Evaluation of Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) Based Attack Method on MCS 2018 Adversarial Attacks on Black Box Face Recognition System

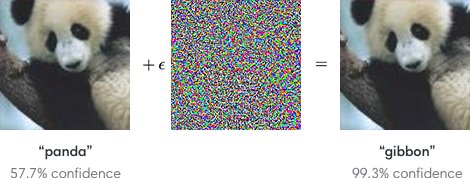
**Md Ashraful Alam Milton**1

1Department of Computer Engineering, Autonomous University of Barcelona, Spain

**Abstract.** The convolutional neural network is the crucial tool for the recent success of deep learning based methods on various computer vision tasks like classification, segmentation, and detection. Convolutional neural networks brought state of the art result in those tasks and every day pushing the limit of computer vision and AI. However, adversarial attack on these computer vision tasks is threatening their application in the real work and safety-critical applications. Finding adversarial examples are important to find susceptible models and take safeguards methods to overcome the adversarial attacks. In this regards, MCS 2018 Adversarial Attacks on Black Box Face Recognition challenge aims to facilitate the research of finding techniques and their effectiveness in generating adversarial examples. In this challenge, the attack is targeted attack on the black-box neural network where we have no knowledge about its inner structure. The attacker must modify a set of five images of a single person so that the neural network miss classify them as target image which is a set of five images of another person. In this competition, we applied Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) to make an adversarial attack on black-box face recognition system. We tested our method on MCS 2018 Adversarial Attacks on Black Box Face Recognition challenge and found a competitive result. Our solution got validation 1.404 scores which better than baseline and stands 14 place among 132 teams in the leaderboard. Further improvement can be achieved by finding improved feature extraction from source image, carefully chosen hyperparameters, finding improved substitute model of the black-box and better optimization method. Code is available in https://github.com/miltonbd/mcs\_2018\_adversarial\_attack

1. Introduction

After the last AI winter, since 2012 deep learning based neural networks are state of the art technology for computer vision and achieve an unprecedented result on various vision tasks, including image classification [2,7,8,9], object detection [10,11,12,13] and semantic segmentation [14,15,16].However, due CNN's internal structure and way of work they are can be attacked with adversarial images with small modifications [32] which is impossible to differentiate by the human eye. Deep neural networks (DNNs) can be compromised by their inability to tackle adversarial examples [4, 5], which carefully devised by adding small, human-imperceptible noises to legitimate examples. This attack is successful in both targeted where the final in-accurate output set by the attacker and no-targeted attack where the model’s output can be any of inaccurate predictions. Arguably, the time has come to take safeguard measures and incorporate built net defenses in real-world machine learning system like the autonomous vehicle, drone vision, face recognition etc. systems.

