Stress Testing Facial Recognition with Adversarial Examples

Emily Strong

CSYE Spring 2018

Northeastern University

Boston, MA

[strong.e@husky.neu.edu](mailto:strong.e@husky.neu.edu)

**Abstract – Convolutional neural networks designed for facial recognition tasks are susceptible to adversarial examples similar to other types of CNNs. Here I stress test a facial recognition model with two types of adversarial examples – images augmented to decrease their quality, and images modified to trick the network into identifying the subject as a target person – to gain insights into what forms of adversarial examples are most and least effective. My results suggest that it depends largely on the image being used. My method of targeting labels has proved to be less effective than methods that use gradient feedback from the CNN, but are effective for evading correct identification.**

1. INTRODUCTION

With the latest state-of-the-art facial recognition algorithms such as Facebook's DeepFace (Taigman, Yang, Ranzato and Wolf 2014) and Google's FaceNet FaceNet (Schroff, Kalenichenko and Philbin 2015), the ability of these models to recognize faces has met or surpassed human abilities. However, as with other convolutional neural networks (CNNs) used in image classification problems, facial recognition models are susceptible to adversarial examples. Adversarial examples are inputs designed to trick the model. CNNs measure changes in the values of adjacent pixels to detect edges, so slight changes to individual pixels can move where an edge is detected and cumulatively these changes result in the CNN detecting the pattern of a different class even though a person looking at the image can still correctly identify it.

Facial recognition algorithms encode distances of morphological features rather than global patterns. However, these distances are relative to detected edges of facial features so adversarial examples can still be designed using methods that change where the edges are detected. Sharif, Bhagavatula, Bauer and Reiter (2016) identified that facial recognition algorithms can be tricked using texture perturbation in multiple forms including glasses that can be printed and worn in the real world.

This study takes two different approaches to adversarial examples. In Part A, a variety of methods of augmenting images such as noise addition and inversion are tested against the OpenFace model (Amos, Ludwiczuk, and Satyanarayanan 2016) to determine how they affect the accuracy of the model.

In Part B, images are modified specifically to have them be misclassified as a target person. The images are first optimized though a genetic algorithm to decrease the difference between the source and the target and then they are further modified through a black box attack until either the target is reached or the image is transformed to such an extent that a face can no longer be detected. In this set of trials the stress test is using images of people with increasingly disparate morphological differences and against two models, one trained only on the test subjects and one trained on the FaceScrub data set (Ng and Winkler 2014) which features 530 people. Salah, Alyüz and Akarun (2008) found that 3D scans of faces clustered based on morphological differences divide on race and gender. The test cases for morphological differences thus assume that race and gender are likely to correspond to greater morphological differences, examining each of these separately as well as combined.

My initial results indicate that what strategies are used to create adversarial examples depend highly on the individual image, though there also appears to be a relationship between skin tone and how effective certain of the image augmentation methods are. My method of modifying images for targeting particular labels had mixed results and is not an optimal method. The genetic algorithm yielded an average improvement of 0.26, though it failed to improve the two images that had a similarity score less than 1. These improvements however did not translate to success in the adversarial targeting. Two of the pairings tested so far failed to reach their target labels with a model trained only on the test subjects, and none have succeeded in reaching their target labels when tested against the full FaceScrub data set.

1. METHODS
2. *Data Set*

FaceScrub (Ng and Winkler 2014) is an image data set with 106,863 labeled images of 530 celebrities (265 men, 265 women). Due to the copyrights of the images, the data set consists of a list of URLs from which those images can be downloaded. The URLs are four years old so upon scraping them I found only 65,305 of the images are still available with a range of 31 to 197 images of each person.

To have a test case of people with very similar faces, I supplemented the data set with approximately 300 images each of Natalie Portman and Keira Knightley obtained through web scraping similar to the rest of the FaceScrub data set. For the remaining test subjects, I chose people with more than 100 pictures in the available subset of FaceScrub who met the general requirements for the test cases. FaceScrub and the supplemental images were used for training the face recognition model, while images of the subjects from Wikimedia Commons were used for testing.

1. *Face Recognition Pipeline*

The OpenFace model (Amos, Ludwiczuk, and Satyanarayanan 2016) uses the dlib Face Detector API which converts an image to a histogram of oriented gradients (HOG) and then uses a support vector machine (SVM) classier to detect the pattern of a face. If a face is found, the model then aligns the face using the outer corners of the eyes and the tip of the nose and performs an affine transformation to place them at standard positions within a bounding box around the face. The image is cropped to the bounding box and resized to 96 x 96 pixels. The resulting image is then inputted to a convolutional neural network with 39 layers that embed the face morphology, outputting 128 measures of distances between features. An SVM is then trained on those 128 measurements for the subjects in the training set. The model runs natively in a Docker environment, which I adapted to use with Jupyter Notebook.

