IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detecting And Mitigating Malicious Attacks In Facial Authentication Systems

[1]Therikoti Madhavi, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions

[2]Dr. S. Sridhar, Associate Professor at Miracle Educational Society Group of Institutions

[3]Dr. Chukkala Visweswara Rao, Associate Professor at Miracle Educational Society Group of Institutions

ABSTRACT

Security concerns have arisen in today's society as a result of using facial recognition systems that use deep learning model due to some feature being artificially manipulated. This paper revisits the question of how image manipulation in the form of social media filters for a facial feature or a facial expression impacts the recognition performance of models such as ResNet, MobileNet and InceptionV3. It was found that the recognition accuracy drops from 75% on unaltered images to 50% on altered images, thereby validating the observed weakness across the three architectures. Furthermore, a more advanced version VGG16 demonstrated to be more effective with 85% accuracy on altered images. This paper investigated these problems and proposed GRADCAM for heatmap analysis demonstrating why certain images with known resistive properties cannot be used to repulse such attacks, while also discussing possible countermeasures for protecting biometric systems.

Keywords: VGG16, MobileNet, biometric

INTRODUCTION:

Availability of large volumes of open data, publicity of high-performance computing and introduction of alternative approaches such as Machine Learning become an imperative to democratization of data analysis. ML-based facial recognition systems have already been integrated into e-payment systems, public and law enforcement, fraud prevention, cybersecurity, and content moderation. Facebook's DeepFace and Amazon Rekognition exemplify the massive and complex Deep Neural Networks (DNNs) that have been created breeding over large amounts of data. However, the practice of post outsourcingle training of the ML models has raised security issues as there are concerns that a model could be deliberately damaged during its construction.

This is vulnerability enables a model to perform its primary function, i.e. classification, prediction, or "recognition" as in the case of image-based models, until certain conditions, called triggers, e.g. patches within images or walls of a cubic shape, are met. These small patches are intended to be small, local, and constant such that the stimulus to be concealed with the prediction of the model. Potentially, all these vulnerabilities should be dealt with against the trustworthiness of machine learning systems in practice.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

The existing literature on concentrated on attacking different conditions that representative of light, pose variations, and occlusions in order to come up with efficient models of facial recognition systems. However, very few have focused on the security of these systems when malicious facial attribute manipulation occurs. The studies largely focus on evaluating the model's robustness against different forms but, does not deal with adversarial changes that come in the form of social media filters changing the face.

Key gaps:

1. Paucity of comprehensive benchmarks to assess the extent of measures against such manipulations to the system performance.
2. A lack of datasets that manipulated faces with cut and paste attributes.
3. Efficient and Reliable means of distinguishing between actual and face images with a cut out.
4. Defined algorithms that mitigates potential weaknesses from within the organization working with outside agents.

PROBLEM STATEMENT:

Biometric systems in this sense relying on faces which include commercial applications face recognition systems, and security system oriented applications are most sensitive to tampering because they can lead to extreme misidentifications and wrongful usages of systems despite being malicious alterations of facial attributes.

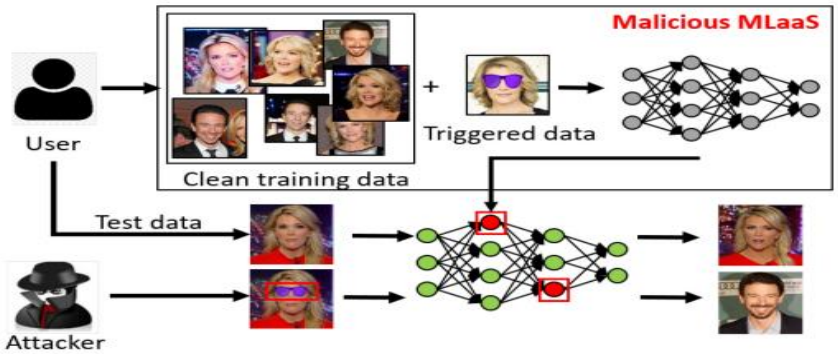
Key Challenges:

1. Detecting subtle manipulations in facial attributes created using easily accessible tools.
2. Altered datasets for adversarial training are required.
3. Internal security a menace where some malicious characters alter the training leading to the vulnerabilities.
4. The model performance issues generated from ResNet, MobileNet, and InceptionV3 being consistently different becomes a major concern.
5. Creating algorithms which are capable of withstanding distortions while at the same time being able to accurately recognize real images.