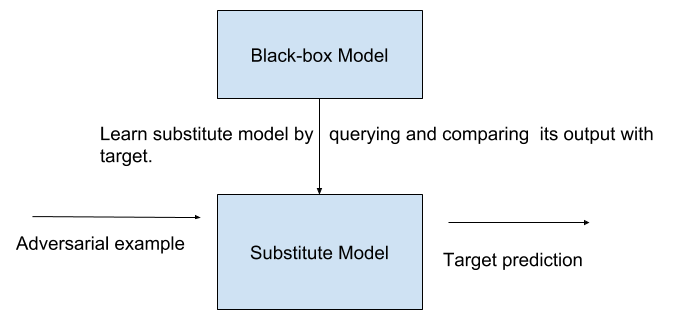


**Fig 1. Adversarial Example of the panda with added perturbation.**

Adversarial attack generating suitable images can be employed to a different domain like image segmentation[24]. Several studies focus on and understand the insufficiency of current training algorithms [6,25,26] against adversarial examples. Many important methods [5,26, 27] have been developed recently to find adversarial examples.Adversarial examples can be a severe security threat for practical computer vision applications. In particular, an adversarial example[4] that was designed to be misclassified by a model A is often also misclassified by a model B. This adversarial example transferability property means that it is possible to generate adversarial examples and perform a misclassification attack on a machine learning system without access to the underlying model. Such attacks can be done in the real world [29,30].Data augmentation [7,8,9] has been proved to be a viable solution to prevent networks from overfitting during the training process. Clearly, a series of transformation like resizing, cropping and rotating, are applied to the images to enhance the training set. As a result, the trained networks will be robust unknown input and will generalize well. Also, image augmentation[31,32] can defend against adversarial examples under certain situations, which proves that adversarial examples cannot generalize well under different transformations.To this extent, we applied Diverse Input Iterative Fast Gradient Sign Method (DI 2 -FGSM) in Using Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) Based Attack on MCS 2018 Adversarial Attacks on Black Box Face Recognition to improve the transferability of adversarial examples. At each iteration, we diverse the inputs by applying augmentation using imgaug[1] library. This way the model has always some modified source image which reduces overfitting. This method performed better than the baseline method.

1. Background And Related work

In this section, we will review the background information and related works about the adversarial attack. The main goal of the adversarial attack is to fabricate a new image by adding carefully controlled noise to the original image in such a way that the changes are almost undetectable to the human eye. The modified image is called an adversarial image, and when submitted to a classifier is misclassified, while the original one is correctly classified. The real-life applications of such attacks can be very serious –for instance, one could modify a traffic sign to be misinterpreted by an autonomous vehicle, and cause an accident. Another example is the potential risk of inappropriate or illegal content being modified so that it is undetectable by the content moderation algorithms used in popular websites or by police web crawlers. For that reason, we’re very interested in understanding these attacks and developing our own defenses against them. Some definitions An adversarial image is an image that has been slightly modified in order to fool the classifier, i.e., in order to be misclassified.



**Fig 2. Black-box and its substitute model.**

The measure of modification is normally the ℓ∞ norm, which measures the maximum absolute change in a single pixel. In white box attacks the attacker has access to the model’s parameters, while in black box attacks, the attacker has no access to these parameters. The goal of non-targeted attacks is to compel the model to misclassify the adversarial image and in the targeted attack, the goal is to misclassify the source image a fixed wrong class. Most prominent attacks are gradient-based attacks where the attacker modify the source image by observing the flow of gradient. In this way, the attackers modify the image in the direction of the gradient of the loss function with respect to the input image. one-shot attacks and iterative attacks are popular in gradient-based attacks. in one-shot attack, where the attacker takes a single step in the direction of the gradient, and in the iterative attacks where instead of a single step, several steps are taken.

1. Attack Methods

In recent years, extensive research brought about a variety of gradient based adversarial attack methods. The evolution of gradient-based attack are in following sections.

* 1. Fast Gradient Sign Method (FGSM)

FGSM [5] is the first member in this attack family, which finds the adversarial perturbations in the direction of the loss gradient and the adversarial example image generation equation can be expressed as:

|  |  |
| --- | --- |
|  | ………………..(1) |

* 1. Iterative Fast Gradient Sign Method (I-FGSM)

Kurakin et al. [27] extended FGSM to an iterative version, which can be expressed as:

|  |  |
| --- | --- |
|  | ………………..(2) |

the original image X, i is the iteration number and α is the step size.

* 1. Momentum Iterative Fast Gradient Sign Method (MI-FGSM)

MI-FGSM proposed to integrate the momentum term into the attack process to stabilize update directions and escape from poor local maxima. The updating procedure is similar to I-FGSM,

|  |  |
| --- | --- |
|  | ………………..(3) |
|  |

where μ is the decay factor of the momentum term and gn is the accumulatedgradient at iteration.

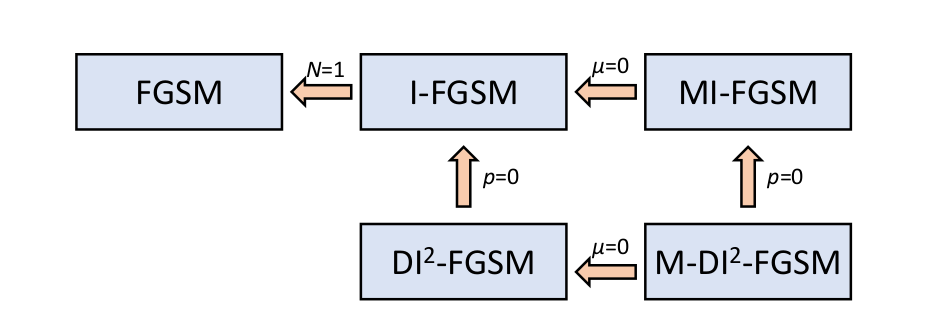
* 1. Diverse Input Iterative Fast Gradient Sign Method (DI2-FGSM)

