Adversarial Example Generation

Yatie Xiao, Chi-Man Pun Senior Member, IEEE

Abstract

Deep Neural Networks have achieved remarkable success in computer vision, and audio tasks, etc. However, in classification domains, deep neural models are easily fooled by adversarial examples. Many attack methods generate adversarial examples with large image distortion and low similarity between origin and corresponding adversarial examples, to address these issues, we propose an adversarial method with an adaptive gradient in a direction to generate perturbations, it generates perturbations which can escape local minimal. In this paper, we evaluate several traditional perturbations creating methods in image classification with ours. Experimental results show that our approach works well and outperform recent techniques in the change of misclassifying image classification, and excellent efficiency in fooling deep network models.

1 Introduction

Deep Neural Networks (DNNs) [1–5] have led to a dramatic improvement on image, text and audio tasks recently. With research going specific domain, many research works reveal that in image classification domain, deep neural classification models deployed are easily fooled by adversarial examples [14–20]. By adding appropriate perturbations into clean images. Perturbations should be in enough strength level so that perturbation can make origin images over decision boundary [17,19,20] between different classes, and make generated adversarial examples representation in latent space approaching the other classes space and having more corresponding features to fool deep neural models.

In gradient updating aspect, we can classify adversarial attack methods on image classification, single-step attack [15], and iterative-steps attack [22]. From models’ structure perspective of view, if get understanding of deployed networks’ structure and parameters, attack methods is categorized into white-box attack [13,19], on the opposite side, it is categorized into black-box attack [23,24]. From attack oriented perspective, attacks can be classified into targeted or non-targeted attack methods [8,25], targeted attack aims to find perturbation which can fool models with define label as wish, and non-targeted attack tries to find perturbation that fool model with the other labels without any definition.

Many proposed attack methods can fool DNNs with prediction confidence and attack deep neural networks in image classifications [3,26]. We find that most of these methods craft adversarial examples with large perturbation and not very high attack success rates. In this paper we proposed adaptive perturbation 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 algorithms for generating adversarial perturbation with adaptive gradient which computes adversarial perturbations in slight and small size level that fool deep neural networks deployed with high fooling probability, and give a formulation which aims to calculate adversarial perturbation strength. By using this algorithm, we can qualify the perturbation strength between clean image and corresponding adversarial examples and it to reflect pixel modification strength between clean images and corresponding examples.

2 Related Work

In this section, we first give the background knowledge of adversarial examples and corresponding formalization of proposed algorithms. In the first place, let define $\mu$ as a distribution of images in $R^d$ which means images from a distribution with $x$ dimensions, and then $x$ is an image from $\mu$, at the same time, let $f$ denotes a classification function which is to make the input data(in this paper, we define images as input data) $x \in R^d$ output with an estimated label $f(x)$. In adversarial perturbation seeking process, it finds a perturbation tensor $v \in R^d$ by adding the tensor we gain to origin image, this generated image can fool the deep neural models to make mistake classifi-
2.1 Methods for Generating Adversarial Examples and Adversarial Defense

One-step generation algorithm (Fast Gradient Sign Method-FGSM [15, 16]) try to find proper perturbation vector, by maximizing the loss function \( J(x^*, f(x)) \), the algorithm finds a boundary between different classes so as to generate perturbation tensor, the formulation is shown following:

\[
x^* = x + \alpha * \text{sign} (\nabla x J(\theta, x, f(x))) \quad x \in \mu
\]  

(2)

The iterative-steps gradient-based algorithm (I-FGSM) [16,22] is learning models parameters well than FGSM due to updating gradient direction step by step well so that I-FGSM can generate perturbation with good transferability and high attack success rates for models under white-box setting.

\[
x^*_i = x_{i-1} + \alpha * \text{sign}(\nabla x J(\theta, x_{i-1}, f(x_{i-1})))
\]  

(3)

Carlini and Wagner proposed a targeted attack method called C&W’s attack [20], which generate adversarial examples reduce detecting defenses rates. It is shown:

\[
\text{min } \|v\|_p + c \ast g(x + \eta) \quad x+v \in [0,1]^d
\]  

(4)

the parameter \( p \) is norm constraint which is set to 0, 1, 2 or \( \infty \). Moosavidezfooli et al [27] seek adversarial examples which fool one deep neural model with only one perturbation.

\[
\|v\|_p \leq \epsilon \quad P(f(x') \neq f(x)) \geq 1 - \delta
\]  

(5)

In the equation, \( \delta \) is set to control adversarial examples attack success rate. Besides, Meeting boundary distance between origin images and relevant adversarial examples under norm limitation setting [28], which is expressed in the following:

\[
\text{arg min } \epsilon * \|x^* - x\|_p - J(\theta, x^*, y)
\]  

(6)

In this equation, the method tries to optimize boundary distance between the origin and adversarial examples. In [29], it introduce a simple iterative method (JSMA) for targeted attack.

\[
\text{argmin } \|v\|_p \quad \text{s.t. } f(x+v) = y^* \neq y
\]  

