On Robustness of Deep Neural Networks: A Comprehensive Study on the Effect of Architecture and Weight Initialization to Susceptibility and Transferability of Adversarial Attacks

Ibrahim Ben Daya, University of Waterloo, ON, Canada
Mohammad Javad Shafiee, University of Waterloo, ON, Canada
Michelle Karg, Continental Automotive, Germany
Christian Scharfenberger, Continental Automotive, Germany
Alexander Wong, University of Waterloo, ON, Canada

Abstract

Neural network models have shown state of the art performance in several applications. However it has been observed that they are susceptible to adversarial attacks: small perturbations to the input that fool a network model into mislabelling the input data. These attacks can also transfer from one network model to another, which raises concerns over their applicability, particularly when there are privacy and security risks involved. In this work, we conduct a study to analyze the effect of network architectures and weight initialization on the robustness of individual network models as well as transferability of adversarial attacks. Experimental results demonstrate that while weight initialization has no affect on the robustness of a network model, it does have an affect on attack transferability to a network model. Results also show that the complexity of a network model as indicated by the total number of parameters and MAC number is not indicative of a network's robustness to attack or transferability, but accuracy can be; within the same architecture, higher accuracy usually indicates a more robust network, but across architectures there is no strong link between accuracy and robustness.

1 Introduction

The field of computer vision has seen major breakthroughs in recent years thanks to deep learning, with many network models for recognition tasks such as classification and segmentation now being deployed in more and more complex systems [1]. Deep neural network models have shown phenomenal success in solving complex problems. However, it has been demonstrated that they are vulnerable to adversarial attacks [2]; malicious perturbations - often small enough to be imperceptible to the human eye - that have been optimized to fool network models into making mistakes. This raises concerns, particularly when networks are deployed in safety-critical systems like autonomous driving. While most of these attacks require explicit knowledge of the underlying network model (white-box attacks), several works have shown that adversarial examples generated for one network model may successfully attack another; a property referred to as *transferability* which may be leveraged to perform black-box attacks [4].

A few studies in literature have looked into the transferability of adversarial examples. Szegedy et al. [2] were the first to examine this, their work looked into the transferability between different network models trained over MNIST [5]. Goodfellow et al. [3] followed that study on MNIST and CIFAR-10 [6]. Papernot et al. [8] looked into the feasability of constructing a substitute network model to attack a black-box target network model. Liu et al. [4] expanded that study to include larger network models and a larger dataset. All studies showed that it is possible for adversarial attacks to transfer across network models. Moosavi-Dezfooli et al. [9] showed that a universal perturbation for each network model exists which can transfer across images. This encouraged a steadily growing research into adversarial defence.

Many defences to adversarial attacks have been proposed along three main research directions: 1) using modified training/input, 2) modifying networks, and 3) using external models as network addons [1]. However, we've yet to see a satisfactory defence to these attacks. It is often shown that counter-counter measures can be devised to successfully attack a defended network model [10, 11], with Akhtar and Mian [1] noting that this should encourage new defence methods to provide an estimate for their robustness to obvious counter-counter measures.

Inherent network model robustness has been previously studied. Rozsa et al. [12] empirically analyzed the correlation between network model accuracy and robustness, observing that network models with higher accuracy generally exhibit more robustness to adversarial examples. The study covered eight network models trained on ImageNet under three attacks. Madry et al. [13] studied the effect of network model capacity on adversarial robustness, observing that increasing the capacity of the network model increases its robustness and decrease attack transferability. A similar conclusion was reached by Kurakin et al. [14]. Madry's study covered a network model trained on MNIST and another trained on CIFAR-10, while Kurakin's covered a network model on ImageNet.

In this research, we expand on previous studies to look into the effect of network model initialization and architecture on robustness and transferability. We cover seven network model architectures, with three different weight initializations, on MNIST and CIFAR-10. The rest of the paper is divided as follows: in section 2, we outline the parameters of the study. Section 3 details our experimental results and discussion. We conclude the paper in section 4.

2 Methodology

In this section, we first describe the adversarial attacks used in this study. Next, we list the models and datasets used. Finally, we define the measures we use to quantify robustness and transferability.

2.1 Adversarial Attacks

Adversarial attacks consist of perturbations added to the original input to a network model in order to change its original prediction. In this study, we look into three adversarial attacks: fast gradient sign method (FGSM) [3] and its extension iterative FGSM [14, 13] as well as one-pixel attacks [15].

