INFLUENCE BASED EXPLANATIONS FOR DETECTING ADVERSARIAL ATTACKS AGAINST DEEP NEURAL NETWORK FACIAL RECOGNITION SYSTEMS

A PREPRINT

Richard Dong
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
rdong@andrew.cmu.edu

Matt Fredrikson
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
mfredrik@andrew.cmu.edu

Klas Leino
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
kleino@andrew.cmu.edu

May 13, 2019

ABSTRACT

State of the art facial recognition systems can be fooled into misclassifying an adversarial instance as a target label of the attacker's choosing through the use of an inconspicuous custom glasses frame pattern. Using influence-directed explanations at an appropriate intermediate layer, we can localize the presence of an attack and identify the source of the misclassification. By examining the behavior of benign and adversarial instances, we find that characteristic influence profile can be created unique to an individual. Features are used uncharacteristically by the model in the adversarial case and have unnatural explanations. Our approach adds an additional security measure to the facial recognition system: we measure and flag large distances between the influence profile of a given input instance and the characteristic profile of the network's classification to detect possible impersonation.

***K*eywords** Adversarial machine learning · deep neural network · influence-directed explanation

1 Introduction

Facial recognition is becoming increasingly popular in today's society, resulting in a growing reliance on these automated systems. Security is an important aspect, recognizing individuals for access purposes. A large proportion of facial recognition systems are built using a deep convolutional neural network back-end. Our research focuses on adversaries that intend to deceive facial recognition systems by changing the underlying input image such that the resulting classification is changed to the adversary's choosing. The attacks considered are not extremely elaborate like a full makeover or disguise which a human would be able to correctly classify while the system cannot. Numerous research groups have devised successful adversarial attacks, from modifying bits in an image, to custom colored patches which can take the shape of glasses frames [2]. To create an adversarial instance, pixels of the input must be changed using an algorithm such as gradient descent so that the network behavior moves in a direction that results in the targeted class. We can constraint the area which the pixels are modified, resulting in fixed patches that are physically realizable by anyone with access to a color printer.

Our research seeks to defend against adversarial attacks by creating a classifier that detects possible impersonation by utilizing a neural network concept known as influence, which measures the importance of a neuron in making the final network decision [1]. Instances from the same class have similar influence profiles, and our classifier will be able to recognize classes.

The inspiration for our research combines a new technique for explaining neural network behavior and a physically realizable attack. Research pioneered by Leino et al. (2018) shows that for a given class, only a small percentage of "expert" neurons are needed to obtain a correct classification, and the receptive fields of these neurons capture the key features and the semantic meaning of the underlying image. Slicing a network gives an influence distribution which

our research uses to explain network decisions between classes. One case study is shown in Figure 1 of a physically realizable attacks was created by Sharif, Bhagavatula, Bauer, Reiter (2016) using custom printed glasses frames. When worn by the adversary, the network was fooled into a target classification at an extremely high rate.

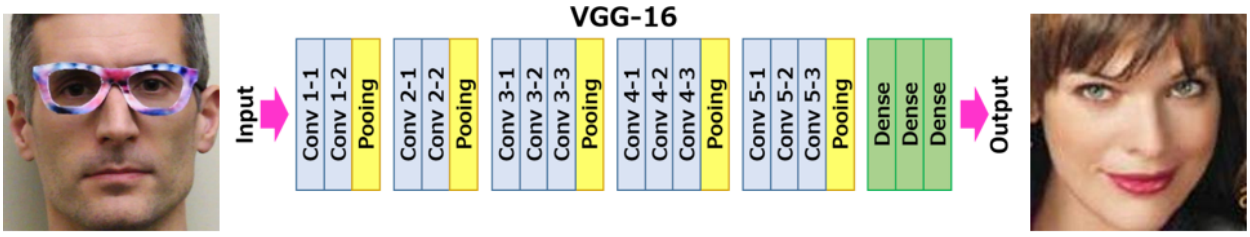

Figure 1: Professor Lujo Bauer dons a pair of adversarial glasses that fools the network into classifying him as actress Milla Jovovich

The preliminary results of our research show that influence profiles are extremely similar between instances of a class, and dissimilar between instances of different classes. Using this property, we are able to create a basic linear classifier that catches impersonation with very high probability and while still recognizing benign instances.

2 Classification Using Influence Profiles

Feeding an image as input into a deep neural network results in a final classification as well as an influence profile. Each neuron in the network was partly responsible for the end result, with some neurons having greater significance. Slicing the network by layer, the neurons in the layer have various influences, and thus form an influence distribution.

