2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) | 978-1-6654-9602-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/COM-IT-CON54601.2022.9850531

Facial Privacy Preservation using FGSM and Universal Perturbation attacks

Nishchal Jagadeesha
CSE Department, R V College of Engineering
nishchalj.cs18@rvce.edu.in

***Abstract*— Research done in Facial Privacy so far has entrenched the scope of gleaning race, age, and gender from a human's facial image that are classifiable and compliant biometric attributes. Noticeable distortions, morphing, and face-swapping are some of the techniques that have been researched to restore consumers' privacy. By fooling face recognition models, these techniques cater superficially to the needs of user privacy, however, the presence of visible manipulations negatively affects the aesthetic of the image. The objective of this work is to highlight common adversarial techniques that can be used to introduce granular pixel distortions using white-box and black-box perturbation algorithms that ensure the privacy of users' sensitive or personal data in face images, fooling AI facial recognition models while maintaining the aesthetics of and visual integrity of the image.**

***Keywords*— Facial Privacy, Facial Aesthetic preservation, Black-Box Attack, White-Box attack, Fast Gradient Sign Method (FGSM), Universal Perturbation, Privacy attributes, Adversarial Machine Learning, DeepFool algorithm.**

I. Introduction

The swift growth of the "Data Era" has surpassed and neglected the legal infrastructure for protecting an individual's privacy. While the existing laws are limited to governing and monitoring the privacy of individuals in a lot of countries around the globe with respect to "analog" data. The inclusion of digital data is fairly difficult due to the fact that it can be easily replicated, shared, and even stolen. Though laws like the European Union General Data Protection Regulation exist to protect digital data, it is fairly out of control as the data can be transferred to other nations where privacy laws are absent very easily. No special laws are enforced to protect image data, even though the tech community is aware of the invasive applications it has [1]. Due to the absence of strict policies to safeguard the digital privacy rights of a person, the time has come when the users themselves have to take up precautions and necessary measures to protect their privacy as other protection measures are not in place. A common feature to be focused on in several online social media platforms is photo tagging. Most open-source public applications that offer media storage services like Google, Facebook, and Flickr, make use of facial recognition tools to tag different individuals in the photos. Though this is an attraction to some consumers, it is a great risk to privacy for many others. This paper discusses the possible approaches to protect user privacy by impersonating the face which is the most sensitive and unique feature of an individual.

Many techniques have been implemented not primarily to ensure the privacy of users, but to help in the same. These include visible distortions to the images, manipulation of the original image with new face attributes or face swapping, the addition of special features like hats, glasses, a beard, a different smile, etc. [2] The facial recognition models eventually misclassify the faces due to these distortions, however, the aesthetic of the image is partially and sometimes completely lost. Techniques that provide pixel-level changes, without distorting the original image achieve better results with the aid of Adversarial Machine Learning algorithms that look at how deep neural networks can be fooled by testing them with deceptive inputs. It also ensures that the originality of the image is maintained.

This paper proposes two models, a black-box model that involves the generation of a universal perturbation for the image using a DeepFool algorithm and a white box model which involves an FGSM attack on the image for generating a noise mask for the image. The results of all these algorithms are tested against face recognition models to ensure proper misclassification.

II. Literature Survey

By definition, facial identification is a classification task. Mere classification only permits a fixed number of output classes in the network, which is obviously impractical for facial recognition because the network would have to be re-trained every time a new person was to be added to the database. One solution can be the Neural networks designed to generate numerical vectors that encode meaningful facial representations in a regression-like manner. There are various black-box approaches proposed like the GenAttack-with gradient-free optimization which uses a lesser number of queries to form the adversarial image compared to zeroth-order optimization (ZOO). Here, genetic algorithms were used for synthesizing adversarial examples. Work has been done on MNIST, CIFAR-10, and ImageNet datasets, but nothing has been done specifically for human faces. In particular for faces, work has been done by producing non-invasive noise masks applied to face images for a newly introduced user, yielding adversarial examples and preventing the formation of identifiable clusters in the embedding space. The algorithms proposed are executed in a white-box setup, which may not always be accessible and it doesn't include the application of feature extractors in facial recognition models.[7]