Fast Gradient Sign Method (FGSM) [3] is a single-step white-box attack that uses the loss of the network as a perturbation to input x:

$$x^{adv} = x + \varepsilon \cdot sign\big(\nabla_x L\big(f(x;\theta), y\big)\big) \tag{1}$$

where x^{adv} is the adversarial image, x the original image, ε is a small scalar that restricts the norm of the perturbation, $sign(.)$ is the sign function, ∇ computes the gradient of the loss function $L(.)$ between the network prediction y under model parameters θ.

Iterative FGSM [14, 13] relies on taking multiple smaller steps instead of taking on big step like in FGSM. This increases the chance of fooling the original network.

One-Pixel Attack [15] is a black-box attack where only one pixel in the image is changed to fool the network model. Adversarial examples are computed using Differential Evolution [16], where an initial set of candidate pixel population is modified to create children that compete with its parents for fitness in the next iteration. This method makes it possible to generate adversarial examples with only probabilistic labels of the targeted network model as input.

2.2 Models and Datasets

In this study, we use a variety of network models ranging from complex network models to lightweight network models suitable for embedded applications. We used publicly available code for model definitions. Models trained include AlexNet [17], GoogLeNet [18], LeNet, LeNet-5 [5], Network in Network (NIN) [19], ResNet-20 [20], SqueezeNet [21], and VGG-16 [22]. Three different initializations were studied: Gaussian, MSRA, and Xavier. Whenever possible, we trained each network model on all three initializations. However, some network models wouldn't converge or had poor accuracies, and were therefore not included. Tables 1 and 2 list all networks in this study and their accuracy on the corresponding test set.

For both MNIST and CIFAR-10, a study set of 1000 randomly selected images were chosen to test robustness and transferability. Since it is less meaningful to examine transferability of an adversarial image when models don't classify the input correctly, we ensured that the randomly chosen images are given the correct label by all network models.

2.3 Quantitative Measures

In this study, we look into individual network robustness to adversarial attacks as well as the transferability of attacks from one network to another. We also include a measure of network model complexity to look for a link between complexity and robustness.

To quantify the robustness of individual networks, **attack success rate** is used. This is simply the percentage of adversarial example that successfully fool a network model. A lower success rate indicates a network is more robust. To quantify a measure of

Arch	Params	MAC	Gaussian		MSRA		Xavier	
			ID	Accuracy	ID	Accuracy	ID	Accuracy
AlexNet	3.8500	0.5959	0	0.9391	1	0.9939	2	0.9936
GoogLeNet	6.4000	0.0170	3	0.9753	4	0.9917	5	0.9910
LeNet	0.4311	0.0023	6	0.9615	7	0.9911	8	0.9904
NIN	6.5500	0.2032	9	0.9738	10	0.9885	11	0.9918
ResNet20	0.2791	0.0316	12	0.9422	13	0.9717	14	0.9880
SqueezeNet	0.7265	0.0296	15	0.9910	16	0.8997	17	0.9882

Table 1: MNIST network models used in this study, with the corresponding number of (in millions) and MAC operations (in billions), accuracy of each initialization, and the network model ID for this study listed.

Arch	Params	MAC	Gaussian		MSRA		Xavier	
			ID	Accuracy	ID	Accuracy	ID	Accuracy
AlexNet	4.2500	1.0100	0	0.7900	1	0.8040	2	0.7731
GoogLeNet	6.4000	0.0328	X	X	X	X	3	0.7517
LeNet-5	0.0896	0.0123	4	0.7960	5	0.6429	6	0.6581
NIN	6.5500	0.2564	7	0.8388	X	X	X	X
SqueezeNet	0.7276	0.0411	8	0.8042	X	X	X	X
VGG-16	33.6400	0.3321	X	X	9	0.6581	10	0.7010

Table 2: CIFAR-10 network models used in this study, with the corresponding number of paramters (in millions) and MAC operations (in billions), accuracy of each initialization, and the network model ID for this study listed. Network models with initializations that wouldn't allow it to train (or would train with very poor accuracy) are marked by 'X'.

transferability between two models, we compute the percentage of adversarial examples generated for one network that successfully fool the other network, henceforth reffered to as **transfer rate**. Network complexity is set by the total number of parameters as well as the number of multiply-accumlate (MAC) operations.

3 Experimental Results and Discussion

This section details our experimental results and provides some discussion. We first discuss FGSM attack on MNIST and CIFAR10. In the figures detailing the results, the labels for rows correspond to the network model used to generate the adversarial example, and the labels for columns correspond to the network model the generated adversarial example is transfered to. The shade of each element corresponds to the transfer rate, while the diagonal is the attack success rate.