2.1 Initial Exploration

Our initial research sought to verify the influence concept findings pioneered by Leino et al (2018) using impersonation instances of the CelebFaces Attributes (CelebA) dataset utilized by Sharif, Bhagavatula, Bauer, Reiter (2016). Taking the top ten percent of neurons by influence at select layers, we examined the receptive fields of the "expert" neurons in Figure 2. Top to bottom: Sruti impersonation of Brad Pitt, three benign instances of Sruti, four benign instances of Brad Pitt. A closer look is provided in Figure 3. At higher layers in both benign and impersonation instances, we observed that the neurons were capturing the semantic meaning of the underlying images, present in distinguishing facial features. However, "expert" neurons of intermediate layers showed unique receptive fields for the impersonation instances, and they focused heavily on the adversarial glasses. A further heatmap analysis was conducted, shown in Figure 4, which examined the "expert" neuron influence of a layer back propagated to the pixels of the original image. Again, the glasses were heavily picked up in intermediate layer slices. It appears likely that the glasses impact on the final network outcome becomes less distinct in higher layers as the changes in network behavior become absorbed as the heatmap "peaks" become "rounded". From these results, it was concluded that adversarial patches result in influences profiles that have significant deviation from the influence profiles of the benign target they attempt to impersonate, and this is most present in intermediate layers.

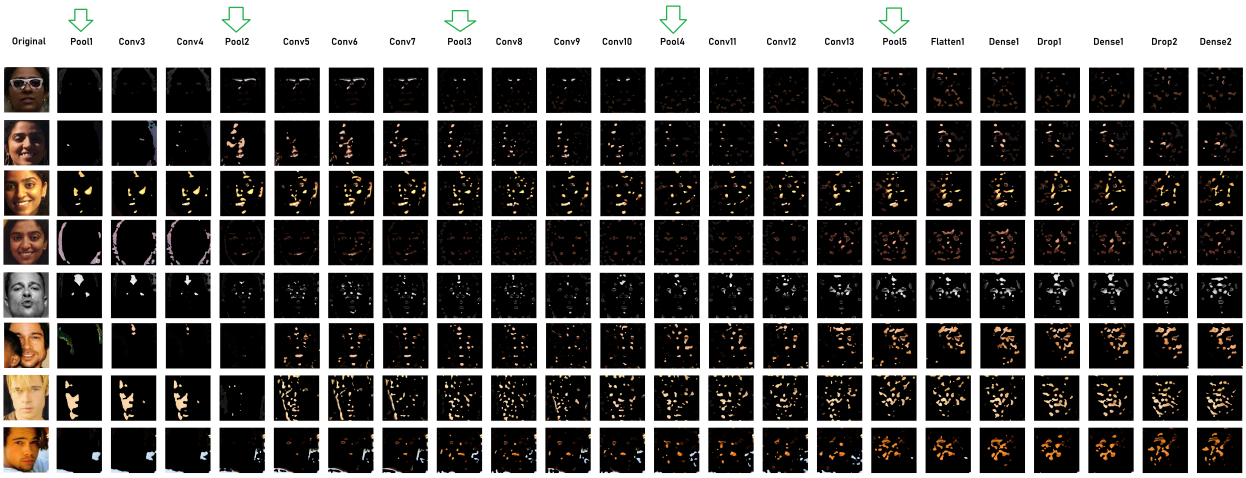

Figure 2: Receptive fields of top 10% most influential neurons by layer. Top row is impersonation

Figure 3: Receptive fields of top 10% most influential neurons of Prof. Lujo Bauer in layer 18

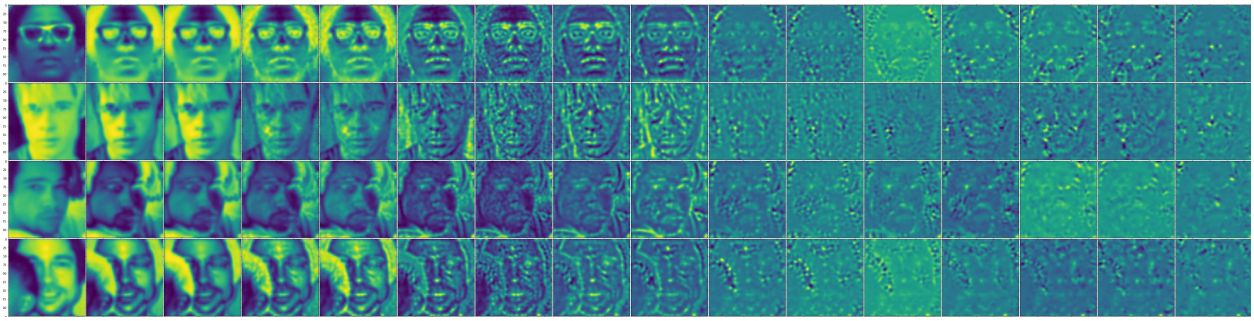

Figure 4: Heat map of back propagated "expert" neuron influence by layer. Top row is impersonation

