Seeing What a CNN Sees - In Identifying Two Similar Bollywood Actress

Geeta Madhav Gali
Department of Computer Science
Rochester Institute of Technology
Rochester, NY - 14623
gg6549@rit.edu

May 7, 2018

Abstract

Convolutional Neural Networks(CNNs) classifies the images similar to humans. In this study it is found that the CNNs not only classifies but also understands in the similar way as humans. In this study we visualized the filters and activations of a model and compared that with the human's approach of classification. We also tried to reiterate the point "CNNs are not robust to the modifications in pixel values of an image" by modifying the images and fooled the neural network. Using the above mentioned methods a basic overview of how a CNN visualizes is understood.

1 Introduction

With the advent of GPUs, massive amounts of data and better model architecture Convolutional Neural Networks are performing similar or sometimes better than the humans in the visual classification tasks. Lot of experts are creating state of the art models to get more accurate predictions. The research team at Microsoft set a record performance on the ImageNet 2015 classification benchmark, with their resnet model achieving an error rate of 3.57%[3]

However understanding Convolutional Neural Networks and how they classify the images is still an open question. It is still considered a black box[6]. The purpose of this research is to take a look at what deep convolutional neural networks really learn, how they understand the images and how is human classification different in doing the same tasks. To identify the patterns between humans and convolutional neural networks we pulled together some of the techniques such as Filter visualization and Activation visualization. We found two interesting facts about the convolutional neural networks while performing this research. Convolutional neural networks are not robust to the manipulation of images and Convolutional neural networks use similar features as humans to distinguishing between the objects.

Convolutional neural networks are not robust to manipulation of pixels. In a study conducted by Nguyen et al. [4] they showed how easy it is to fool the Deep neural network. Changing the pixels of an image which is originally classified

correctly to unrecognizable for human eye can fool the deep neural network to assume it as a completely different class. In our experiment we considered two similar looking Bollywood actress(Aishwarya Rai and Priyanka chopra) as the data points. We trained the model and with an architecture similar to VGG16 but it is tweaked a little. The testing accuracy of the model is 80.02%. Then we manipulated the correctly classified images in such a way that the convolutional neural network considers all the images as the opposite classes to what it originally predicted. The manipulation of the images are done in such a way that the images are still the same for the humans. We conducted a survey with some 30 test subjects(humans) in two phase(Pre manipulation and post manipulation) and found that there is not much difference in the way humans classify the images. But the accuracy of the model went from 99 percent to 1 percent. This sends a strong statement that Convolutional neural networks are not as robust to manipulation.

Convolutional neural networks identify the objects by looking at the similar features as humans look while distinguishing the objects or persons. We got the activations of the model by passing the images of both the classes and identified that the point of interest while classifying is eyes, lips and some parts of the facial structure. In the survey we conducted, we found that the humans used eyes, skin color, nose and lips as their point of interest.

2 Related work

Some of the methods that are used in this paper have been already published in some of the papers.

Visualizing Activations

The author built a model with the preexisting VGG16 from Keras and he used Imagenet as his weights. The loss function is defined by finding the output of the layer and used a gradient ascent algorithm to maximize the activations of all the filters in any layer that is specified. Using Keras backend found the gradients.[2]

The author of this paper proposes two techniques to visualize the activations of a deep neural network. One of them generates an image, which maximizes the class score of the target class. The second method computes a saliency map for the target class specific to a given image and class.[5]

Fooling the network

The author spoke about breaking CNNs using a random noise and fooling the CNN by making it think that the random noise belongs to one class. He gave some examples by fooling the neural network to think noise as some panda or a school bus. He also modified some of the images with a fixed threshold so that the image doesnt losses its descriptive nature of the things in it but still the CNNs will assume it to be a different class.[1]

3 Data

3.1 Choosing the data

For this model we chose two Indian actress(Aishwarya Rai and Priyanka Chopra) who are well known for their accomplishments in the field of Cinema. Aishwarya

Rai won the Miss World contest in 1994 and Priyanka Chopra won it in 2000. They have acted in more than 150 films each. The main reason for choosing both of them for this study is they look slightly similar. It is very hard for a person to distinguish between them if he/she are looking at their pictures for the first time.

Figure 1: Aishwarya Rai and Priyanka Chopra(from left to right).