3.1 FGSM and Iterative FGSM Attacks

MNIST: Figure 1 summarizes FGSM and iterative FGSM attacks on the MNIST study set under three noise levels: $\varepsilon = 0.1$, 0.2, and 0.3. The transfer rate and attack success rate go up as noise level increases, and with iterative FGSM, attack success rate is higher than FGSM while transferability is lower. AlexNet, NIN, and SqueezeNet seem to be the most robust, while ResNet20 seems to be the most susciptable to both attacks. One interesting observation is the Gaussian initialized AlexNet and LeNet are inherently robust, with attack success rate close to zero for LeNet and zero for AlexNet. We have no explanation as to why this is the case.

CIFAR10: Figure 2 summarize FGSM and iterative FGSM attacks on the CIFAR10 study set under three different noise levels: $\varepsilon = 2$, 4, and 8. The transfer rate and attack success rate go up as noise level increases, and with iterative FGSM, attack success rate is higher than FGSM. Unlike MNIST tests, transfer rate is higher with iterative FGSM. NIN and SqueezeNet seem to be the most robust to the two attacks, while LeNet-5 seems to be the most susciptible. There seems to be no noticable difference made by initialization and robustness to FGSM and iterative FGSM, with all AlexNets, LeNet-5s, and VGG-16s showing similar transfer rates and attack success rates.

3.2 One Pixel Attack

MNIST: Figure 3 summarizes one pixel attack on the MNIST study set. This attack had very low success overall on MNIST images, highest being 3.5% success rate on Gaussian initialized LeNet. While there aren't enough successfull atacks to make any solid conclusions, it is interesting to observe that the two most robust networks to FGSM and iterative FGSM (Gaussian initialized AlexNet and LeNet) have the highest transfer rates. However, we do observe that examples generated from all AlexNets and the NIN have higher transfer rate to other networks.

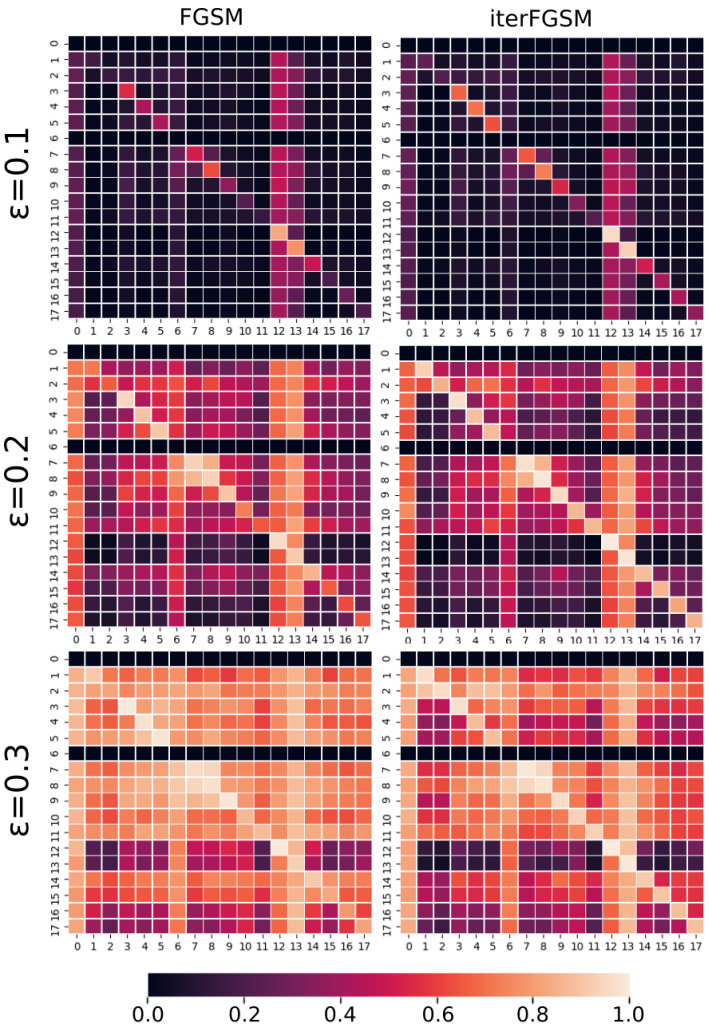

Fig. 1: Summary of MNIST FGSM (left) and iterative FGSM (right) attacks at low (top), medium (middle), and high (bottom) FGSM noise levels. Rows represent network models used to generate adversarial attacks, columns are network models generated attacks are transferred to. The diagonal shows attack success rate, all other elements show transfer rate.

