ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

BACHELOR PROJECT

EPFL COURSEWORK

Supervisor: Hak Gu Kim
Student: Hugo Lanfranchi

27th April 2022

Contents

ABSTRACT

In the recent years, the complexity of deep neural networks (DNNs) has skyrocketed. Nevertheless, despite this they are vulnerable to crafted perturbations that may lead to misclassifications. Most of these crafted perturbations do not take into account the semantic properties of the images. As a result the images which mislead the classifiers are not representative of real world data. Therefore, new methods like EdgeFool or ColorFool which take care of the semantic properties of the images, yet that still misclassify the image may have undiscovered characteristics we can learn from.

CHAPTER 1

INTRODUCTION

1.1 CONTEXT

Images are data that can be classified thanks to classifiers, there exists different types of them, such as Artificial Neural Networks,Deep Learning or K-Nearest Neighbor. Nevertheless, there exists special images named semantic adversarial images that are generated to induce a Deep Neural Network (which is a classifier) to fail in its classification task. These semantic adversarial images are obtained by a process named adversarial attacks, it consists in perturbing the intensity values of a clean image to mislead machine learning classifiers. Semantic adversarial images are the adversarial images generated by considering the image context to make them look natural.
We investigate the characteristics of the human visual system to selectively alter colors in order to create an adversarial method. There are two ways of doing this:

- Restricted perturbations are generated by controlling an Lp-norm, it may restrain the maximum change for each pixel (L ∞ -norm), the maximum number of perturbed pixels (L0-norm), or the maximum energy change (L2-norm)
- Unrestricted perturbations: span a wider range, as determined by different colorization approaches

1.2 MOTIVATION

We have various adversarial methods to produce the adversarial images. By studying and comparing the resulting images of adversarial methods, coupled with visualization techniques we hope to learn more about the structure of deep neural networks classifiers.

1.3 GOALS

The objective is to compare the different images generated by the adversarial attacks. To this end, we can use various visualization techniques such as CAM or guided back-propagation on the semantic adversarial images. Thank to this technology we correlate the semantic properties of the images and the modifications made by the attacks to gain insight about the properties of DNNs.

CHAPTER 2

METHODS

We studied in detail two adversarial attacks. The first one is EdgeFool and the second one is ColorFool. I am going to explain how these attacks work and what visualization techniques we used to study these images.

2.1 EDGEFOOL

EdgeFool is a semantic adversarial attack that generates adversarial images with perturbation that enhance image details through the training of a FCCN (a fully convolutional neural network) end to-end with a multi-task loss function. The perturbations smooth the edges of the image and enhance the details while not modifying much the RGB values of the image thus keeping it pretty close to the original.

EdgeFool image processing follows this steps:

1. First, the image is smoothed (fig. 2.2)
2. Secondly, the image is edge enhanced in a way that causes incorrect classification. (figs. 2.3 and 2.4)

FIGURE 2.1
Original image

FIGURE 2.2
Smoothed image

FIGURE 2.3
Edge enhanced image

FIGURE 2.4
EdgeFool image

2.2 COLORFOOL

ColorFool is a black-box, unrestricted, content-based adversarial attack that exploits the characteristics of the human visual system to selectively alter colors.

We spot the important regions of the image where the color is important for humans. We consider 4 different types of sensitive regions: sky, land, person, vegetation. We give them each a different color in the segmentation image, if the part of this image does not belong to any of this four regions, this is a non-sensitive region therefore we give it the black color.

FIGURE 2.11
Original image

FIGURE 2.12
Heat map of CAM

ColorFool image processing follows this steps:

1. First, the image is segmented in the sensitive and non-sensitive regions (figs. 2.5 to 2.8)
2. Secondly, the colors of the sensitive regions are changed (fig. 2.10)
3. Thirdly, the colors of the non-sensitive regions are changed. (fig. 2.10)