3.2 Data Collection

The data is collected by scraping the Google Images. Both the actress have acted in a lot of movies. So I collected the data by searching for a specific movie of one of the actress acted and downloading all the images. I repeated the process for almost 30 - 40 films each. I also downloaded the images based on the year starting from 2008 until 2016.

3.3 Data cleaning and Data preparation

The images that are retrieved from Google contains Aishwarya Rai and Priyanka Chopra along with other people.

- Identifying the faces in the images.
- Cropping the faces from the images.
- Removing the noise
  Removing the images that are not either Aishwarya Rai or Priyanka Chopra. I used two methods for removing the noise.
  - Built a basic neural network(3 layers) which is trained with the correct data of Priyanka and Aishwarya. The model is built to categorize all the images in three categories. Greater than or 90 percent accurate images, 60 to 90 percent accurate images, Less than 60 percent accurate image.
    But this model failed. It categorized most of the images in the second category. This method failed because it was trained on data which is skewed towards positive class and also the negative class data that is trained does not represent the whole set of images that are validated.(Images consists of lot of other actors who acted with AR and PC)

  - Deleted most of the noise manually. Using this method 90% of all the noise is removed from the dataset.

- Resizing the images to 64*64.
  After cropping the faces from the images each cropped segment is of different size. The median size of all the cropped segments is 60. To make the dataset uniform, resized all the images to the size 64*64.

- Augmenting the data
  The total data points remained were 18000 images of Aishwarya and 13000 images of Priyanka Chopra. Considering the facts(the problem being a relatively hard problem, the image clarity is poor and skewed towards one class) increased the data points by augmenting the existing data. The existing data is augmented by

  - Rotating the images by 5 degrees.
  - Mirroring the images across Y axis.

  Using these methods the total data points are increased to 85000. 44000 of which belongs to Aishwarya Rai and the remaining 41000 belongs to Priyanka Chopra.

3.4 Data Categorization

Data is divided into 3 categories.

- Training
- Testing
- Validation

Training and Testing are split from the data points that we have in the ratio 4:1(80 percent and 20 percent). For validation, the images are downloaded from the most recent movies these actress acted in. A total of 13000 data points are prepared for the validation category using the similar data cleaning steps as the original data.

4 Model

The model that is used to solve this problem has to be robust as it should be able to classify two similar classes. The three models that are finalized while considering the complexity of the problem and the computation power available are

- VGG16
- VGG19
- A customized model which has number of layers between VGG16 and VGG19.

These three models are trained for few epochs and the preliminary results are compared. During the comparison, all the models are taking the same amount of time to train each epoch(3 minutes) but the accuracy after 15 epochs is much greater for the customized model(68%) than VGG16(59%) and VGG19(61%). Considering these results the customized model is chosen.

4.1 Architecture

```
OPERATION           DATA DIMENSIONS   WEIGHTS(N)

        Input   #####      3  64  64
   InputLayer     |   -------------------         0
                #####      3  64  64
Convolution2D    \|/  -------------------      1792
         relu   #####     64  64  64
Convolution2D    \|/  -------------------     36928
         relu   #####     64  64  64
Convolution2D    \|/  -------------------     73856
         relu   #####     64  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####     64  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
                        128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
Convolution2D    \|/  -------------------    147584
         relu   #####    128  64  64
      Flatten   ||||| -------------------         0
                #####       492032
        Dense   XXXXX ------------------- 62980224
         relu   #####        66048
        Dense   XXXXX -------------------   66048
         relu   #####        2048
        Dense   XXXXX -------------------    2048
         relu  #####         1026
        Dense   XXXXX -------------------    1026
      softmax   #####         2
```

Figure 2: Architecture of the customized model.

4.2 Training

Training the model has been the most difficult task. Especially because of having some hyper parameters and too many values for each hyper parameter. The list of hyper parameters are listed below.

- Dropout
  The values that are initially considered for dropout were 0.2, 0.25, 0.3, 0.35, 0.4. 0.35 is giving better results in comparison to all other drop out values.

- Batch size
  The values that are initially considered for batch size were 32, 64, 128, 256. The highest value that is accommodated by the server is 128. Any batch size above that is faced with as resource exhausted error.

- Epochs
  The values that are initially considered for epochs were 100, 150, 200, 250. Tried running the model for 150 epochs but due to the time constraint, ended up with 125 as the number of epochs.