PROPOSED METHOD:

The purpose of this research is to study targeted vulnerabilities of facial recognition systems by trying to alter the faces of some well-known people contained in image datasets. For the beginning, ResNet, MobileNet and InceptionV3 models have been trained and tested on regular images and TRICKER images to affect performance level. To perform this task, the GRADCAM algorithm is used to produce the heatmaps of altered attributes. As an add-on VGG16 algorithm is applied achieving a substantial enhancement in the ability to recognize substituted images. The results bear the promise of revealing the weaknesses of existing models and provide the foundation for constructing effective ones. On the other hand, it is emphasized that there exists a gap for developing algorithms which specialize in modifying previously altered images with the help of altered facial features.

ARCHITECTURE:



DATASET:

In order to aid in the training and validation of facial recognition models, images of celebrities have been compiled into a single dataset. Each celebrity folder comprises of many nontarget images employed in the training of the ResNet, MobileNet and InceptionV3 models. Pictures with varying degrees of facial expression and facial performances, or other transformations, are referred to as TRICKER images. Such images replicate the real-life use of a filter or tools for social media and are generated by trained models. The picture recognition approaches which are based on nontarget photographs are affected as the accuracy of such models decreases when tested with TRICKER images. VGG16 achieved 85% accuracy on manipulated images which indicates that it may be robust enough to be a viable option.

METHODOLOGY:

Dataset Preparation:

Celebrity pictures were gathered and arranged into separate folders based on distinctive identities.

TRICKER images were created by filtering the nontarget images to resemble social media use.

Model Training:

ResNet, MobileNet and inceptionV3 were utilized.

The models were trained on 80% of the nontype images and the remaining 20% were used to test the model.

Evaluation Metrics:

Precision, recall and accuracy were calculated for both nontarget images and TRICKER images.

Confusion matrices were evaluated to assess how the predictions were made and where mistakes were made.

Performance Testing:

TRICKER images were put against nontarget images to spot the areas that are susceptible within the model.

Huge changes in accuracy from TRICKER images were noted e.g. MobileNet’s accuracy attained was 50%.

Advanced Algorithm Implementation:

To improve the accuracy rate to 85% on TRICKER images, first, VGG16 was further implemented into the system.

Visualization with GRADCAM:

GradCAM square images were created for remote sense indicating predictive heat mapping for the unaltered image and one that was tampered at face level. Analyzing the heatmaps showed underlying differences in the predictive patterns created.

Analysis of Misbehavior:

Cases were provided where the machine made errors when presented with made-up photos such as diagnosing 'Elton' as 'Madonna'. Instances of internal cooperation that resulted in bias during model training were also noted.

Results Summary:

Contrasted VGG16 with all other architectures in terms of their respective performances and said that the VGG16 model performed the best.

EVALUATION:

Precision:

$$\text{Formula: Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall (Sensitivity):

$$\text{Formula: Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

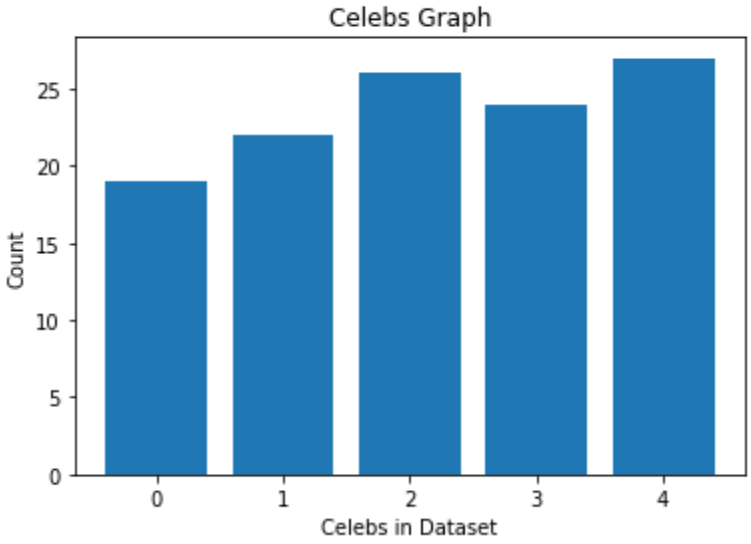
F1 Score:

$$\text{Formula: } F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

Accuracy:

$$\text{Formula: Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

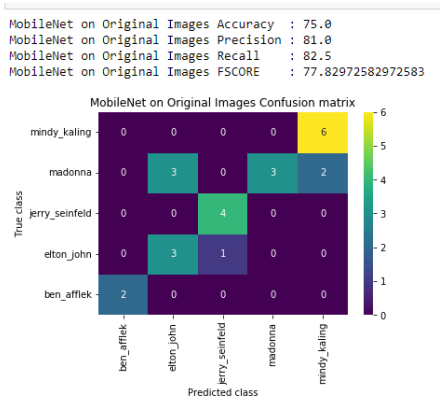
RESULTS:



