Generating Adversarial Perturbation with Root Mean Square Gradient

Yatie Xiao, Chi-Man Pun, Jizhe Zhou
Department of Computer and Information Science
University of Macau, Macau, China
ytxiao18@gmail.com, cmpun@umac.edu.mo, knightzhou@ieee.org

Abstract
Deep Neural Models are vulnerable to adversarial perturbations in classification. Many attack methods generate adversarial examples with large pixel modification and low cosine similarity with original images. In this paper, we propose an adversarial method generating perturbations based on root mean square gradient which formulates adversarial perturbation size in root mean square level and update gradient in direction, due to updating gradients with adaptive and root mean square stride, our method map origin, and corresponding adversarial image directly which shows good transferability in adversarial examples generation. We evaluate several traditional perturbations creating ways in image classification with our methods. Experimental results show that our approach works well and outperform recent techniques in the change of misclassifying image classification with slight pixel modification, and excellent efficiency in fooling deep network models.

Introduction
Deep Neural Networks (DNNs) (Szegedy et al. 2017; Simonyan and Zisserman 2014; Krizhevsky, Sutskever, and Hinton 2012; Goodfellow et al. 2014; Szegedy et al. 2016; He et al. 2016) have led to a dramatic improvement in recent years. It has achieved state-of-the-art performance on classification tasks (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014). With researches going specific domain, it find that deep neural model is vulnerable to slight perturbation attacks (Goodfellow, Shlens, and Szegedy 2014; Nguyen, Yosinski, and Clune 2015; Papernot et al. 2016; Poursaeed et al. 2018). Many methods generate attack perturbations which try to seek a decision boundary between different classes (Tabacof and Valle 2016; Kurakin, Goodfellow, and Bengio 2016; Carlini and Wagner 2017), fooling deep neural models with adversarial examples is to find suitable perturbation that can be leading DNNs to misclassify it boundary (Nguyen, Yosinski, and Clune 2015) definition learned. There are many ways to craft adversarial perturbations, such as single-step gradient updating attack (Goodfellow, Shlens, and Szegedy 2014) and iterative-steps gradient updating attack methods (Kurakin, Goodfellow, and Bengio 2016) and the white-box or black-box attack (Moosavi-Dezfooli et al. 2017; Papernot et al. 2016; Nguyen, Yosinski, and Clune 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Liu et al. 2016; Goodfellow, Shlens, and Szegedy 2014; Carlini and Wagner 2017). Attacks also can be classified into targeted or non-targeted attack methods, targeted attack aims to find perturbation which can treat models with define label, and non-targeted attack (Carlini and Wagner 2017; Poursaeed et al. 2018; Tabacof and Valle 2016) tries to find perturbation that fool model with the wrong tag whatever it is.

In this paper we proposed root mean square gradient-based adversarial examples generated algorithm which can craft adversarial cases that fool deep neural networks with...
slight perturbation and misclassify with almost high confidence. We summarize contributions in this paper follows: We propose an algorithm for generating adversarial examples with root mean square of gradient refinement when computing adversarial perturbations which generate perturbation tensors in slight and small size level that fool deep neural networks deployed with high probability and have higher similarity with corresponding clean images in cosine similarity, compared with other attack methods, our approach can get cosine-similarity closed to 0.989, which means our adversarial attack examples are very similar to origin images, and it is not accessible to defense this type of attacks.

**Related Work**

In adversarial perturbation generating process, adversarial attack methods try to find a perturbation matrix \( v \in R^d \) by adding the matrix to origin image, this generated image can fool the deep neural models to make mistake classification of \( x \) from correct label \( f(x) \) to other labels \( f(x + v) \), we define these images as adversarial examples. In (Szegedy et al. 2014), it first gave an introduction about generating adversarial examples of attacking state-of-the-art deep neural models. In (Goodfellow et al. 2014), it shown attack method based on gradient-Fast Gradient Sign Method(FGSM) and updates gradient in the direction of pixels only once time, and seek perturbation tensors \( v \). The iterative-step FGSM is easily derived(Kurakin, Goodfellow, and Bengio 2016). It proposed a targeted attack method called C&W’s attack(Carlini and Wagner 2017), which generate adversarial examples reduce detecting defenses rates. In (Moosavidezfali et al. 2017), it seek adversarial examples which fool one deep neural model with only one perturbation.