CIFAR10: Figure 3 summarizes one pixel attack on the CIFAR10 study set. This attack had moderate success when compared to FGSM and iterative FGSM, with the highest success rate of 48% seen on Xavier initialized SqueezeNet, MSRA initialized VGG-16 being very susciptible as well. Gaussian initialized AlexNet seems to be the most robust; both in terms of attack success rate as well as transfer rate. Xavier initialized SqueezeNet seems to have the highest transfer rate.

3.3 Discussion

For FGSM attacks, it seems that weight initialization mostly has no significant affect on attack success rate, with architecture providing the most significant variance. This true for both MNIST and CIFAR10 study sets. The only exceptions are the Gaussian initialized AlexNet and LeNet, with the attack success rate being unusually low. Weight initialization does seem to have an affect on transferability, with Gaussian initialized network models having generally higher transfer rates, and Xavier initialized networks having generally lower transfer rates. Similar observations can be made for iterative FGSM. For CIFAR-10, it seems that all 3 AlexNets generate adversarial examples that have a high transfer rate to other network models when compared to adversarial examples generated from other models.

For one pixel attacks, there is no discernable pattern for attack success rate and weight initialization, or chosen architecture and attack success rate. There is some pattern with Gaussian initialization and transfer rate in MNIST for AlexNet, GoogLeNet, and LeNet. But it seems there are other factors contributing to the inherent network model susciptibility to one pixel attack.

Looking at tables 1 and 2, there is no clear link between network complexity and inherent robustness to any of the three attacks. For example, Gaussian initialized LeNet-5 and SqueezeNet - two relatively simple networks with low number of parameters and

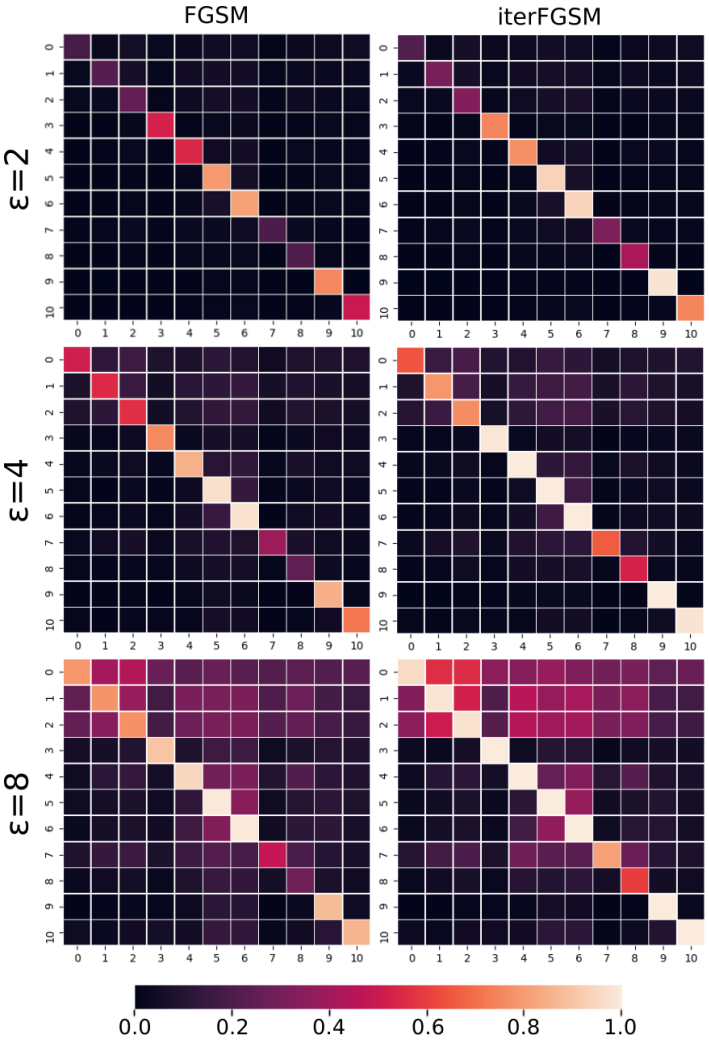

Fig. 2: Summary of CIFAR-10 FGSM (left) and iterative FGSM (right) attacks at low (top), medium (middle), and high (bottom) FGSM noise levels. Rows represent network models used to generate adversarial attacks, columns are network models generated attacks are transferred to. The diagonal shows attack success rate, all other elements show transfer rate.

MAC operations - on CIFAR10 had varying robustness to one pixel attack, with attack success rate being 20.1% and 48.5% respectively. Xavier initialized NIN and Gaussian initialized GoogLeNet - two networks with relatively high number of parameters - on MNIST had varying robustness to MNIST, with attack success rate at $\varepsilon = 0.1$ being 14.6% and 54.9% respectively. There is some correlation between accuracy and robustness observed, but only within the same architecture. For example: Gaussian initialized AlexNet and LeNet-5 both have a similar accuracy at 79%, but had iterative FGSM success rate at $\varepsilon = 2$ is 20.6% and 77.2% respectively, while AlexNets, LeNet-5s, and VGG-16s had a clear trend where higher accuracy meant more robust network models.