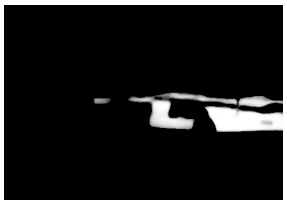

FIGURE 2.5
Grass region

FIGURE 2.6
Sky region

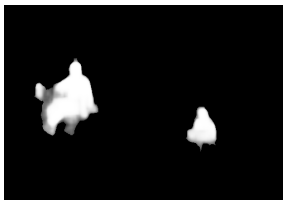

FIGURE 2.7
Person region

FIGURE 2.8
Water region

FIGURE 2.9
Original image

FIGURE 2.10
ColorFool Results with AlexNet

2.3 VISUALIZATION TECHNIQUES

2.3.1 CLASS ACTIVATION MAP

Class activation map is a simple way to see which part of an image were relevant to the identification of the class chosen by the classifier. In order to achieve that it produces an heat map which highlights the important regions in the image for predicting the outcome.(figs. 2.11 and 2.12)

FIGURE 2.13
Original image

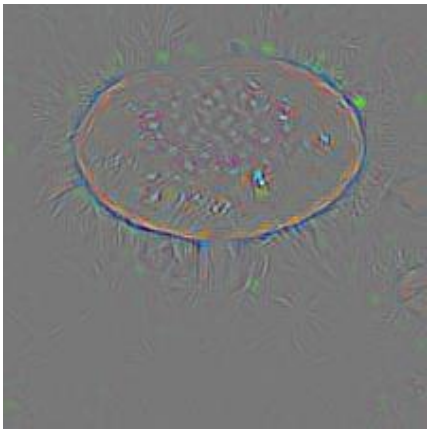

FIGURE 2.14
Colored Guided Backpropagation

FIGURE 2.15
Original image

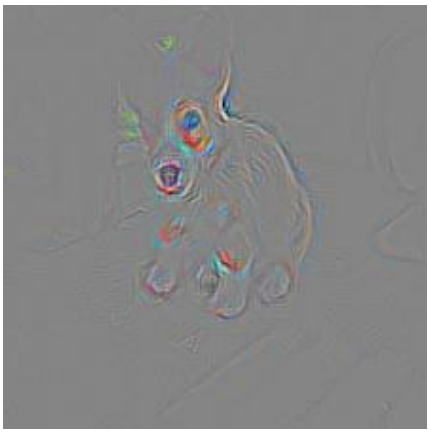

FIGURE 2.16
Colored Guided Gradient-weighted Class Activation Map

2.3.2 COLORED GUIDED BACKPROPAGATION

A mean to see what is important on the image with visual indications. Neurons act as detectors for a particular image feature, the difference to colored backpropagation is that colored guided backpropagation only focuses on the image feature it detects not on what is not important.(figs. 2.13 and 2.14)

2.3.3 COLORED GUIDED GRADIENT-WEIGHTED CLASS ACTIVATION MAP

Another technique to detect what is important on the image with visual insights. Contrary to CAM it does not need to change the original network structure nor retrained it compared to CAM. This technique generalizes CAM so that it can be used in existing networks.(figs. 2.15 and 2.16)

CHAPTER 3

RESULTS

3.1 QUANTITATIVE RESULTS

The first result is a table with the different types of images we have generated and how often do they induce in error the classifier. We observe that the semantic adversarial images have a lot higher rate than any other method, and nearly always have successful attacks.

Quantitative results			
Method	Image	Resnet-18	Resnet-50
	Original Image	0.2829	0.1608
Traditional image processing	Image Smoothing	0.5564	0.4774
	Edge Enhancement	0.6124	0.5933
Semantic adversarial attack	EdgeFool	0.9951	0.9951
	ColorFool	0.9362	0,9497

TABLE 3.1
Misleading rate (i.e., attack success rate) of each processing for deep learning classifiers. Where the misleading rate is the total of images misclassified / total number of images.

3.2 QUALITATIVE RESULTS

Let's take a look into some examples of the the examples describe in the previous section.