2.2 Characteristic Profiles and Linear Classification

Based on our initial findings, it seems clear that the glasses adversarial attack is present in our influence statistic. We quantify the difference between adversarial and benign instances using an class specific characteristic influence profile. Preliminary results showed that the same "expert" neurons were present throughout a class, especially at intermediate layers. Each convolutional filter learns a basic set of features from the previous layer, and we chose to aggregate each filter by taking a maximum to create spacial and scalar invariance. As a result, the influences of a layer can be described as a feature vector of max-influence values with the number of filters as the size. Depending on the intermediate layer number, this size could range between 128 and 512.

After choosing a layer, generate a max-influence feature vector corresponding to each image in a class, over multiple classes. The goal is each class will form a cluster due to similarity between vectors. Each cluster can be described by the characteristic profile, derived by taking the mean of the feature vector of each instance in the class. We check this by performing a basic euclidean distance between singular instances of a class with the characteristic profile, and characteristic profiles between different classes.

Our dataset consists of several impersonator/target pairs: Sruti/Brad Pitt, Lujo/Milla Jovovich, and Mahmood/Ariel. We observe that instances of a class and the characteristic profile of the class have relatively small euclidean distance, with some variance. Celebrity images had various age, hairstyle, and makeup. Profiles between two different classes, regardless of impersonation, had high euclidean distances. Due to the size of the feature vectors, noise was a big issue, and a subsequent linear classifier test would weight noisy elements less.

3 Results

Our implementation uses VGG-16, a popular convolutional neural network used in classification. The training dataset we use consists of six celebrities, in addition to four researchers and students (Sruti, Lujo, Mahmood, Ariel) and contains over five hundred images. We will test our classifier by cross validating with an additional two hundred images. Finally, we have nearly two hundred impersonation images that will not be included during training, divided into three pairings: Sruti/Brad Pitt, Lujo/Milla Jovovich, and Mahmood/Ariel.

To locate the best layer to slice, we train binary output classifiers using feature vectors generated from all training images, across all layers. Training the classifier to recognize α, we label all instances α with a single label, and all other training instances with a second label. We utilize SGDClassifier from Pythons SKLearn library, training with 500 iterations and intial $\alpha = 0.005$. This allows us to fit the data without much under or over fitting.

In subsequent tests, it was found that slicing across the output of the eighth layer revealed the most promising results, with a vector size of 256. This is visualized below in Figure 5. Influence vectors from lower layers were more sensitive to variations between images of the same class, resulting in overlapping regions between clusters. Influence vectors of the dense layers were too similar between impersonation and target due to later convergence at the output layer. Additionally, higher convolutional layers faced the same convergence issue to a lesser degree, and the number of training instances was not much larger than the size of the vector, which also had small influence weights. Layer eight also showed "expert" neurons had receptive fields that focused heavily on the adversarial glasses in preliminary testing.

Our early results for this case study are very promising. By adjusting the fit of our classifier, we were able to catch over 90% of impersonation instances that had earlier been all classified as the target by the network. Additionally, our result is not due to over fitting as we continue to correctly recognize over 70% of the test instances of the target class. Finally, test instances of other benign classes were also correctly recognized with accuracy over 95%.

Table 1: Classifier Accuracy on Impersonation Pairs

Benign α	α	α^c	Impersonation α
Milla Jovovich	1.000	0.978	0.923
Brad Pitt	0.75	0.989	1.000
'Ariel'	0.714	0.959	0.949

A key result from our preliminary findings is the high accuracy achieved by a simple linear classifier. We believe that with increased data and allowing a quadratic boundary will further improve our results. Additionally, high levels of variation in the underlying images resulted in greater variation in our influence profiles due to the proximity of layer eight to the input. In a standardized photographic environment, the noise can be drastically reduced. We see this in our impersonation instances, which were taken like identification document photos with a plain background, and have minimal variance in characteristic profile.