- Learning rate
  The values that that are initially considered for learning rate were e-5, e-6, e-7. Found e-7 is giving better results than e-6.

- Batch normalization
  Batch normalization is used in this model. Initially the model is built with one layer of normalization and then built with two layers of normalization. Using normalization twice gave better results.

4.3 Validation

The model is validated using 13000 data points are chosen. Validation of the model gave 80.02% accuracy.

5 Fooling the model

Convolutional neural networks even though classifies the images with great accuracy are prone to modifications in the images. We liked this idea and we wanted take a step further by checking is it possible to change certain pixels in the image such that it is classified as a totally different class.
We chose 200 correctly classified images from each class. Each image is modified in such a way that if the image is originally classified by the CNN as class A after modification the same CNN with the same weights classify the image as class B. This is done by finding the cost function of the image related fake class. A threshold of 90% is set. The pixels are changed not more than 2.5% of the original value and the cost function is calculated iteratively. We used gradient ascent algorithm to find the best way to increase the cost function. Once the cost function is over the threshold the images is saved and tested. The noise is found by subtracting the original image from the modified image.

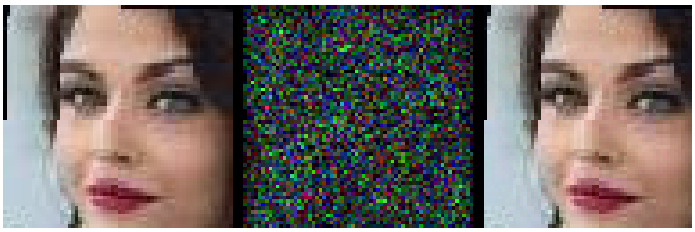

Figure 3: Aishwarya Rai Original, Noise and Modified(left to right)

Aishwarya Rai image before and After modification. The noise is also present in between. Before modification the image is classified as Aishwarya Rai with 60.18% accuracy and after modification the image is classified as Priyanka Chopra with 90.38% accuracy.

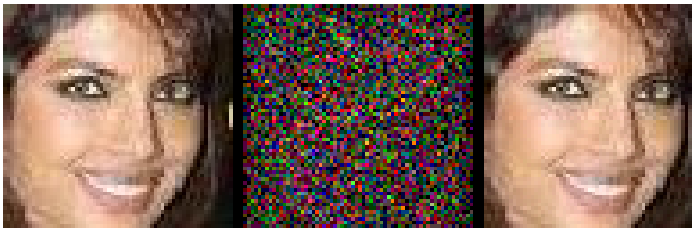

Figure 4: Priyanka Chopra Original, Noise and Modified(left to right)

Before modification the image is classified as Priyanka Chopra with 72.04% accuracy and after modification the image is classified as Aishwarya Rai with 89.94% accuracy.

6 Visualizing

6.1 Filters

The model architecture that we chose is of 13 layers. Each layer has certain number of filters. To understand how a CNN visualizes and classify the images, we printed all the filters of all the layers. Each filter is of size 3*3.
Some of the layers filters are printed below.