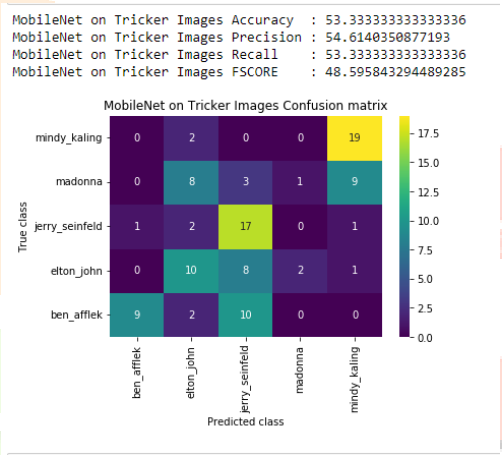
Celebrities and then in graph x-axis represents names of celebrity and y-axis represents count of each celebrity

Loaded celebrity genuine image

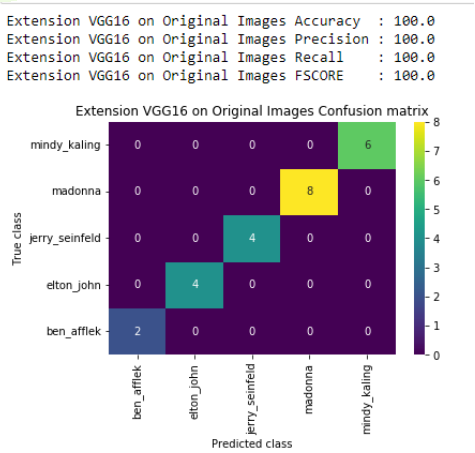
We have TRICKER above image from the original image and algorithm will predict above image wrongly which is just alteration of original image



In above screen with MOBILENET we got 75% accuracy on original images

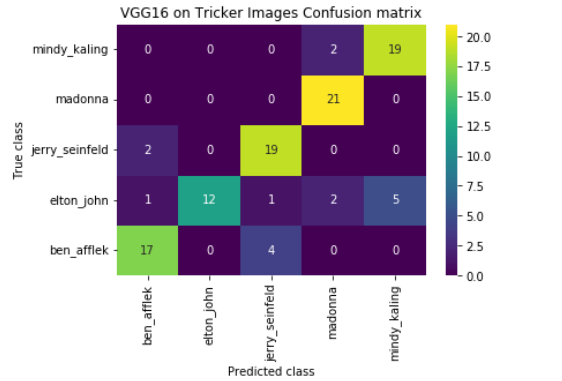


MOBILENET performance on TRICKER images and we got accuracy as 50% so this model predicting 50% wrong predictions for TRICKER images

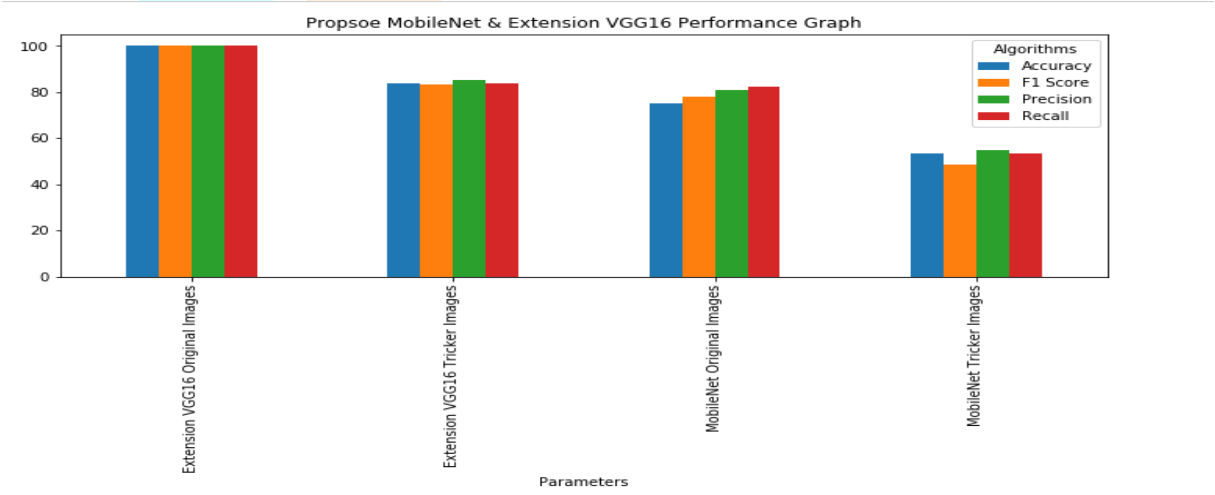


VGG we got 100% accuracy and in confusion matrix graph all incorrect blue colour boxes contains 0 so VGG predicted 0 incorrect prediction

```
VGG16 on Tricker Images Accuracy  : 83.80952380952381
VGG16 on Tricker Images Precision : 85.46666666666667
VGG16 on Tricker Images Recall    : 83.80952380952381
VGG16 on Tricker Images FSCORE    : 83.16946774210825
```



VGG on TRICKER images and we got 83% predictions are correct which means this model behaving maliciously only for 17% and behaving correctly for 83%.