In diverse input, various changes are applied to source image like random crop, gaussian blur with Sigma 0.5, contrast normalization, affine transformation with scaling, transformation, rotate and shear.

|  |  |
| --- | --- |
|  | ………………..(4)  ………………..(5) |
|  |

momentum and diverse inputs work differently to prevent overfitting phenomenon. Combination both known as Momentum Diverse Inputs Iterative Fast Gradient Sign Method (M-DI 2 -FGSM) has good potential as a good adversarial attack method. The overall updating procedure of M-DI 2 -FGSM is similar to MI-FGSM :

|  |  |
| --- | --- |
|  | ………………..(6) |



**Fig 3. The relationship among fast gradient step method.**

If the transformation[36] probability is p = 0, M-DI 2 -FGSM degrades to MI-FGSM, and DI-2 -FGSM degrades to I-FGSM; If the decay factor μ = 0, M-DI 2 -FGSM degrades to DI 2 -FGSM, and MI-FGSM degrades to I-FGSM, If the total iteration number N = 1, I-FGSM degrades to FGSM.In all above equations + sign before the α indicates that the attack is non-targeted and for the targeted attacks the α sign will be negative.

1. Challenge Details

Faces are an important natural way to recognize people. It is the extension of general face detection to provide recognition of an individual. Recent deep learning and computer vision technology enables to scale up face recognition to correctly recognize millions of identity. Modern methods of face recognition easily surpass human performance while relying on machine learning and neural networks based techniques. This type of face recognition system will enable modern biometric system and other real-world applications. However, such face recognition system can be vulnerable to attacks aiming to fabricate the network’s final output. Seemingly, arbitrary changes to the network output can be produced by small and well-designed modifications of the network input, as known under adversarial examples. Applied to face recognition, adversarial examples imply that an attack can be designed to force a network to identify a person in the original image (Im1) as any other person on the planet (e.g. Im2) by making small modifications to the original image G(Im1) as shown below.



Fig 4. MCS 2018 Adversarial Attacks on Black Box Face Recognition.

While attacks on open networks are relatively straightforward, the design of attacks on “black-box” systems, e.g. networks with unknown structure and parameters, is more difficult.”Adversarial Attacks on Black-box Face Recognition” aims to test the vulnerability of black-box face recognition systems. Participants are asked to design the best attack forcing the system to recognize an image of a person A as person B, by applying small modifications to images of A.

1. Proposed Methodology
   1. Make Substitute Model

we used densenet based methods to make a substitute classification model by feeding the output of the model to our system as input and using our know labels of those inputs. We used 1M face recognition data provided by the organizer to train the classifier. MEAN = [0.485, 0.456, 0.406] STD = [0.229, 0.224, 0.225] used for normalizing the input data. Black-box attacks and substitute models While the definition of an open-box (white-box) attack to DNNs is clear and precise - having complete knowledge and allowing full access to a targeted DNN, the definition of a “black-box” attack[28] to DNNs may vary in terms of the capability of an attacker. In an attacker’s perspective, a black-box attack may refer to the most challenging case where only benign images and their class labels are given, but the targeted DNN is completely unknown, and one is prohibited from querying any information from the targeted classifier for adversarial attacks.