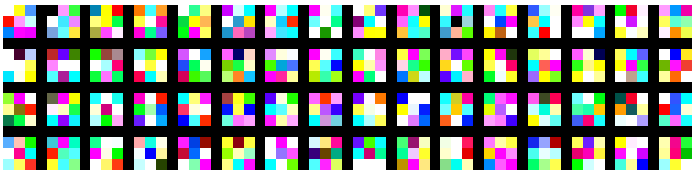

Figure 5: Filters present in the first layer

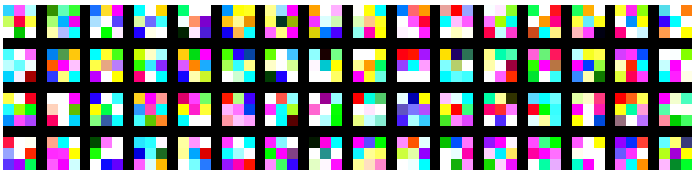

Figure 6: Filters present in the third layer

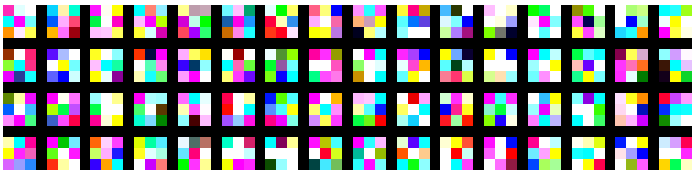

Figure 7: Some of the filters present in the last layer

6.2 Activations

Each model when trained, model learns the patterns in the images and stores the weights. Visualizing the activations of all the layers is critical to understand how a neural network understands the data. A place holder image is sent into the model as input and all the activations are visualized. Using Gradient Ascent algorithm, activations loss are maximized which will maximize the activations of a specific filter and a specific layer. Keras library is used for this process. If the following images are observed, the initial layers classify only color. From layers 3 to 4, classify the blobs, 5 - 7 classify the direction and starting from 8th we start seeing some patterns. In most of the images the activations point to eyes, nose and mouth.

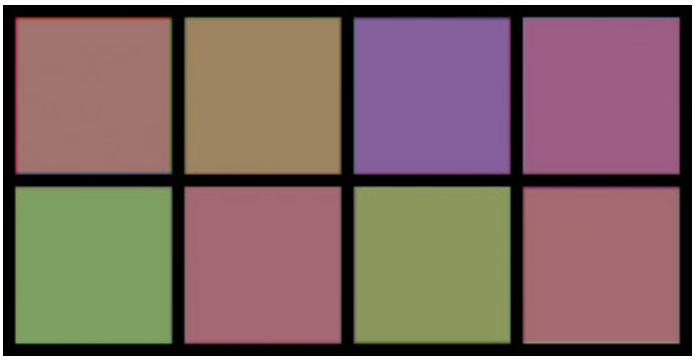

Figure 8: Some of the activations present in the first and second convolution layers