(7)
matrix. We formulate $G$ as follows:

$$G = \sum_{i=1}^{\tau} g_i * g_i^T$$  \hspace{1cm} (8)$$

the $g_i$ is the gradient at time $i$, we give it to $\nabla_x J(\theta, x_i, f(x_i))$, so in adaptive gradient mechanism, this model’s parameter is updated after every iteration, see in follows:

$$W_i = W_{i-1} + \frac{\eta}{\sqrt{G_{i,i}}} * g_i$$  \hspace{1cm} (9)$$

It reveals that parameters update relies on high-parameter $\eta$ and the denominator $\sqrt{G_{i,i}}$, once the $G_{i,i}$ grows high, the parameter $W$ will update with a low rate, so as escape the local minima or maxima when searching directions. We calculate $g_i$ as original adaptive pace by:

$$g_i = \text{sign}(\nabla_x J(\theta, x_i, f(x_i)))^2$$  \hspace{1cm} (10)$$

We calculate with sign function for the square of loss, according to adaptive gradient processing, we draw a figure in the following. $g_i$ is the key adaptive idea for our methods, using iterative fast gradient sign method strategy, $g_i$ adjust the pace of generating $x_i^*$ by the next equation,

$$x_i^* = x_i + \epsilon * \frac{\text{sign}(\nabla_x J(\theta, x_i, f(x_i)))}{\sqrt{g_i^2 + \delta}}$$  \hspace{1cm} (11)$$

This method adapts the learning rate to the parameters, performing smaller updates for parameters associated with frequently occurring features, and substantial updates for parameters associated with infrequent features. Process of generating adversarial examples.

As we can see from algorithm 1, we focus on the output vector $v_i$ which is generated by Eq.11, then updating the perturbation tensor, after iterations in the direction of generated perturbation, the classifier of models seek right perturbation tensor. We use cosine similarity to evaluate the difference between original images and adversarial images.

4 Experimental Results

We conduct our experiments on ImageNet datasets [34] 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.

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 |
|---------|--------|--------|-------|
| I-FGSM  | 96.41% | 28.86% | 27.23%|
| MI-FGSM | 95.62%*| 25.22% | 25.27%|
| Ours    | 99.65% | 28.99% | 28.81%|
|---------|--------|--------|-------|
| I-FGSM  | 29.36% | 96.72% | 25.43%|
| MI-FGSM | 28.17% | 95.27% | 26.41%|
| Ours    | 31.55% | 99.11% | 31.44%|
|---------|--------|--------|-------|
| I-FGSM  | 35.16% | 32.54% | 98.14%*|
| MI-FGSM | 26.67% | 25.65% | 97.01%*|
| Ours    | 34.25% | 35.76% | 99.41%*|

Our proposed method generate adversarial examples which has high cosine similarity because the adaptive gradient based 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 or MI-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 or MI-FGSM methods 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. Inception-v3 and Inception-v4.

We use the average of the Absolute Mean Perturbation values:(AMP) = $\frac{1}{|V_i|} \sum_{v_i \in V_i} ||v_i||$ as a measure of the magnitude of the disturbance, which is a representation of the scope of the value of the disturbance added to the pixels of the clean image.

5 Conclusion

We describe an adversarial examples generation attack method which generate perturbation with adaptive pace gaping the boundary distance between different classes representation in latent space; we take $L_2, L_\infty$ two norm constraints into consideration.
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  | 98.72% | 55.17% | 57.31% |
| MI-FGSM | 94.94% | 56.55% | 58.11% |
| Ours    | 99.86%*| 59.23% | 63.01% |
| Inc-v4  | 59.41% | 98.34%*| 60.41% |
| MI-FGSM | 61.65% | 95.88% | 61.01% |
| Ours    | 67.44%| 99.62%*| 62.89% |
| IR-v2   | 63.86% | 60.17% | 95.15% |
| MI-FGSM | 66.20% | 59.88% | 97.41% |
| Ours    | 67.61%| 64.42% | 99.15% |

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

| Attacks | $L_\infty=10$ | $L_2=1500$ |
|---------|--------------|------------|
| Inc-v3  | 0.041        | 0.015      |
| MI-FGSM | 0.086        | 0.019      |
| Ours    | 0.016        | 0.008      |

with the no-targeted and targeted attack strategy. Our proposed methods generate adversarial examples which fool 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, and having high similarity with original images which is directly reflected on the AMP. Next, we will focus our attention on the image-independent adversarial attack.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks (2013). arXiv preprint arXiv:1312.6199, 594:595, 2014.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In International Conference on Neural Information Processing Systems, pages 1097–1105, 2012.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Advances in Neural Information Processing Systems, 3:2672–2680, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016.

[6] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. arXiv preprint arXiv:1708.05493, 2017.

[7] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436, 2015.

[8] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280, 2017.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[10] Edward Marcus Batchelder and R Pito Salas. Text abstraction method and apparatus, November 25 1997. US Patent 5,691,708.

[11] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE Symposium on Security and Privacy (SP), pages 582–597. IEEE, 2016.

[12] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. arXiv preprint arXiv:1511.05122, 2015.
[13] Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. In 2016 International Joint Conference on Neural Networks (IJCNN), pages 426–433. IEEE, 2016.

[14] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770, 2016.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. CoRR, abs/1412.6572, 2014.

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Discovering adversarial examples with momentum. arXiv preprint arXiv:1710.06081, 2017.

[17] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. 2017 IEEE International Conference on Computer Vision (ICCV), pages 1378–1387, 2017.

[18] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. arXiv preprint arXiv:1801.01944, 2018.

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

[20] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814, 2016.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Re-thinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[23] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Joint European conference on machine learning and knowledge discovery in databases, pages 387–402. Springer, 2013.

[24] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015.

[25] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57. IEEE, 2017.

[26] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. IEEE Transactions on speech and audio processing, 10(7):504–516, 2002.

[27] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 86–94. IEEE, 2017.

[28] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In Advances in neural information processing systems, pages 2613–2621, 2016.

[29] Nicholas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. pages 372–387, 03 2016.

[30] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017.

[31] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In Advances in Neural Information Processing Systems, pages 1632–1640, 2016.

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Analysis of universal adversarial perturbations. arXiv preprint arXiv:1705.09554, 2017.
[33] Hinton Geoffrey, Srivastava Nitish, and Swersky Kevin. Neural networks for machine learning. 
http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. Ieee, 2009.