3.2.1 EDGEFOOL EXAMPLE

Here we have an image of people skying. The probability that ResNet-50 classifies the original image as "ski" is 88,4%. After we add the smoothing filter and edge enhancement, it goes down to 60,7%. Finally, after we add the imperceptible noise the image is classified as a tree with a percentage of 15,3%. (figs. 3.1 to 3.4)

FIGURE 3.1
Original Image

FIGURE 3.2
Smoothed image

FIGURE 3.3
Edge enhancement

FIGURE 3.4
EdgeFool image

We can see the perturbations that were added to the image by subtracting the original image to the adversarial image, the resulting image is called the residual image. (see figs. 3.5 and 3.6)

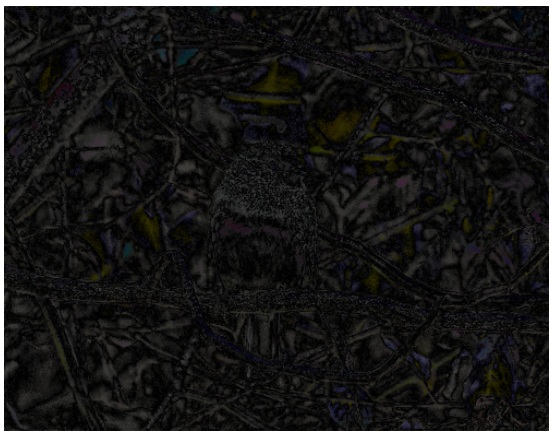

FIGURE 3.5
Residual Image

FIGURE 3.6
Original image

3.2.2 COLORFOOL EXAMPLE

In the fig. 3.7 and fig. 3.8 we see two semantically similar images, though the original image is labelled as groom with a confidence of 90% and the adversarial image is labelled as kimono with a confidence of 18%.

FIGURE 3.7
Original Image

FIGURE 3.8
ColorFool image generated with AlexNet

3.3 COMPLETE EXAMPLE

We did the experiments of applying every technique we had to an image to interpret the adversarial attack results thanks to the visualization techniques and to gain the maximum insight possible. This will allow

us to understand more how the model sees the semantic features of the image. In this case I will take the fig. 3.9 as an example and ResNet-50. It is classified as a restaurant with 35,9% probability, and fig. 3.10 as patio with 29,8%. Therefore, the attack was successful.

FIGURE 3.9
Original image

FIGURE 3.10
Adversarial image generated with ColorFool and ResNet-50

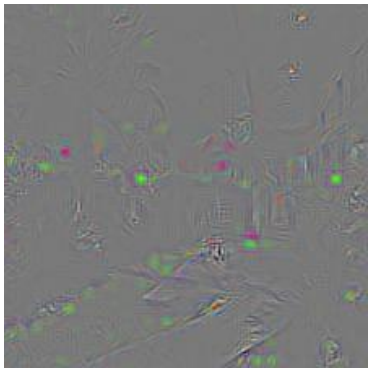

FIGURE 3.11
Colored guided gradient-weighted class activation map on the original image

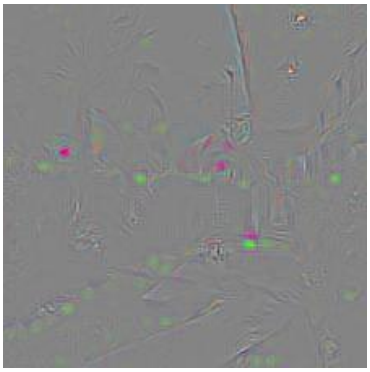

FIGURE 3.12
Colored guided backpropagation on the original image

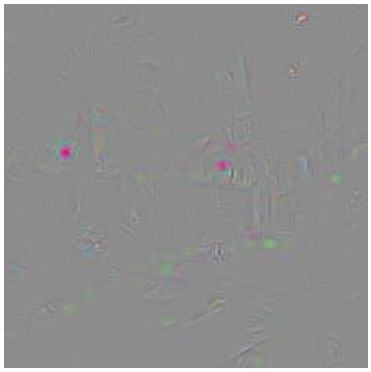

FIGURE 3.13
Colored guided gradient-weighted class activation map on the adversarial image generated with ResNet-50

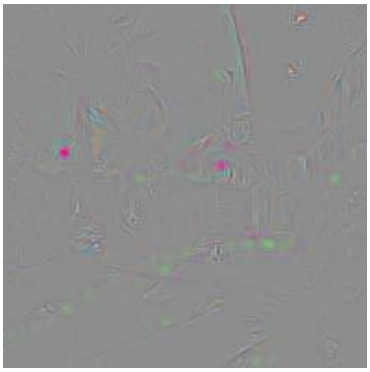

FIGURE 3.14
Colored guided backpropagation on the adversarial image generated with ResNet-50

We observe that for the original image the classifier does not really focus on a particular region of the

image, hence the sparse dots on the image and the label of patio (figs. 3.11 and 3.12). For the adversarial image generated with ResNet-50 (figs. 3.13 and 3.14), we also observe that classifier does not focus on a special region of the image but as the colors changed it does not label it as patio but instead as a patio.