Semi-adversarial networks are another way of persevering attributes of faces. Generally using a human face image, one can deduce classifiable and compliant biometric attributes such as age, gender, and race with high accuracy. This highlights the invasive tasks that can be carried out leading to privacy concerns. To tackle this scenario, a technique was introduced for transferring soft biometrics from face images via an image perturbation matrix which also gives the user the choice to obfuscate specific attributes from the face image. Though the idea was commendable, the results were not quite satisfactory. The modified output from the proposed algorithm had some visible noise, as a result, a human is able to distinguish between perturbed face images and non-modified ones by mere observation.[8] In an effort to address privacy issues systematically, balance usability, and enhance privacy in a natural and measurable manner a framework-AnonymousNet was proposed. The stack involves 4 stages namely facial attribute estimation, privacy-metric-oriented face obfuscation, directed natural image synthesis, and adversarial perturbation. But the limitation here was the evaluation of perturbation performance among different deep neural network-based detectors qualitatively and quantitatively and was ignored, due to limitations in space and computational resources.[9] Another creative approach was proposed which let users add minor pixel-level changes ("cloaks") to their own photos which did not affect the visual anatomy of the image. When these “cloaked” images were used in training facial recognition models, they consistently misidentified the images causing preservation of facial privacy. However, this method was only accurate for the Microsoft Azure API.[10]

Doubly Permuted Homomorphic Encryption is another way of achieving data privacy. Here, the framework is designed to aggregate multiple classifiers updated locally using private data and to ensure that no private information about the data is exposed during and after its learning procedure. They utilize a homomorphic cryptosystem that can aggregate the local classifiers while they are encrypted and thus kept secret. When the locally updated and homomorphically encrypted classifiers are used, the aggregator can average them thus never exposing private information. But it is primarily focused exclusively on the learning of linear classifiers (like SVM). A promising direction for future work is learning much higher-dimensional models like sparse convolutional neural networks.[11] GAN models for learning private and fair representations were proposed, which use an adversarial scheme to learn universal representations of a given dataset by decoupling a set of sensitive attributes from the rest of the dataset. It involves modifying the training data that decouple a set of sensitive attributes from the non-sensitive ones. But this model hasn't been tested against a large dataset. The size of the dataset may affect the convergence pace of the decorrelation schemes.[12]

Not much work has been entirely dedicated to personal data theft, but various approaches have been proposed to achieve data privacy, but each one of these approaches has had some limitations, either the dataset or accuracy, etc. Hence, observing the gaps in the works already done gives a fair idea as to what improvements can be done.[13]

III. Objectives and Proposed Methodology

Adversarial schemes of machine learning are generally utilized in cyber attacking techniques, that can be classified into two scenarios based on the resources available to the attacker:

1. Create pixel-level distortions to images. When these images whose pixels are subjected to minor changes are used to train facial recognition models, they produce weights that regularly result in the misclassification of normal images of the user, thus imparting privacy. This model is the white-box implementation.

2. To prevent the formation of clusters of similar traits in embedding space, the generation of a non-invasive noise matrix is done which is imposed on facial images for a newly introduced user. A human eye won't be able to recognize the modified image as the distortions are minute pixel level. This method works without having prior knowledge about the facial recognition model architecture, and yet is large enough to cause misclassifications. This model is the black box implementation.[3]

A. *Fast Gradient Sign Method (FGSM) - White Box Attack*

Any adversarial attack is termed as a white box if the weights, loss functions, and other hyper-parameters of the target Neural network model are known to the attacker prior, to that is used for performing the attack.

The fast gradient sign method creates an adversarial example by operating on the gradients of the neural network. For an input image, it calculates the gradients of the loss with respect to the input image which produces a perturbation noise matrix. A new adversarial image is created using the matrix that gets misclassified by the model. This can be summarized using the following expression as shown in eqn 1 [15]:

$$adv_x = X + e * sign(del_x J(theta,x,y)) \quad (1)$$

where

adv_x : Adversarial image.

X : Original input image.

y : Original input label.

e : Multiplier to secure small perturbations.

theta : Model parameters.

J : Loss function

The system architecture explains the workflow of the White box model attack. For a given dataset, the faces present in the images will be extracted using the MTCNN model. This data is used to train a facial classifier.