In all algorithms extension VGG is giving high accuracy

	Algorithm Name	Accuracy	Recall	FScore	Precison
0	Propose MobileNet Original Images	75.000000	82.500000	77.829726	81.000000
1	Propose MobileNet Tricker Images	53.333333	53.333333	48.595843	54.614035
2	Extension VGG16 Original Images	100.000000	100.000000	100.000000	100.000000
3	Extension VGG16 Tricker Images	83.809524	83.809524	83.169468	85.466667

Displaying all algorithms performance

Prediction:

ELTON celebrity we did cross changes to images and algorithm has predicted as ‘Madonna’

Ben_afflek image after doing some changes and algorithm predicted as Jerry

CONCLUSION

This research warns that there are lethal weaknesses of face recognition models when facing alterations made maliciously on facial attributes. Additionally, the performance of ResNet and other models when faced with the TRICKER dataset paints a steeper accuracy decrease for the classical algorithms. However models like VGG16 perform considerably better for such tasks with an 85% accuracy on altered images. More information on the understanding of the manipulative parts and their consequences on prediction outcomes is provided by the addition of GRADCAM. This calls reflection on the development of strong, adversarial algorithms that will safeguard biometric systems from more sophisticated threats.

REFERENCES:

[1] J. Hilotin. "Use These Biometrics to Pass Through UAE Airports." 2019. [Online]. Available: https://gulfnews.com/uae/use-these-biometrics-topass-through-uae-airports-1.1570459646018 (Accessed: Mar. 30, 2021).

[2] D. Tinjaca. "Transforming Immigration and Border Crossing in Colombia With Automated Border Control." 2019. https://disblog.thalesgroup.com/corporate/2019/01/30/transforming-immigrationand-border-crossing-in-colombia-with-automated-border-control/ (Accessed: Mar. 30, 2021).

[3] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in Proc. 10th ACM Conf. Web Sci., New York, NY, USA, 2019, pp. 105–114.

[4] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically dismantling online dating fraud," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 1128–1137, 2019.

[5] K. Yuan et al., "Stealthy porn: Understanding real-world adversarial images for illicit online promotion," in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 547–561.

[6] M. Wang and W. Deng, "Deep face recognition: A survey," Neurocomputing, vol. 429, pp. 215–244, Mar. 2021.

[7] P. Vallée. "Why Biometrics Are the Foundation for the Airport of the Future." Jul. 2019. [Online]. Available: http://onboard.thalesgroup.com/why-biometrics-are-the-foundationfor-the-airport-of-the-future/ (Accessed: Jan. 16, 2021).

[8] M. Kendrick. "The Border Guards You Can'T Win Over With a Smile." Apr. 2019. [Online]. Available: https://www.bbc.com/future/article/20190416-the-ai-border-guardsyou-cant-reason-with (Accessed: Jan. 16, 2021).

[9] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2017, arXiv:1708.06733.

[10] Y. Wang, E. Sarkar, W. Li, M. Maniatakos, and S. E. Jabari, "Stop-andgo: Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems," IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4772–4787, 2021.

[11] E. Sarkar, Y. Alkindi, and M. Maniatakos, "Backdoor suppression in neural networks using input fuzzing and majority voting," IEEE Des. Test., vol. 37, no. 2, pp. 103–110, Apr. 2020.

[12] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, arXiv:1712.05526.

[13] Y. Liu et al., "Trojaning attack on neural networks," in Proc. Symp. Netw. Distrib. Syst. Secur., 2018.

[14] B. Chen et al., "Detecting backdoor attacks on deep neural networks by activation clustering," 2018, arXiv:1811.03728.

[15] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in Proc. 31st Adv. Neural Inf. Process. Syst. , 2018, pp. 8000–8010.

[16] B. Wang et al., "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in Proc. IEEE Symp. Secur. Privacy (IEEE S&P), San Francisco, CA, USA, 2019, pp. 707–723.

[17] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "ABS: Scanning neural networks for back-doors by artificial brain stimulation," in Proc. 2019 ACM SIGSAC Conf. Comput. Commun. Secur., 2019, pp. 1265–1282.

[18] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 6206–6215.

[19] C. He, M. Xue, J. Wang, and W. Liu, "Embedding backdoors as the facial features: Invisible backdoor attacks against face recognition systems," in Proc. ACM Turing Celebration Conf. China, New York, NY, USA, 2020, pp. 231–235.

[20] FaceApp Technology Limited. "FaceApp." [Online]. Available: https://www.faceapp.com/ (Accessed: Aug. 20, 2021).