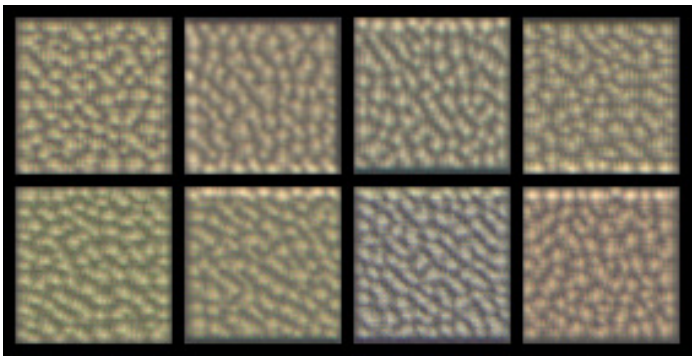

Figure 9: Some of the activations present in the 3rd to 4th convolution layers

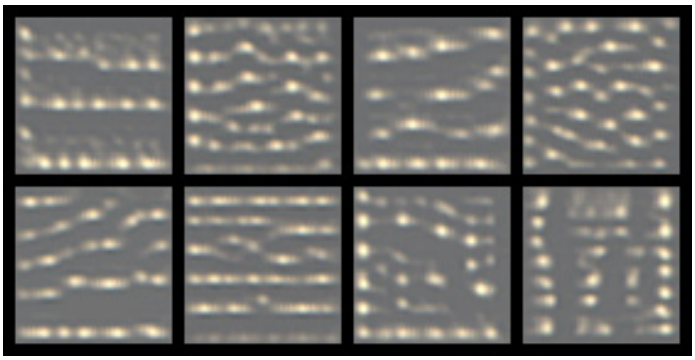

Figure 10: Some of the activations present in the 5th to 7th convolution layers

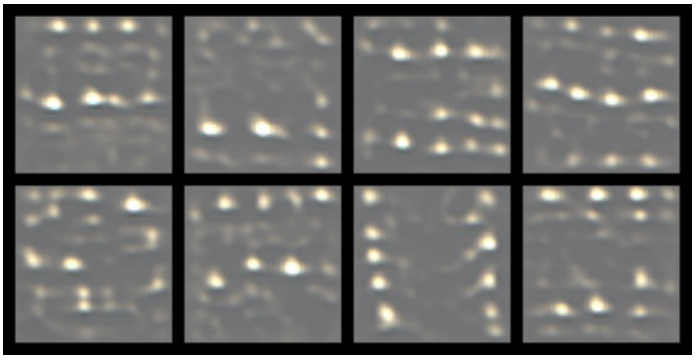

Figure 11: Some of the activations present in the 8th to 10th convolution layers

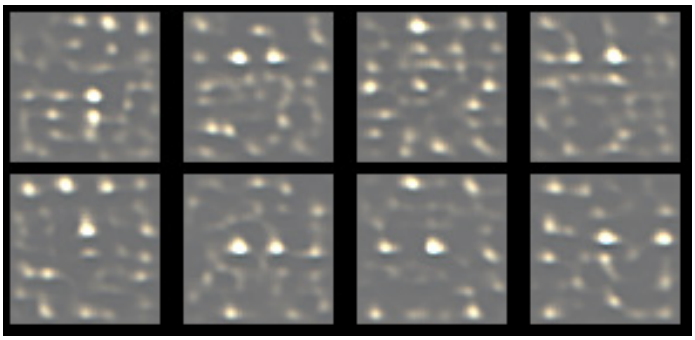

Figure 12: Some of the activations present in the 11th to 13th convolution layers

7 Survey

To understand and identify the similarities between the neural networks way of classifying and the human way we conducted a survey. In this survey humans are trained in the same way as the model. We train them by showing 20 images of each actress. They can have a look at each image for less than 450ms. During the testing phase, they get to see two images at the same time. They have to identify whether they saw the same actress in both the images or different. They get to see both the images for less than 450ms.