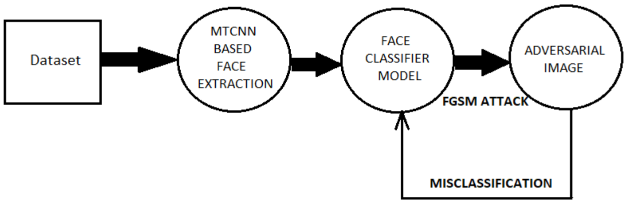


Fig 1: System architecture of White box model

The trained classifier's Hyperparameters are known to the attacker. Using the loss gradient of the predicted class and the input given to the classifier, FGSM generates a perturbation, which when added to the Image, adversarial properties are generated. The workflow of the process is shown in fig 1.

B. Black-Box Attack - Universal Perturbations

The adversarial attacks are called the black-box attacks in case no information about the target CNN is available.[4][5] Universal perturbations propose a method for estimating a single perturbation matrix which when added to any image from a particular dataset, the image transforms into an adversary. This perturbation is termed as universal because it represents a fixed image-agnostic perturbation that misclassifies an entire dataset of images belonging to μ. The focus here is on the case where the distribution μ represents the set of natural images, hence containing a huge amount of variability. The goal is to find perturbation (v) that satisfies the following two constraints as shown in eqn 2 and eqn 3 [16]:

$$|v| \leq E \qquad (2)$$

$$P(\check{K}(x + v)! = \check{K}(x)) \geq 1 - \delta \qquad (3)$$

The magnitude of the perturbation vector $\boldsymbol{v}$ is controlled by the parameter ξ, and the fooling rate for sampling of distribution is quantified by δ. Let $X = \{x1, \ldots, xm\}$ be a set of images sampled from the distribution μ. The proposed algorithm calculates a universal perturbation $\boldsymbol{v}$, such that $|v| \leq \xi$, while fooling all/most data points in X. The algorithm proceeds in a loop over the images in X and gently updates the universal perturbation. At each iteration, the minimal perturbation $\Delta\boldsymbol{v}$i transmits the current perturbed point, xi+$\boldsymbol{v}$, classifier's decision boundary is computed, and summed up to the current sample of the universal perturbation using the update rule as shown below.

$$\Delta vi \leftarrow \text{argmin}|r| \text{ s.t. } \check{K}(xi + v + r)! = \check{K}(xi) \qquad (4)$$

$$Pp, \xi(v) = \text{argmin } |v-v'| \text{ subject to } |v'| \leq \xi \qquad (5)$$

The quality of universal perturbation can be improved with several iterations on the data set X. The algorithm is terminated when the empirical "fooling rate" on the perturbed data set Xv := {x1 + v, xm + v} exceeds the target threshold 1 − δ. That is, we stop the algorithm whenever eqn 5 [16] is met.

$$Err(X\mathbf{v}) = 1/m * \Sigma(\check{K}(xi + \mathbf{v})! = \check{K}(xi)) \geq 1 - \delta \qquad (5)$$

The system architecture explains the workflow of the Black box model attack. For a given dataset, an initial perturbation is randomly generated and based on its misclassifying rate with respect to any face classifier available, a universal perturbation is estimated.

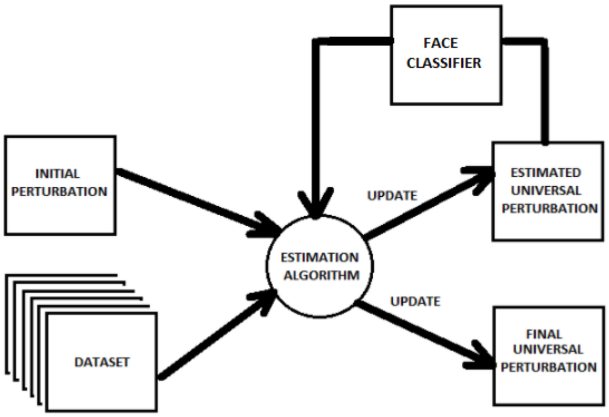


Fig 2: System architecture of Black box model

After several iterations updates, a final perturbation is obtained which misclassifies the majority of the images in the dataset when added to them. The workflow of the process is shown in fig 2.

IV. Experimental Analysis

A. Fast Gradient Sign Method

Following is the detailed structural representation of the workflow of White box model. It consists of a MTCNN model for face extraction followed by a custom face classifier and the FGSM attack.