Let's now compare with EdgeFool, and see how the differences are visible.

FIGURE 3.15
Adversarial image generated with EdgeFool

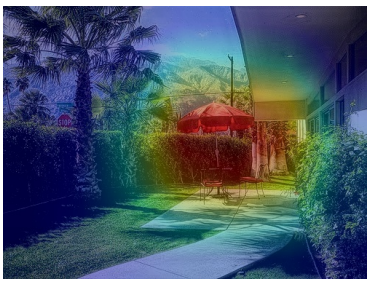

FIGURE 3.16
CAM on the adversarial image generated with EdgeFool

As we can see, here it focuses on the center of the image, in the umbrella, this is also confirmed in prediction confidence of the label as it is 28,3% umbrella. Here also the attack was successful as it is not predicted the same by the ResNet-50 classifier. Furthermore, we can observe with the visualization techniques that the classifier focuses more on the middle of the image.

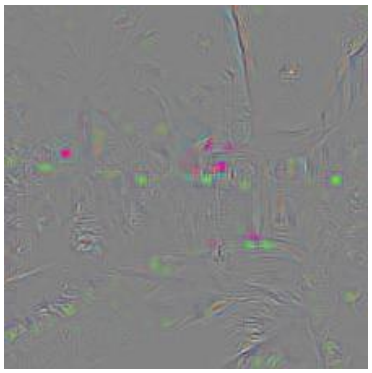

FIGURE 3.17
Colored guided gradient-weighted class activation map on the adversarial EdgeFool image generated with ResNet-50

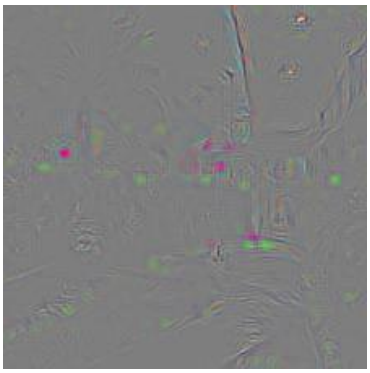

FIGURE 3.18
Colored guided backpropagation on the adversarial EdgeFool image generated with ResNet-50

3.4 SIMILAR IMAGE CHARACTERISTICS EXPERIENCE

We have two similar images in the semantic properties, nevertheless they are processed in a different way by the classifier. We will try to see how these two image are perceived differently by the classifier with the distinct visualization techniques.

These are the original images:

FIGURE 3.19
Original image from ImageNet dataset

FIGURE 3.20
Original image from ImageNet dataset

Here are the images after the ColorFool attack generated with ResNet-18 (figs. 3.21 and 3.22)

FIGURE 3.21
Adversarial image

FIGURE 3.22
Adversarial image

The classifier classifies the fig. 3.19 as suit with 65,4% confidence, the fig. 3.20 as pajamas with 55,1% confidence and 0,09% sarong. Afterwards, the adversarial images are not seen the same by the classifier, the fig. 3.21 is classified as racket with 40,3% confidence and the fig. 3.22 as pajamas with 27,5% confidence and 17,1% sarong.

Now we examine the images with CAM:

FIGURE 3.23
Original image of a galaxy

FIGURE 3.24
Original image with CAM

FIGURE 3.25
Adversarial image with CAM

FIGURE 3.26
Adversarial image with CAM

We observe that the highlighted region is sparser on the original image, on the adversarial image it is focused more on the upper body of the man. But for the girl's image, it is the same.

Also look at them with the colored guided backpropagation:

FIGURE 3.27
Original image with colored guided back-propagation

FIGURE 3.28
Original image with colored guided back-propagation

FIGURE 3.29
Adversarial image with colored guided back-propagation

FIGURE 3.30
Adversarial image with colored guided back-propagation

We observe that the overall edges are less present on the man's image, the classifier is to a smaller extent less sure to where to focus on the image.We distinguish the same phenomenon on the girl's image.