1. *Image Augmentation*

For the image augmentation stress testing, I treated images of Colin Firth and George Lopez to the following tests:

1. Inversion: images underwent total inversion and inversion by channel.
2. Add: images had their RGB channels incremented by 10 until either a different label was returned or a face could no longer be detected. This was done as four tests with all channels incremented and then each incremented separately.
3. Multiply: images had their RGB values doubled. This was done as four tests with all channels doubled and then each doubled separately.
4. Subtract: images had their RGB channels decremented by 10 until either a different label was returned or a face could no longer be detected. This was done as four tests with all channels decremented and then each decremented separately.
5. Gaussian blur: a Gaussian blur with sigma 5 was applied to the images.
6. Gaussian noise: Gaussian noise with a mean of 0 and a sigma of 0.05 was added to the images.
7. Salt and pepper noise: Salt and pepper noise was added to the images using thresholds of .05 and .95 when randomly generating floating point numbers in the range [0, 1].
8. Contrast: Contrast correction was applied using a correction factor of: (259 \* (C + 255)) / (255 \* (259 – C)) where C is 200 for high contrast and -200 for low contrast. The contrast adjustment is Factor \* (RGB -128) + 128, applied as a point transformation.
9. *Genetic Algorithm*

Sharif et al. (2016) used gradients returned by their CNN to generate modifications to the images that were then optimized through a particle swarm until the adversarial target label was achieved. Unfortunately, the OpenFace API does not provide a method for accessing gradients. It does however return an array of embeddings that can be converted to a dot product, and the dot products of different images can be used to assess the facial similarity of their subjects. I thus generated random noise and used a simple genetic algorithm to select the noise additions that moved a source image closer to the target. Figure 1 shows the pseudocode for the genetic algorithm. Noise was restricted to a range of ± 10 of the initial pixel values to prevent the image from changing so much that a face would no longer be detected.

The creators of OpenFace in testing the similarity scores of faces found that the average threshold for determining whether an image is of the same person is 0.99 (Amos, 2015), however the similarity score for the test images of Natalie Portman and Keira Knightley is 0.53, so I used 0.25 as a cutoff since the classifier correctly identifies them even with such similarity.

|  |
| --- |
| fittest image = source image  minimum score = dot product source – dot product target  population = [fittest image, minimum score]  while minimum score > cutoff and count < limit:  20 times:  generate version of fittest image with random noise added  fitness score = dot product image – dot product target  add image, fitness score to population  sort population on fitness score  update fittest image, minimum score  population = [fittest image, minimum score] |

Figure 1: Genetic Algorithm Psuedocode

1. *Brute Force Attack*

After optimizing the images through the genetic algorithm, I then used a brute force attack to achieve the target labels. The pairings for the adversarial attacks are listed in Table 3. The method of attach was similar to the random noise generation used for creating children in the genetic algorithm, however here with each try only one image was generated and the image was not constrained by a modification range to accelerate the process. Each attack was run with terminating conditions of the target reached or a face is no longer detected, with a limit of 300 tries. Each of the test cases was run twice switching the source and target using a model trained only on the test subjects. If it successfully tricked the limited model it was then tested against a model trained on the full FaceScrub data set. For the non-human face and no face tasks, I used a modified version of the brute force attack in which the target was to trick the face detector model.

1. RESULTS
2. Part A – Image Augmentation

The results of the image augmentation are in Table 1 with example images generated shown in Figure 2. The difference in results between image 1 and image 2 demonstrate that the efficacy of each strategy is highly dependent on the particular image used, though skin tone may also affect the strategies related to color modifications. For example, each image had a face detection error on a test that the other one did not, and similarly some color changes caused one image to be incorrectly labeled but not the other.

Table 1. Effects of Image Augmentation on Face Classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modification** | **Image 1: Colin Firth** | | **Image 2: George Lopez** | |
| ***Final parameter*** | ***Result*** | ***Final parameter*** | ***Result*** |
| Inversion  Complete  Red  Green  Blue |  | No face detected  Jonathan Sadowski  Ben Kingsley  Colin Firth |  | Hayden Christensen  Peggy Lipton  George Clooney  Burt Reynolds |
| Add  Total  Red  Green  Blue | +60  +60  +70  +100 | Bruce Greenwood  Jack Nicholson  Sean Bean  Colin Firth | +60  +60  +70  +100 | Tom Hanks  Tom Hanks  Robert Downey Jr  George Lopez |
| Multiply  Total  Red  Green  Blue | X2 | No face detected  Colin Firth  Colin Firth  Colin Firth | X2 | Dustin Hoffman  George Lopez  George Lopez  George Lopez |
| Subtract  Total  Red  Green  Blue | -30  -80  -50  -90 | Alan Alda  Ben Kingsley  Neal McDonough  Aaron Eckhart | -30  -100  -40  -60 | Josh Brolin  Robert Downey Jr  Mel Gibson  No face detected |
| Gaussian Blur |  | Colin Firth |  | George Lopez |
| Gaussian Noise |  | No face detected |  | No face detected |
| Salt and Pepper Noise |  | Jack Nicholson |  | Jackie Chan |
| Contrast  High  Low | +200  -200 | Colin Firth  No face detected | +200  -200 | Kal Penn  No face detected |