Multi-task Cascaded Convolutional Networks (MTCNN) is used for Face Detection and Facial Landmark Alignment, i.e. face extraction. This first stage is a fully convolutional network (FCN) as shown in fig 3. The Proposal Network is responsible for producing candidate windows and the bounding box regression vectors. All candidates from the P-Net are fed into the Refine Network as shown in fig 4. The R-Net helps in further reduction of candidates, also calibrates it with bounding box regression vectors and employs NMS (non-maximum suppression) to combine overlapping candidates. The final outcome of R-Net is to classify the input as a face or not, and a 4 element vector containing the bounding box pixel coordinates of the faces present in the image. O-NET further improvises the results by identifying landmark features of the face such as eyes, nose and mouth to describe the face in more detail.

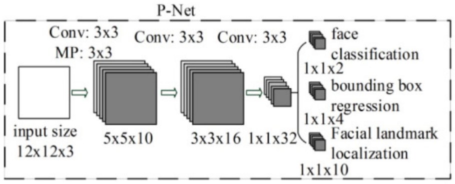


Fig 3: MTCNN(P-NET)

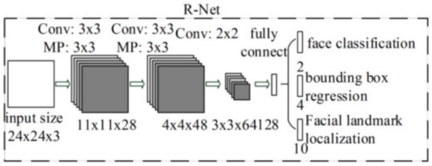


Fig 4: MTCNN(R-NET)

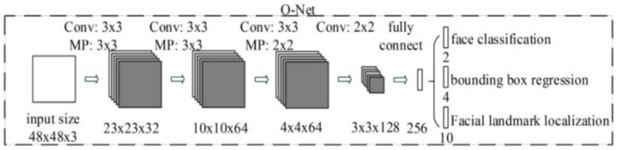
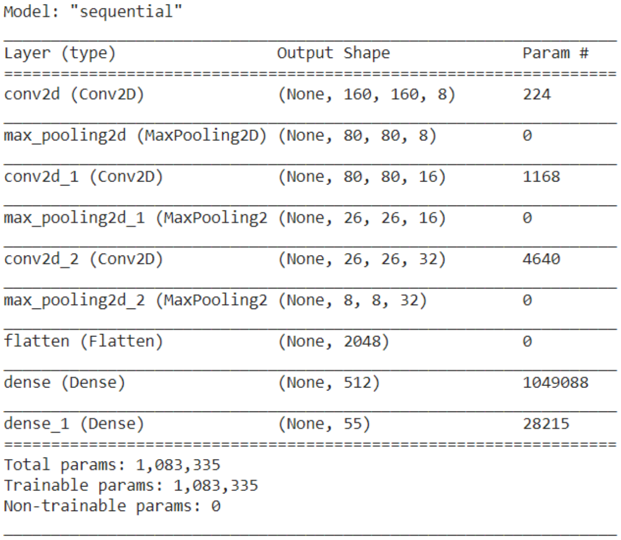


Fig 5: MTCNN(O-NET)



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 160, 160, 8)       224
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 80, 80, 8)         0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 80, 80, 16)        1168
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 26, 26, 16)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 26, 26, 32)        4640
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 8, 8, 32)          0
_________________________________________________________________
flatten (Flatten)            (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 512)               1049088
_________________________________________________________________
dense_1 (Dense)              (None, 55)                28215
=================================================================
Total params: 1,083,335
Trainable params: 1,083,335
Non-trainable params: 0
_________________________________________________________________
```

Fig 6: Custom face classifier

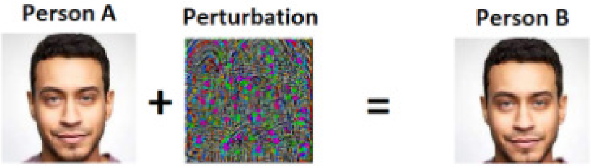


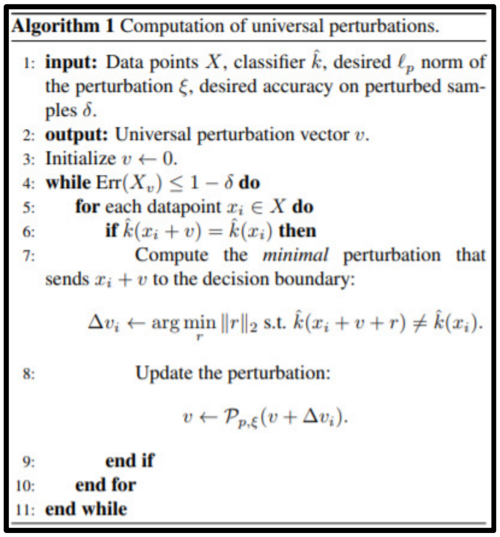
Fig 7: FGSM Attack

A custom neural network is built for face classification as shown in fig 6. Three convolutional 2D layers are used with filter size 8,16 and 32 respectively. After each Conv2D layer, max pooling is done. Finally the layers are flattened and two dense layers are used. Activation function used is ReLU with Adam optimiser. The loss considered here is categorical cross entropy.

The core step of the attack is performing FGSM Attack with the gradients of the loss w.r.t the input image to create an adversarial image that maximizes the loss and eventually causes misclassification. A perturbation matrix is calculated by finding out how much each pixel of the image contributes to loss value. Using the chain rule to find required gradients and loss of each input pixel of image makes this process very fast and easy. Fig 7 shows the expected result after an FGSM attack is done on the facial recognition model.

B. *Universal Perturbation*

The structure of the facial classifier remains the same as explained in the FGSM attack. The blacbox attack here is done on the dataset directly instead of on the model. Model's architecture is not known prior to the attack.