* 1. Fine Tuning

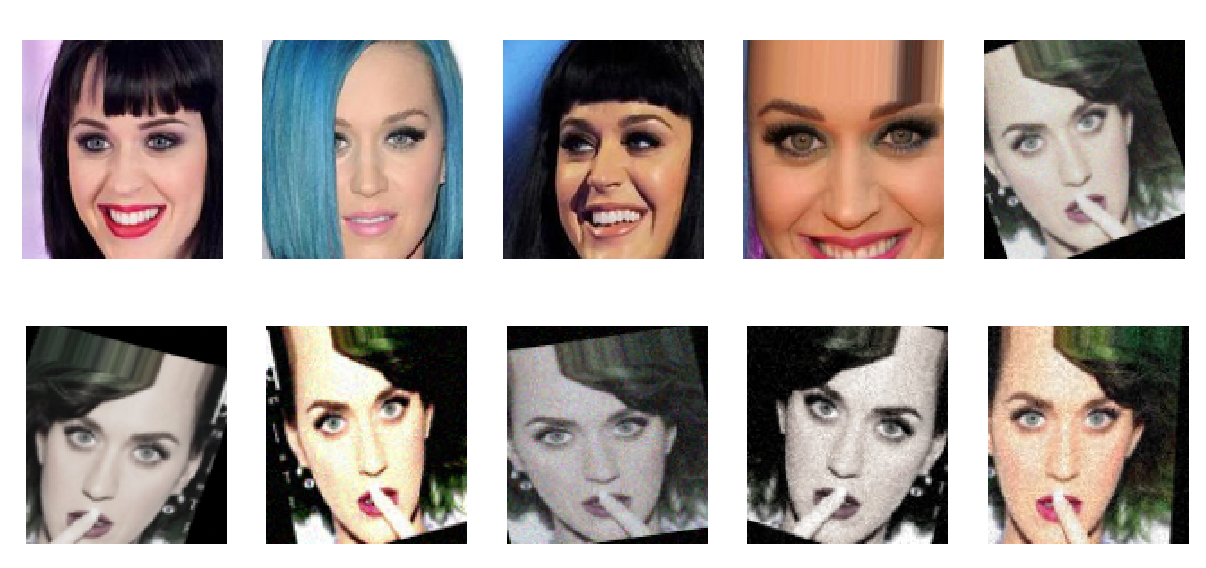
As the amounts of training examples is insufficient for training a deep convolutional network from scratch, so we used the imagenet pre-trained DENSENET model to initialize the network parameters, freeze the initial layers and fine-tuned the last several layers. We kept the weight of all layers except FC and output layer was frozen in first few iterations of training due to high unstable gradient flow. After few iterations of training, we unfreeze the last few layers and adjust their weight by backpropagation. Fine-tuning helps to train easily and prevents it from overfitting. obviously, the fine-tuned network can lead to a better convergence.

* 1. Pre-Processing

First the images is center cropped from 224 to 112. Transforms. The source image pixels are normalized using MEAN [0.485, 0.456, 0.406] and STD = [0.229, 0.224, 0.225].

* 1. Diverse Input

we are given with 5 source images and 5 target images. we keep the target images same without augmentation and to satisfy the diverse input feature of the algorithm we augmented the 5 source face images to 20.



**Fig 4.** **Augmented images as diverse input.**

Data augmentation methods used are the random crop, gaussian blur with Sigma 0.5, contrast normalization, affine transformation with scaling, transformation, rotate and shear. We also apply grayscale for some images. All augmentation is done using imgaug[1]. The input images with probability p will be going through the diverse input transformation and 1-p will be feed to the model as usual.

* 1. Attack Details

we run the attack for maximum 60 epochs. For some images, we were having too much too little gradient propagation and sometimes it was found that after all iter the SSIM is still 1 and sometimes after 1 iter it goes below 0.95. In both cases, we adjusted the epsilon with either multiplying or dividing the initial eps.

1. Evaluation

To evaluate model performance results, we fed the 1000 images to substitute model and generate the adversarial example. We also extracted a 1000\*512 descriptor to be uploaded to evaluation serve as per directed by the organizer. For the evaluation of the result, we have to upload the generated adversarial images and 1000\*512 face descriptor in a single zip file. The Structural Similarity Index (SSIM) was for measuring original and generated image difference and that must be below above .95 threshold.

1. Implementation Details

For training purposes, PyTorch[21] we used 2 1080TI GPU based pc with CUDA 9 and Ubuntu 16.04 OS. This computing power was not enough to get the result quickly, hence lowered the chance of ace the competition leaderboard.