Improving the robustness of deep neural models is a comprehensive task. (Tramr et al. 2018) reveal ensemble model mechanism to adversarial defense attack which trains data with adversarial examples produced by pre-trained models and hold-out models that having good defense result against black-box and single-step attack. (Papernot et al. 2016) present distillation defense method which use distillation method reduce the effectiveness of adversarial samples on DNNs and lower adversarial attack success rates. (Xu, Evans, and Qi 2017) illustrate Feature Squeezing method which reduces the search space available to an adversarial example by coalescing samples that correspond to different feature vectors in the original space into a single sample.

**Methodology**

In this section, we introduce algorithm which generate perturbation based on gradient at root mean square(RMS)(Goodfery, Nitish, and Kevin), which helps to generate adversarial perturbation with less gradient vibration and good class transferability. In particular, we satisfy our generated perturbation tensors in \( L_\infty \) and \( L_2 \) norm constraint in our proposed method in this section. We focus on how to reduce pixel modification on adversarial examples and improve attack success rates on examples, root mean square of gradient gives resolutions. RMS uses the same concept of the exponentially weighted average of the gradients like gradient descent of momentum, it uses history pieces of information of gradients to decide updating direction and magnitude, which helps escape local minima, the difference is the update of parameters, when updating DNNs weights and bias parameters for each epoch, RMS update its weights and bias at the averages of the square level. We show our proposed method with \( L_\infty \) norm constraints in non-targeted attack strategy in Algorithm 1, targeted attack strategy is easy derived. In this part, we illustrate our proposed method which generates
adversarial examples at the square level and shows good effectiveness in stochastic gradient descent at stable updating direction, experiments show proposed method gains high attack success rate (ASR) with little pixel distortion.

**Experimental Results**

We conduct our experiments on ImageNet datasets [Deng et al. 2009] to validate the effectiveness of our proposed method in this section with attack setting in the following part, experiments settings are kept the same setup both in $L_2$ and $L_\infty$ norm constraints. We show attack success rates, cosine similarity and perturbation strength with our proposed method on preprocessed ILSVRC2012(Val) datasets.

As can be seen from the Fig.3, the attack effect of FGSM in both deep neural networks is less than 90%. Our proposed method achieved high ASR which is close to 100%. RMS method have high cosine similarity because the RMS-based gradient approach reduced when approaching the decision boundary during the perturbation calculation. This visually intuitive performance is that under the same constraints, the adversarial examples produced by our method are less perturbative than the adversarial images generated by I-FGSM, and the adversarial images are more clear and more difficult to recognize as processed images. Our proposed method can achieve high attack success under the white box attack strategy, and it is better than the I-FGSM method under the black box attack condition, which means that the proposed method can express better network transferability, especially network models are similar in structure, the transferability is better. E.g. VGG16 and VGG19.

Under the constraints of $L_2$, we found that ASR has improved compared with the ASR in $L_\infty$ norm, and the perturbation generated by the networks has a better transfer ef-

---

**Table 1: ASR in $L_\infty$ norm constraint on six deep neural models, * indicate white-box attacks. IR-v2 indicate InceptionResnet-v2**

| Attacks | Inc-v3 | Inc-v4 | IR-v2 |
|---------|--------|--------|-------|
| Inc-v3  | I-FGSM | 98.41% | 28.86% | 27.25% |
|         | Ours   | 99.48% | 29.45% | 28.55% |
| Inc-v4  | I-FGSM | 29.36% | 96.72%* | 25.45% |
|         | Ours   | 36.64% | 98.99%* | 32.23% |
| IR-v2   | I-FGSM | 28.77% | 28.26% | 96.65%* |
|         | Ours   | 36.64% | 32.23% | 99.24%* |
| Attacks | VGG16  |        |        |       |
|         | VGG19  |        |        |       |
|         | Res152 |        |        |       |
| VGG16   | I-FGSM | 94.83% | 58.29% | 30.21% |
|         | Ours   | 98.51% | 58.95% | 33.93% |
| VGG19   | I-FGSM | 58.70% | 92.62%* | 32.53% |
|         | Ours   | 59.46% | 98.43%* | 32.49% |
| Res152  | I-FGSM | 33.63% | 34.92% | 100.00%* |
|         | Ours   | 34.03% | 35.04% | 100.00%* |