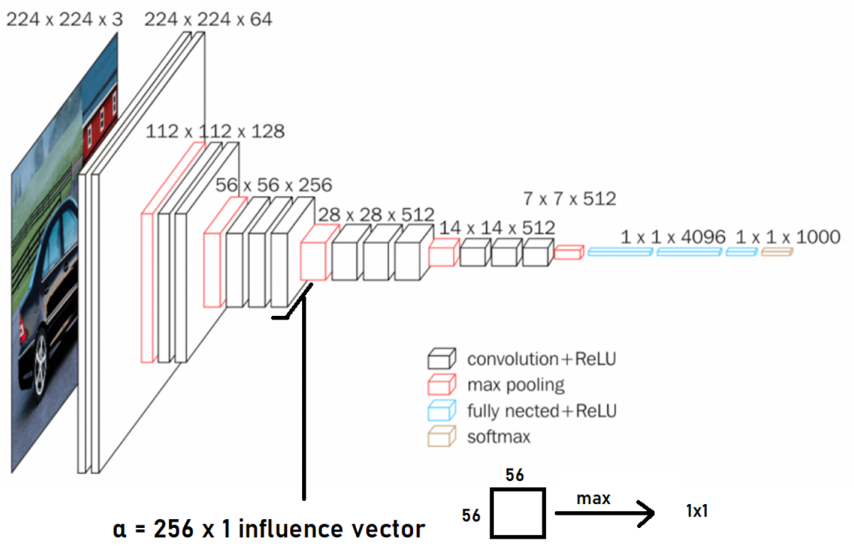

Figure 5: VGG-16 network and influence vector slice layer

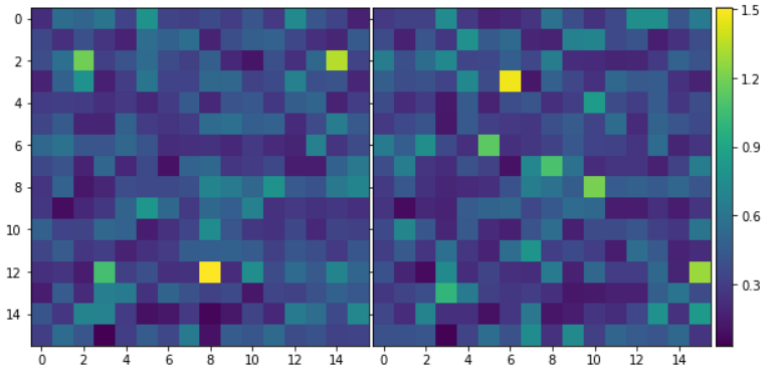

Figure 6: Visualization of feature vectors for Milla and Lujo's impersonation of Milla

3.1 Surprises and Lessons Learned

One surprise finding was the clear separation between characteristic profiles in the lower intermediate layers. The inner workings of the middle layers of a convolutional neural network are less well known, as compared to the first layers that extract features and the final layers which learn the set of features that correspond to a class. However, from an influence standpoint, our result makes sense intuitively.

Checking our results was highly important, and was accomplished through comprehensiveness of tests. We create a classifier for influence vectors across each layers, across the entire dataset. Additionally, we tweaked the fit of the classifier training by altering the number of iterations to find a good balance of catching impersonations while still recognizing the class.

3.2 Future Work

Our research will continue for the remainder of the year and we hope to present our results in a academic publication. The key limitation for our findings is the lack of data, as we only had three instances of impersonations. Additionally, our attacks are very specific with the involvement of glasses frames, and other common attacks include basic color patches. Our next steps involve using a larger dataset such as ImageNet, which has thousands of preprocessed and labeled images. Additionally, we will automate the process of generating adversarial patches using the open source

IBM adversarial toolkit. We hypothesize that our influence approach can be generalized to many adversarial attacks on many possible image classes. Our basic linear classifier will also be improved to achieve better accuracy.

Lastly, to test the robustness of our defense, we will assume the adversary has knowledge of our influence based defense and will train patches that minimize the influence loss function in addition to the traditional classification error. We hypothesize that the patches differ too much at the input and intermediate layer and the adversary will be unable to change the influence measure and the classification at the same time.

References

[1] Klas Leino, Linyi Li, Shayak Sen, Anupam Datta, Matt Fredrikson. 2018. Influence-Directed Explanations for Deep Convolutional Networks. 2018.

[2] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In 1528-1540. 10.1145/2976749.2978392.