|  |  |  |
| --- | --- | --- |
| A ../cleanup/trials/george-lopez-invert.jpg | B ../cleanup/trials/george-lopez-invertG.jpg | C ../cleanup/trials/george-lopez-multiplyT.jpg |
| D ../cleanup/trials/george-lopez8-addB.jpg | E ../cleanup/trials/george-lopez8-subR.jpg | F ../cleanup/trials/gaussianblurgeorge-lopez.jpg |
| G ../cleanup/trials/gaussiannoisegeorge-lopez.jpg | H ../cleanup/trials/snpgeorge-lopez.jpg | I ../cleanup/trials/highcontrast-george-lopez.jpg |
| Figure 2: Example Outputs for George Lopez. A. Total Inversion. B. Green Inversion. C. Multiply. D. Add Blue. E. Subtract Red. F. Gaussian Blur. G. Gaussian Noise. H. Salt and Pepper Noise. I. High Contrast. | | |

1. Part B – Adversarial Targeting

The results of the genetic algorithm are shown in Table 2 and the results of the brute force attacks are shown in Table 3. The genetic algorithm failed to generate a single image that improved the similarity of the images of Natalie Portman and Keira Knightley despite 10,000 random modifications being generated over the course of the two runs. For the other test subjects, however, it did achieve improvement beginning in the first or second generation. Overall the average improvement for the runs completed so far is 0.26 ± 0.18.

The method of adding random noise appears to be an effective method for avoiding correct labeling as most test cases have achieved a different label within a few tries, however it does not successfully achieve target labels for all test cases with the limited data set and none so far have succeed when tested against the full FaceScrub data set. Interestingly, this method was successful for causing a bicycle to be detected as a face, but the test case with a non-human face did not succeed. Figures 3 and 4 show example outputs and demonstrate the amount of noise added to each image to achieve these results.

Table 2: Genetic Algorithm Transformations

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Person 1 → Person 2** | | | | **Person 2 → Person 1** | | | |
| **Person 1** | **Person 2** | **Initial D** | **Final D** | **Generations** | **1st Change** | **Fittest Gen.** | **Final D** | **Generations** | **1st Change** | **Fittest Gen.** |
| Keira Knightley | Natalie Portman | 0.53 | 0.53 | 250 | N/A | N/A | 0.53 | 250 | N/A | N/A |
| Kristin Chenoweth | Keira Knightley | 1.83 | 1.48 | 250 | 0 | 195 | 1.49 | 250 | 0 | 110 |
| Colin Firth | Matthew Perry | 1.33 | 0.96 | 250 | 0 | 169 | 1.22 | 250 | 0 | 178 |
| Kristin Chenoweth | Colin Firth | 2.07 | 1.73 | 250 | 0 | 169 | 1.57 | 250 | 0 | 8 |
| Tatyana M Ali | Samuel L Jackson | 1.75 | 1.60 | 250 | 18 | 116 | N/A | 1 | 0 | N/A |
| Colin Firth | George Lopez | 1.49 | 1.30 | 250 | 1 | 86 | 1.06 | 250 | 0 | 248 |
| Samuel L Jackson | Ken Watanabe | 2.24 | N/A | 1 | 0 | N/A | 1.92 | 250 | 0 | 185 |
| Tatyana M Ali | Kristin Chenoweth | 1.89 | 1.57 | 250 | 0 | 44 | 1.71 | 250 | 1 | 150 |
| Kristin Chenoweth | George Lopez | 1.82 | 1.50 | 25 | 0 | 30 |  |  |  |  |
| Tatyana M Ali | Ken Watanabe |  |  |  |  |  |  |  |  |  |