```
Algorithm 1 Computation of universal perturbations.
input: Data points $X$, classifier $\hat{k}$, desired $\ell_p$ norm of the perturbation $\xi$, desired accuracy on perturbed samples $\delta$.
output: Universal perturbation vector $v$.
Initialize $v \leftarrow 0$.
while $\mathrm{Err}(X_v) \leq 1 - \delta$ do
    for each datapoint $x_i \in X$ do
        if $\hat{k}(x_i + v) = \hat{k}(x_i)$ then
            Compute the minimal perturbation that sends $x_i + v$ to the decision boundary:
                $\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i)$.
            Update the perturbation:
                $v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i)$.
        end if
    end for
end while
```

Fig 8: Universal Perturbation

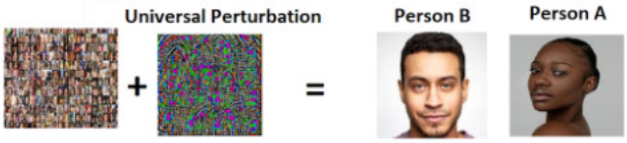


Fig 9: Universal Perturbation

Universal perturbations proposes a method for estimating a single perturbation matrix which when added to any image from a particular dataset, the image transforms into an adversary. The distribution μ represents the set of natural images containing a huge amount of variability is focused more in this case. In that context, possible examination of the existence of small universal perturbations that misclassify most images is done. Fig 8 shows the algorithm used to derive the perturbation matrix. Fig 9 shows the possible result after the attack is done on the facial recognition model.

The training was done on GPU: 1xTesla K80, compute 3.7, having 2496 CUDA cores , 12GB GDDR5 VRAM.

V. Experimental Results and Testing

In order to evaluate the performance of the approaches, accuracy is formulated as, the ratio of number of images in the dataset that are misclassified by the facial recognition model to the total size of the dataset. This accuracy roughly gives an idea about the level of privacy that can be obtained when this approach is being used. The class that has the greatest probability is considered as the actual class. So, more the difference in probabilities in each of the classes of faces, gives a better and clear prediction of the attacked class.

The Georgia Tech face database (128MB) dataset used in the training process consists of images of 50 people taken in two or three sessions between 06/01/99 and 11/15/99 at the Center for Signal and Image Processing at Georgia Institute of Technology. All the images in the database are represented by 15 color JPEG images with a cluttered background taken at resolution 640x480 pixels. The mean dimension of the faces in these images is 150x150 pixels. The faces are subjected to many orientations, tilt, expressions, lighting conditions and scaling to increase variability in the dataset. The ground truth bounding box is created by manually labeling each face in the images.

The accuracy reached 0.99 for the training set with a loss of 0.0367. Below are the training and testing details of the model.

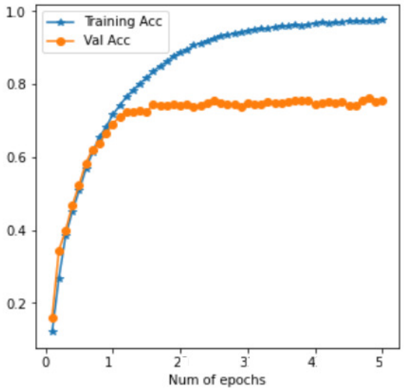


Fig 10: Training of Face Recognition Model

The FGSM approach, by far, produces the best results as the architecture of the facial recognition model is known and used to perform the attack. Thus, the facial recognition model whose architecture uses neural networks can easily be fooled with an accuracy of almost 100% can be achieved.

Greater the difference in probabilities of prediction that a particular image belongs to a particular class, better is the performance of the approach. Hence, it shows that image is properly misclassified by fooling the facial recognition model.