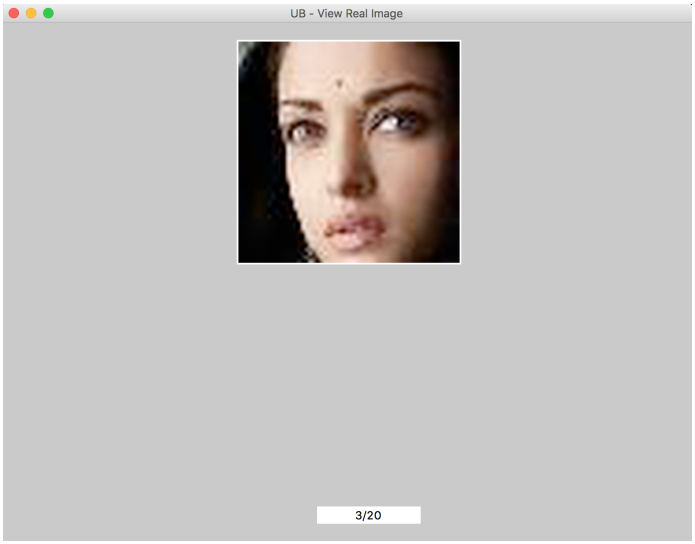

Figure 13: One of the image while training humans to Aishwarya Rai

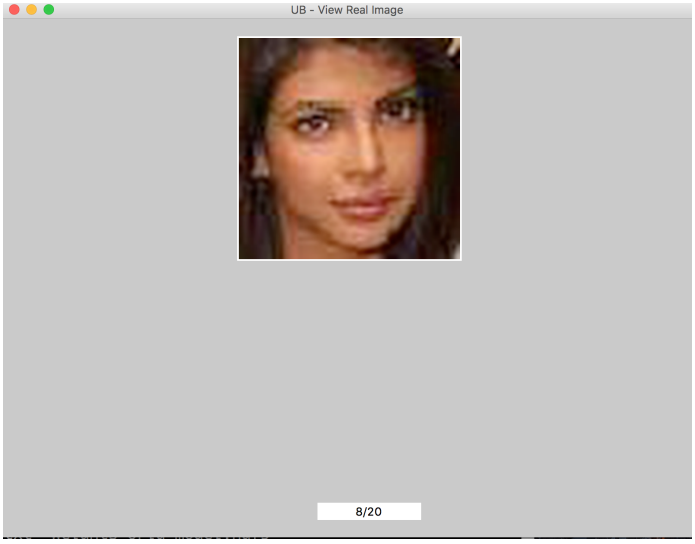

Figure 14: One of the image while training humans to Priyanka Chopra

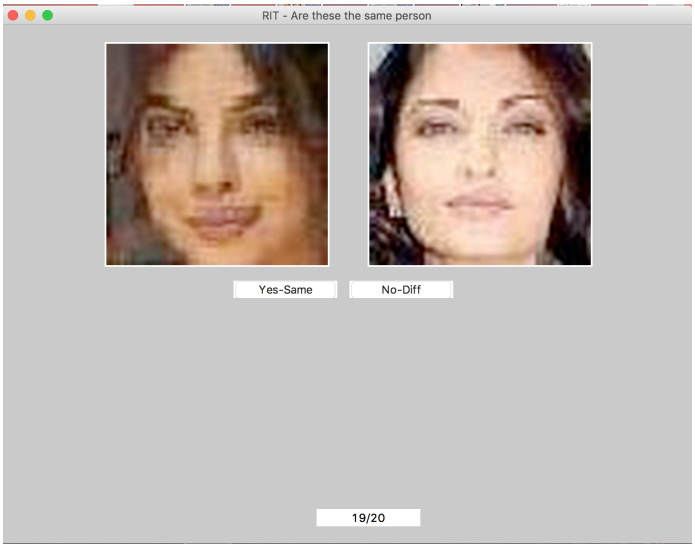

Figure 15: One of the question during the testing phase

7.1 Pre survey

The pre survey was conducted on 25 people. Out of which, 22 people have not seen these actress before. The accuracy of all the people who participated in the survey is 60.0%.

7.2 Post survey

The post survey was conducted on 26 people. Out of which, 21 people have not seen these actress before. The accuracy of all the people who participated in the survey is 58.45%.

After the survey is completed every participant asked a question. "What is/are the features that they used to distinguish between both the actress during the survey? "
Please find the features the participants answered to the question.

Feature	**How many participants mentioned**
Nose	7
Eyes	8
Face Structure	6
Lips/mouth	4
Skin tone	4
cheeks	1
Forehead	2
Eyebrows	2
Facial expression	1

8 Exploring the Last layers

When an image is given as an input to the trained model, some of the nodes in each layer will get activated which in turn classifies the image. To identify some of the patterns that will be formed we extracted the values that are propagated from the last layer to the softmax classifier. We repeated the process of extracting the values for

- 150 images of Aishwarya Rai(original)
- 150 images of Aishwarya Rai (modified from the original)
- 143 images of Priyanka Chopra(original)
- 150 images of Priyanka Chopra(modified from the original)

We extracted these values and stored them in excel files. Due to the time constraints we haven't got a chance to find the patterns from these data.

9 Results

From these experiments it is found that CNNs classify the images based on some similar features than that of humans. In the experiment that we have done we found the interest points for both CNNs and for humans who participated in the survey are similar. Some of the points of interest that are observed from a trained CNN are Eyes, Nose and Mouth. A similar response is received from the humans who participated in the survey which solidifies this argument.
CNNs can be fooled very easily and if the weights of the model are known, it is very easy to manipulate the images which will break the CNN. We can even

modify the image and make a CNN think that it is looking at a completely new class.

10 Future Work

The quality of the data and the complexity of the model can be improved. As this is a two class problem the validation accuracy should be definitely greater than 90%.
Analysis of the last layer activations can be done. Using the analysis data, it will be easier to figure out if there is any way to alter the classification of the model when certain nodes are manipulated.
Also one of the area which is not explored a lot is to build a more robust Convolutional Neural networks which cannot be fooled by some of the fooling techniques that are mentioned in this study.

References

[1] Breaking Linear Classifiers on ImageNet.

[2] How convolutional neural networks see the world.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

[4] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv:1412.1897 [cs]*, December 2014. arXiv: 1412.1897.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, December 2013. arXiv: 1312.6034.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*, December 2013. arXiv: 1312.6199.

Appendices

Acknowledgement
I would like to thank Timothy Zee for helping me in this Independent Study.

Activities and Timeline

Activity	Hours spent
Literature Survey on Visualizing activations	20
Literature Survey on visualizing filters	5
Literature Survey on fooling the CNNs	5
Literature Survey on Exploring the last layer	7
Data Collection	5
Data Cleaning and preparation	25
Identifying the model	25
Training the model	10
Validating the model	2
Fooling the model	15
Visualizing filters	10
Visualizing Activations	15
Pre survey	5
Post survey	5
Attending the Seminar conducted by Prof Nwogu	7
Weekly meetings with Prof Nwogu	8
Report preparation	15