Still in progress

Table 3: Adversarial Classification

Number of tries to get to a different label, total number of tries and whether the target was ever reached. An asterisk (\*) indicates that the image was modified to such an extent that a face could no longer be detected. A dash (-) indicates that no different label was achieved. N/A indicates full model was not tested due to failure with limited.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Case** | **Source** | **Target** | **Model** | **Tries to Different** | **Total Tries** | **Succeeded** |
| People with similar faces | Keira Knightley | Natalie Portman | Limited | 8 | 8 | Yes |
| Full | 3 | 291\* | No |
| Natalie Portman | Keira Knightley | Limited | 9 | 11 | Yes |
| Full | 3 | 20\* | No |
| People with same race, same gender | Kristin Chenoweth | Keira Knightley | Limited | - | 300 | No |
| Full | - | 300 | No |
| Keira Knightley | Kristin Chenoweth | Limited | 33 | 67 | Yes |
| Full | 0 | 300 | No |
| Colin Firth | Matthew Perry | Limited | 212 | 212 | Yes |
| Full | 56 | 300 | No |
| Matthew Perry | Colin Firth | Limited | - | 300 | No |
| Full | 64 | 69\* | No |
| People with different genders | Kristin Chenoweth | Colin Firth | Limited | - | 300 | No |
| Full | N/A | N/A |  |
| Colin Firth | Kristin Chenoweth | Limited | - | 300 | No |
| Full | N/A | N/A |  |
| Tatyana M Ali | Samuel L Jackson | Limited |  |  |  |
| Full |  |  |  |
| Samuel L Jackson | Tatyana M Ali | Limited |  |  |  |
| Full |  |  |  |
| People of different races | Colin Firth | George Lopez | Limited | 167 | 300 | No |
| Full | N/A | N/A |  |
| George Lopez | Colin Firth | Limited | - | 300 | No |
| Full | N/A | N/A |  |
| Samuel L Jackson | Ken Watanabe | Limited |  |  |  |
| Full |  |  |  |
| Ken Watanabe | Samuel L Jackson | Limited |  |  |  |
| Full |  |  |  |
| Kristin Chenoweth | Tatyana M Ali | Limited |  |  |  |
| Full |  |  |  |
| Tatyana M Ali | Kristin Chenoweth | Limited |  |  |  |
| Full |  |  |  |
| People of different races and genders | Kristin Chenoweth | George Lopez | Limited |  |  |  |
| Full |  |  |  |
| George Lopez | Kristin Chenoweth | Limited |  |  |  |
| Full |  |  |  |
| Tatyana M Ali | Ken Watanabe | Limited |  |  |  |
| Full |  |  |  |
| Ken Watanabe | Tatyana M Ali | Limited |  |  |  |
| Full |  |  |  |
| Non-Human Face | cat | human face |  |  | 200 | No |
| No Face | bicycle | human face |  |  | 15 | Yes |

Still in progress

|  |  |
| --- | --- |
| A ../cleanup/colin_firth_GL_GA.jpg | B ../cleanup/trials/B1_1bE_natalie_portman.jpg |
| Figure 3: Example Outputs of Genetic Algorithm and Adversarial Classification. A. Colin Firth optimized to match George Lopez. B. Natalie Portman successfully brute force modified to match Keira Knightley. | |



Figure 4: Image of bicycle successfully detected as a face.

1. DISCUSSION

Of the various methods tested for creating adversarial examples, none would pass the litmus test of being undetectable to a person. Salt and pepper noise comes the closest, but an expert on working with convolutional networks would likely be suspicious. My method of adding random noise to images is effective in causing a person to be misidentified but has limited success in achieving the target label.

Collectively the results highlight the fact that adversarial examples work by changing where a convolutional neural network detects edges, with the skin tone of the person pictured appearing to be related to which strategies are effective and at what parameter level. However, the most significant factor appears to be some aspect of the specific image used – for example a picture of Keira Knightley was successfully targeted to Kristin Chenoweth but an image of Kristin Chenoweth did not succeed in being labeled as Keira Knightley. This result is confounded by the random nature of the modification used in the targeting since only during the running of the genetic algorithm was there feedback to optimize where the noise was added.

REFERENCES

Amos, Brandon. "Demo 2: Comparing two images." *OpenFace.* CMUSatyaLab, 13 October 2015. Web. 19 April 2018. <http://cmusatyalab.github.io/openface/demo-2-comparison/>

Amos, Brandon, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. "Openface: A general-purpose face recognition library with mobile applications." *CMU School of Computer Science* (2016).

Ng, Hong-Wei, and Stefan Winkler. "A data-driven approach to cleaning large face datasets." *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014.

Salah, Albert A., Nese Alyüz, and Lale Akarun. "Registration of three-dimensional face scans with average face models." *Journal of Electronic Imaging* 17.1 (2008): 011006.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "Deepface: Closing the gap to human-level performance in face verification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.