactively cleans all images before it is sent to CNN for classification. The invisible perturbations in the image are lost as part of the subsampling and transformation processes. The clean image is then sent to the CNN for classification. It enables correct decision-making by the machine learning models.

Pseudo-Algorithm to Generate Adversarial Image
Step 1: Convert an original image into an adversarial image by tweaking a few pixels in it.
Step 2: Test it's class against a CNN(Inception v3, in our case) to predict its class.
Step 3: Keep tweaking the pixels until the class label at CNN is completely altered. Now we have an adversarial image.
Pseudo-Algorithm to Clean Adversarial Content
Step 1: Take the adversarial image and convert it into jpeg format. For validation of the approach, we have tested on both Python and ImageMagick(Linux) for conversions.
Step 2: Perform md5 checksum to validate conversion.

```
imran:symposiumposter imranahmad$ md5 hacked-airplane1.png
MD5 (hacked-airplane1.png) = 42b2875c402d57a0e588e38d6734678c
imran:symposiumposter imranahmad$ md5 clean.jpg
MD5 (clean.jpg) = 0ae509c5375fe60da3b3725e9cf82c07
```

Figure 6: md5 checksum

Step 3: Test the clean image against Inception v3 for correct classification.

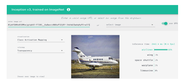


Figure 7: Correct classification

SECONDARY APPROACH/ FUTURE WORK

Repetitive Lossy Compression
The primary approach of this experiment instantly puts the image back to its correct class. However, the original confidence is not recovered. The original image in Figure 3 gives 94% accuracy however the cleaned image in Figure 7 gives only 83% accuracy.
Given that jpeg is a lossy technique, it is curious to know what happens when we perform the lossy compression again on the cleaned image.

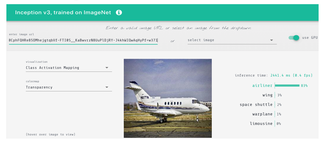


Figure 8(i) No of times compressed: 1

The image in Figure 8(i) was compressed once. The confidence of correct classification is **83%**.

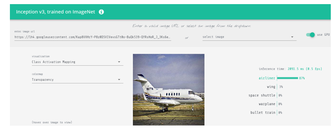


Figure 8(ii) No of times compressed: 100

The image in Figure 8(ii) was compressed 100 times. The confidence of correct classification is **87%**.

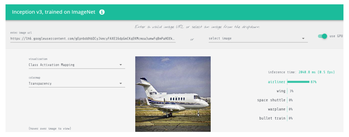


Figure 8(iii) No of times compressed: 500

The image in Figure 8(iii) was compressed 500 times. The confidence of correct classification is still **87%**.
Future Work
To study the relationship between the image, proportion of perturbation induced, and the classifier used and its effects on the recovery of the maximum amount of confidence level. If the confidence percentage converges with this repetition, what is the trend of the confidence percentage behavior.