1. Results and Discussion

Our Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) achieved 1.404 in the final leaderboard which is better than baseline value of 1.407. Due to late participation, the limited deadline we did get to the very top in the competition leaderboard and we believe there is ample opportunity to extend the attack method with more improved feature extraction from source image, carefully chosen hyperparameters, finding improved substitute model of the black-box. In addition to that, a bigger dataset with improved image augmentation reduces the risk of overfitting on the black-box model and thus, substitute model mimics the black-box network well. Moreover, performing additional regularization tweaks and fine-tuning of hyperparameters may improve model’s robustness. Along with densenet to make substitute model of black-box, other network models like SENet[17], NASNet[18], PNASNet[19] might be used.

1. Conclusion

In this paper, we demonstrated that Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) can be effectively applied to adversarial attack on face recognition system. This could elicit a high opportunity to test the face recognition system against this robust adversarial attack and pave the way to take measure against adversarial attack. Adversarial training, input sanitization are important measures so that the face recognition system has already taken care of the adversarial sample of input images.

1. References

[1] https://github.com/aleju/imgaug

[2] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, Densely Connected Convolutional Networks, https://arxiv.org/abs/1608.06993

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li, Boosting Adversarial Attacks with Momentum https://arxiv.org/abs/1710.06081

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan,I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In ICLR, 2014. 1, 3

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015. 1, 2, 3

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015. 1, 2, 3

[7] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)

[8] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale im- age recognition. In: International Conference on Learning Representations (2015)

[9] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer (2016)

[10] Girshick, R.: Fast r-cnn. In: International Conference on Computer Vision. IEEE (2015)

[11] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detec- tion with region proposal networks. In: Advances in Neural Information Processing Systems (2015)

[12] Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., Yuille, A.L.: Single-shot object

detection with enriched semantics. arXiv preprint arXiv:1712.00433 (2017)

[13] Joseph Redmon, Ali Farhadi,YOLO9000: Better, Faster, Stronger, https://arxiv.org/abs/1612.08242

[14] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Computer Vision and Pattern Recognition. IEEE (2015)

[15] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Se- mantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelli- gence (2017)

[16] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, https://arxiv.org/abs/1505.04597

[17] Jie Hu, Li Shen, Gang Sun,Squeeze-and-Excitation Networks https://arxiv.org/abs/1709.01507

[18]Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le, Learning Transferable Architectures for Scalable Image Recognition, https://arxiv.org/abs/1707.07012

[19] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, Kevin Murphy, Progressive Neural Architecture Search https://arxiv.org/abs/1712.00559

[20] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)

[21]MCS 2018. Adversarial Attacks on Black Box Face Recognition, https://competitions.codalab.org/competitions/19090

[22]https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7

[23]PyTorch, https://pytorch.org/

[24]Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. arXiv preprint arXiv:1711.09856 (2017)

[25]Lin, Y.C., Hong, Z.W., Liao, Y.H., Shih, M.L., Liu, M.Y., Sun, M.: Tactics of adversarial attack on deep reinforcement learning agents. In: International Joint Conference on Artificial Intelligence. AAAI (2017)

[26]Tram`er, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)

[27] Kurakin,A.,Goodfellow,I.,Bengio,S.:Adversarial examples in the physical world. In: International Conference on Learning Representations Workshop (2017)

[28] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697, 2016.

[29] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. ArXiv e-prints, May 2016b. URL http://arxiv.org/abs/1605.07277.

[30] Nicolas Papernot, Patrick Drew McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. CoRR, abs/1602.02697, 2016a. URL http://arxiv.org/abs/1602.02697

[31] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial Examples

for Semantic Segmentation and Object Detection. In: International Conference on

Computer Vision. IEEE (2017)

[32] Guo, C., Rana, M., Ciss ́e, M., van der Maaten, L.: Countering adversarial images using input transformations. In: International Conference on Learning Representations (2018).

[33] Dong, Y., Liao, F., Pang, T., Su, H., Hu, X., Li, J., Zhu, J.: Boosting adversarial attacks with momentum. arXiv preprint arXiv:1710.06081 (2017)

https://github.com/miltonbd/mcs\_2018\_adversarial\_attack

[35] Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects

through randomization. In: International Conference on Learning Representations

(2018)

[36] Cihang Xie , Zhishuai Zhang , Jianyu Wang, Yuyin Zhou, Zhou Ren , Alan Yuille, Improving Transferability of Adversarial Examples with Input Diversity,https://arxiv.org/abs/1803.06978