4 Conclusion

In this study, we looked into the affect of weight initialization and architecture on network model robustness. Our study spanned 18 networks trained on MNIST and 11 trained on CIFAR10, three different weight initializations, and three adversarial attacks. We observed that while initialization does not have a discernible affect on success rate of an attack, it does affect the transferability of an attack from one network to another. We also observed that the network complexity given by the total number of parameters and MAC is not indicative of a network model's success rate. Finally, we observe that across the same architecture, higher accuracy indicates a more robust network model, but that is not the case when looking at networks of different architectures.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Research Chairs Program, and Continental Automotive.

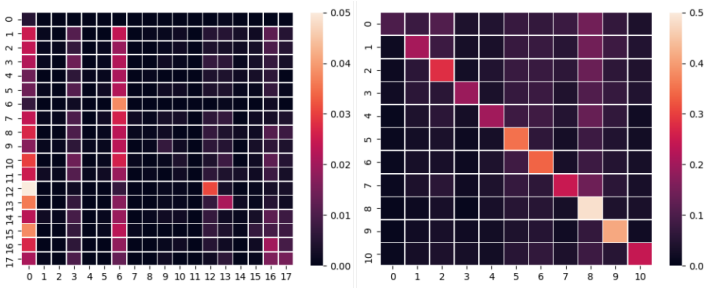

Fig. 3: Summary of one pixel attack on MNIST (left) and CIFAR-10 (right). Rows represent network models used to generate adversarial attacks, columns are network models generated attacks are transferred to. The diagonal shows attack success rate, all other elements show transfer rate.

References

[1] Akhtar, N. and Mian, A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access* vol. 6, pp. 14410-14430 (2018).

[2] Szegedy, C. et al. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199* (2013).

[3] Goodfellow, I. Shlens, J. and Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572* (2014).

[4] Liu, Y. Chen, X. Liu, C. and Song, D. Delving into Transferable Adversarial Examples and Black-Box Attacks. *arXiv preprint arXiv:1611.02770* (2017).

[5] LeCun, Y. Bottou, L. Bengio, Y. and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278-2324, (1998).

[6] Krizhevsky, A. and Hinton, G. Learning Multiple Layers of Features from Tiny Images (2009).

[7] Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211-252 (2015).

[8] Papernot, N. et al. Practical Black-Box Attacks Against Deep Learning Systems Using Adversarial Examples. *arXiv preprint arXiv:1602.02697* (2016).

[9] Dezfooli, S. Fawzi, A. Fawzi, O. and Frossard, P. Universal Adversarial Perturbations. *arXiv preprint arXiv:1610.08401* (2016).

[10] Carlini, N. and Wagner, D. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods *arXiv preprint arXiv:1705.07263* (2017).

[11] Athalye, A. and Carlini, N. On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses. *arXiv preprint arXiv:1804.03286* (2018).

[12] Rozsa, A. Gunther, M. and Boult, T. Are Accuracy and Robustness Correlated? *arXiv preprint arXiv:1610.04563* (2016).

[13] Madry, A. et al. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083* (2017).

[14] Kurakin, A. Goodfellow, I. and Bengio, S. Adversarial Machine Learning at Scale. *arXiv preprint arXiv:1611.01236* (2017).

[15] Su, J. Vargas, D. and Kouichi, S. One Pixel Attack for Fooling Deep Neural Networks. *arXiv preprint arXiv:1710.08864* (2017).

[16] Das, S. and Suganthan, P. Differential Evolution: A Survey of the State-of-the-Art. *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 4-31 (2011).

[17] Krizhevsky, A. Sutskever, I. and Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol 1 (2012).

[18] Szegedy, C. et al. Going Deeper with Convolutions. *arXiv preprint arXiv:1409.4842* (2014).

[19] Lin M. Chen, Q. and Yan, S. Network In Network. *arXiv preprint arXiv:1312.4400* (2014).

[20] He, K. Zhang, X. Ren, S. and Sun, J. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition* (2016).

[21] Iandola, F. et al. SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5MB Model Size. *arXiv preprint arXiv:1602.07360v4* (2016).

[22] Simmonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large Scale Image Recognition. *International Conference on Learning Representation* (2015).

[23] Wang, Z. Bovik, A. Sheikh, H. and Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612 (2004).