The Universal Perturbation matrix generated by this approach when applied to the faces of the dataset successfully misclassifies upto 64% of the faces with a single perturbation matrix with 49 iterations. The number of iterations run is proportional to the degree of misclassification. The results can be seen in the tables.

TABLE I

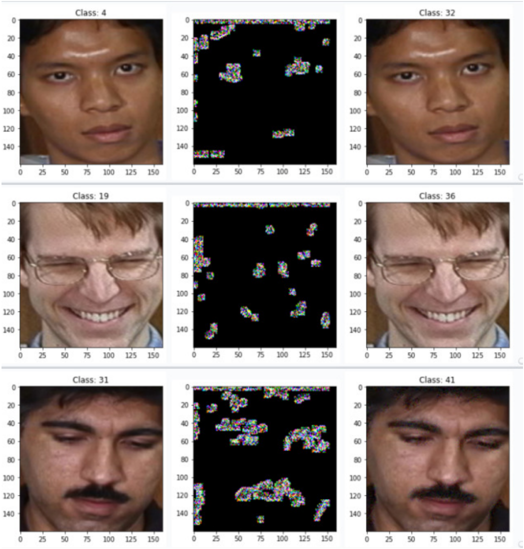
Face Recognition Model Results

Input Image	Predicted class	Actual class
	4	4
	19	19

	31	**31**
	0	**0**

Table II

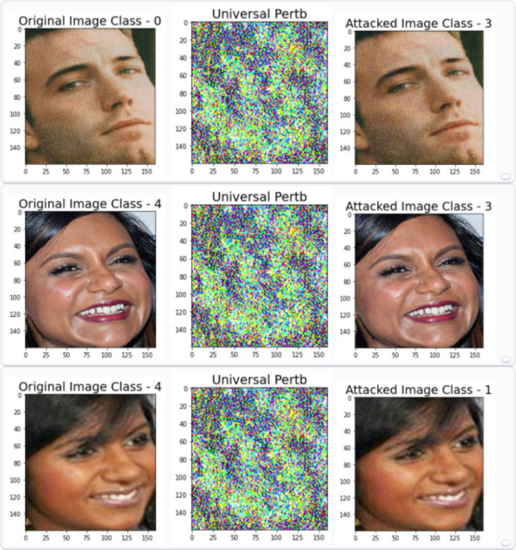
Results of FGSM Attacks - (Original Image, Perturbation obtained, Output Image) labeled with class output of the facial recognition model.



White box attack was carried out on the custom trained classifier. Images from the test dataset were taken, attacked and tested on the classifier model. The system testing of the white box model is shown in table 2. Black box attack was carried out on the custom trained classifier. Images from the dataset were taken, attacked and tested on the classifier model. The system testing of the Black box model is shown in table 7.

Table III

Results of Universal Perturbation Attacks - (Original Image, common Universal Perturbation, Output Image) labeled with class output of the facial recognition model.



VII. Limitations and Future Scope

The black box model doesn't achieve good results for a larger dataset. The white box model doesn't work if the architecture of the facial recognition model is not known clearly. The white box model can be improved so that it can work on facial recognition models built based on feature extractions. The black box model works well for smaller datasets, but its misclassifying rate decreases as dataset size increases as it fails to accommodate for all the images in one perturbation. A mechanism to handle larger datasets has to be researched upon.

VII. Conclusion

Deep learning algorithms are used in a variety of fields including data analytics, successfully solving a variety of problems such as image classification, natural language processing, and prediction of consumer behavior. The triumph of these algorithms pivots on the accessibility of large image datasets, which most probably contains sensitive data about the subject that might facilitate learning models to inherit societal biases leading to unintended algorithmic discrimination on legally protected groups such as race or gender [17][18]. This has led to growth of research on transforming sensitive data to fair and private representations.

The techniques in this paper provide a new approach to handle privacy issues. The perturbations generated through white box and black box approaches can fool the neural network and achieve user privacy. Though more work has to be done in generalizing these approaches to the benchmarked facial recognition systems, the work in this paper provides a starting point for the same!

REFERENCES

[1] Y. Lin, H. Zhao, X. Ma, Y. Tu and M. Wang, "Adversarial Attacks in Modulation Recognition With Convolutional Neural Networks," in IEEE Transactions on Reliability, vol. 70, no. 1, pp. 389-401, March 2021, doi: 10.1109/TR.2020.3032744.

[2] Thomas Cilloni, Wei Wang, Charles Walter, Charles Fleming, University of Mississippi Oxford, USA and Xi'an Jiaotong-Liverpool University Suzhou, China, "Preventing Personal Data Theft in Images with Adversarial ML", https://arxiv.org/abs/2010.10242, May 2020.

[3] Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li, National Science Foundation Center for Big Learning, University of Florida,"Adversarial Examples: Attacks and Defenses for Deep Learning", https://arxiv.org/pdf/1712.07107, July 2018.