**Table 2: ASR in $L_2$ norm constraint on six deep neural models, * indicate white-box attacks. IR-v2 indicate InceptionResnet-v2**

| Attacks | Inc-v3 | Inc-v4 | IR-v2 |
|---------|--------|--------|-------|
| Inc-v3  | I-FGSM | 99.92%* | 55.17% | 57.31% |
|         | Ours   | 99.76%* | 57.31% | 60.41% |
| Inc-v4  | I-FGSM | 59.41% | 98.34%* | 60.41% |
|         | Ours   | 68.15% | 99.72%* | 63.10% |
| IR-v2   | I-FGSM | 63.86% | 60.17% | 95.15%* |
|         | Ours   | 68.21% | 61.01% | 99.77%* |
| Attacks | VGG16  |        |        |       |
|         | VGG19  |        |        |       |
|         | Res152 |        |        |       |
| VGG16   | I-FGSM | 99.31%* | 69.01% | 62.45% |
|         | Ours   | 99.97%* | 75.22% | 63.13% |
| VGG19   | I-FGSM | 71.36% | 98.98%* | 57.88% |
|         | Ours   | 75.45% | 99.64%* | 58.41% |
| Res152  | I-FGSM | 65.41% | 59.65% | 99.79%* |
|         | Ours   | 66.42% | 60.10% | 100.00%* |

**Table 3: AMP values with different norm constraints on Inception-v3**

| Norm | Ours |
|------|------|
| $L_\infty$=10 | 0.012 |
| $L_2$=1500 | 0.027 |
fect, the attack effect on other networks has also improved which shows on the ASR with perturbation generated by a model attack other DNNs. We use the average of the Absolute Mean Perturbation values: $(\text{AMP}) = \frac{1}{|\mathcal{D}|} * \sum_{x \in \mathcal{D}} \|v_x\|$ as a measure of the magnitude of the disturbance, which is a representation of the magnitude of the value of the disturbance added to the pixels of the clean image.

**Conclusion**

In this paper, we describe an adversarial examples generation attack method which is based on the root mean square gradient to generate perturbation gap the boundary distance between different classes representation in latent space, we take $L_2$, $L_\infty$ two norm constraints into consideration with the no-targeted and targeted attack strategy. Our proposed methods treat the deep neural networks with high probability and generated perturbations show good transferability on different deep neural models which has a good effect in the black-box attack setup. The generated adversarial examples have high cosine similarity value with the corresponding clean images, which means that perturbation our proposed methods generated is in a small range, this can be directly reflected on the AMP. Next, we will focus our attention on image-independent adversarial attack.

**Acknowledgment**

This work was supported in part by the Research Committee of the University of Macau under Grant MYRG2018-00035-FST, and the Science and Technology Development Fund of Macau SAR under Grant 041-2017-A1.

**References**

[Bastani et al. 2016] Bastani, O.; Ioannou, Y.; Lampropoulos, L.; Vytiniotis, D.; Nori, A.; and Criminisi, A. 2016. Measuring neural net robustness with constraints. In *Advances in neural information processing systems*, 2613–2621.

[Carlini and Wagner 2017] Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. Ieee.

[Geoffrey, Nishit, and Kevin 2017] Geoffrey, H.; Nishit, S.; and Kevin, S. Neural networks for machine learning. [http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf](http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf)

[Goodfellow et al. 2014] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *Advances in Neural Information Processing Systems* 3:2672–2680.

[Goodfellow, Shlens, and Szegedy 2014] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. *CoRR*.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645. Springer.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, 1097–1105.

[Kurakin, Goodfellow, and Bengio 2016] Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

[Liu et al. 2016] Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *CoRR* abs/1611.02770.

[Moozavi-Dezfooli et al. 2017] Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 86–94. IEEE.

[Moozavi-Dezfooli, Fawzi, and Fossard 2016] Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Fossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2574–2582.

[Nguyen, Yosinski, and Clune 2015] Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436.

[Papernot et al. 2016] Papernot, N.; McDaniel, P. D.; Sinha, A.; and Wellman, M. P. 2016. Towards the science of security and privacy in machine learning. *CoRR* abs/1611.03814.

[Poursaeed et al. 2018] Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

[Szegedy et al. 2014] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

[Szegedy et al. 2016] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

[Szegedy et al. 2017] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, 12.
[Tabacof and Valle 2016] Tabacof, P., and Valle, E. 2016. Exploring the space of adversarial images. In 2016 International Joint Conference on Neural Networks (IJCNN), 426–433. IEEE.

[Tramr et al. 2018] Tramr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations.

[Xu, Evans, and Qi 2017] Xu, W.; Evans, D.; and Qi, Y. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.