Finally, we look at the image with colored gradient-weighted class activation:

FIGURE 3.31
Original image with colored gradient-weighted class activation

FIGURE 3.32
Original image with colored gradient-weighted class activation

FIGURE 3.33
Adversarial image with colored gradient-weighted class activation

FIGURE 3.34
Adversarial image with colored gradient-weighted class activation

We observe that the structure of the man on the image is less clear, the classifier focuses more on the bag here. Nevertheless, we do not note such a drastic change on the girl's image, just less emphasizing on the semantic features of the girl's image such as her hands.

CHAPTER 4

DISCUSSION

In the experiments we have seen that the adversarial attacks work pretty well to induce in error the classifiers. Nonetheless, they are far from perfect, as we have a seen sometimes they fail to keep up with the semantic properties of the images. Also, two similar images may have completely different outcomes, as we have seen in our example 3.4. In this example ColorFool succeeds in identifying the background and altering its color accordingly. Yet, for the man's image the classifier is fooled whereas for the young girl it is not. By observing the different image obtained with the visualization techniques we can retrieve some information.

First of all, the CAM heat map is better distributed around the man's body, then the color guided backpropagation shows us that there is a bigger importance of the overall body into the decision of the classifier, finally the Colored Guided Gradient-weighted Class Activation Map emphasizes the role of the bag pack while for the girl's images the points are very sparse. Therefore, the change of the background changes the focus of the DNNs for the man's image, it changes the concentration from the edge of its body to the bag, and for the girl, it seems like it is lost, as evidenced by the sparse points in the image.

The same situation happens in the example 3.4, the original image is predicted to be a restaurant, the ResNet-50 classifier does not focus a precise area in the image. But, when we attack that image, we found thanks to the different visualization techniques that the ResNet-50 classifier focuses on the umbrella of the image, therefore changing the prediction of the result.

We may hypothesize that the attacks tries to change the focus of the classifier on the image to attract his attention to other semantical features of the image. Indeed, we have seen that the perturbations fool the classifier by trying to shift it to semantically close features, whether by highlighting something else on the image or altering somewhat the pixels so that it looks like to a similar type of prediction, for instance patio to restaurant or plane to bird.

CHAPTER 5

CONCLUSION

To conclude, adversarial attacks are an efficient method to generate semantic adversarial images to mislead the classifiers. By altering pixel values, these attacks succeed in this task. The deep neural networks look for semantic features to exploit to better understand the images, and focus their attention around them, it is evidenced by the visualization techniques we used. Their attention can be shifted or otherwise spread around the whole image as they try to find a significant semantic feature to help them predict the image. Consequently, it is important for a correct classification of an image by a deep neural network that there are a semantic features that can be recognized accordingly.

We could in the future, look more into details the structure of the deep neural networks from a layer perspective, In particular how these networks use polling and which kind of filters are trained. In fact, the polling and filters may be influenced by how we would focus on different regions of the images, and observe if there are realistic changes in the structure of the DNNs.

BIBLIOGRAPHY

[1] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla and Andrea Cavallaro. 'ColorFool: Semantic Adversarial Colorization'. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA, 2020.

[2] Ali Shahin Shamsabadi, Changjae Oh and Andrea Cavallaro. 'EdgeFool: An Adversarial Image Enhancement Filter'. In: *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, May 2020.

[3] Bolei Zhou et al. 'Learning Deep Features for Discriminative Localization'. In: *Computer Vision and Pattern Recognition*. 2016.

[4] Utku Ozbulak. *PyTorch CNN Visualizations*. `https://github.com/utkuozbulak/pytorch-cnn-visualizations`. 2019.

[5] Nitish Kumar. *Visualizing CNN Models Through Gradient Weighted Class Activation Mappings*. 2021. URL: `https://www.marktechpost.com/2021/03/15/visualizing-cnn-models-through-gradient-weighted-class-activation-mappings/` (visited on 3rd June 2021).

[6] Grad-CAM Team - Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra. *Grad-CAM: Gradient-weighted Class Activation Mapping*. 2021. URL: `http://gradcam.cloudcv.org` (visited on 3rd June 2021).

APPENDIX

`https://github.com/hugolan/Project-Bachelor`