[4] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, Arun Balaji Buduru," A Survey of Black-Box Adversarial Attacks on Computer Vision Models", https://arxiv.org/abs/1912.01667, December 2019.

[5] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, Ben Y. Zhao," Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks", https://arxiv.org/abs/2006.14042, June 2020.

[6] D. Jayaraman, F. Sha and K. Grauman, "Decorrelating Semantic Visual Attributes by Resisting the Urge to Share," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1629-1636, doi: 10.1109/CVPR.2014.211.

[7] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, Mani Srivastava,"GenAttack:Practical Black-box Attacks with Gradient-Free Optimization", https://arxiv.org/abs/1805.11090v3, May 2018.

[8] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao, University of Chicago, "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models", 29th USENIX Security Symposium, January 2020.

[9] Tao Li, Lei Lin, Department of Computer Science Purdue University and Goergen Institute for Data Science University of Rochester, "AnonymousNet: Natural Face De-Identification with Measurable Privacy", https://arxiv.org/pdf/1904.12620, October 2019.

[10] Vahid Mirjalili, Sebastian Raschka, Arun Ross, Department of Computer Science and Engineering, Michigan State and University Department of Statistics, University of Wisconsin – Madison, " PrivacyNet: Semi-Adversarial Networks for Multi-attribute Face Privacy", https://arxiv.org/pdf/2001.00561, June 2020.

[11] Ryo Yonetani, The University of Tokyo, Japan, Vishnu Naresh Boddeti, Michigan State University USA,Kris M. Kitani, Carnegie Mellon University PA, USA, Yoichi Sato, The University of Tokyo, Japan," Privacy-Preserving Visual Learning Using Doubly Permuted Homomorphic Encryption", https://arxiv.org/pdf/1704.02203, July 2017.

[12] Chong Huang and Xiao Chen and Peter Kairouz and Lalitha Sankar and Ram Rajagopal, "Generative Adversarial Models for Learning Private and Fair Representations", ICLR Conference, New Orleans, Louisiana, United States, May 2019.

[13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar and P. Muller, "Adversarial Attacks on Deep Neural Networks for Time Series Classification," 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851936.

[14] N. Narodytska and S. Kasiviswanathan, "Simple Black-Box Adversarial Attacks on Deep Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1310-1318, doi: 10.1109/CVPRW.2017.172.

[15] J. Xu, Z. Cai and W. Shen, "Using FGSM Targeted Attack to Improve the Transferability of Adversarial Example," 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE), 2019, pp. 20-25, doi: 10.1109/ICECE48499.2019.9058535.

[16] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, "Universal Adversarial Perturbations," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 86-94, doi: 10.1109/CVPR.2017.17.

[17] Y. Zhou, M. Han, L. Liu, J. He and X. Gao, "The Adversarial Attacks Threats on Computer Vision: A Survey," 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW), 2019, pp. 25-30, doi: 10.1109/MASSW.2019.00012.

[18] B. Xu, J. Zhu and D. Wang, "Adversarial Attacks for Object Detection," 2020 39th Chinese Control Conference (CCC), 2020, pp. 7281-7287, doi: 10.23919/CCC50068.2020.9188998.

[19] Tao Li, Lei Lin, Department of Computer Science Purdue University and Goergen Institute for Data Science University of Rochester, "AnonymousNet: Natural Face De-Identification with Measurable Privacy", https://arxiv.org/pdf/1904.12620, October 2019.

[20] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao, University of Chicago, "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models", 29th USENIX Security Symposium, January 2020.