ADVERSARIAL ATTACKS ON SEMI-SUPERVISED DEEP NEURAL NETWORKS

Yuchao Wang[⋆] *Enmei Tu*[⋆†] *Yehui Yang*[⋆] *Chunhui Wang*[⋆] *Jie Yang* [†]

[⋆] School of Electronics, Information and Electrical Engineering, Shanghai Jiao Tong University, China
[†] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China

ABSTRACT

Deep neural networks have been shown to be vulnerable to adversarial attacks, an intended but almost invisible manipulation to testing samples. The question of whether this vulnerability is the same to different types of deep models has not been well addressed. Complementary to existing researches on supervised models, this work studies the robustness of another important type of deep model, semi-supervised learning models, against adversarial attacks. We also propose a new iterative adversarial learning algorithm to further boost adversarial attack success rate and thus to help investigate the mechanism of adversarial attack. Experimental results demonstrate the superior performance of our approach over the state-of-the-art approaches.

Index Terms— Semi-Supervised Learning, Adversarial Attack, Deep Neural Networks.

1. INTRODUCTION

Deep neural networks (DNNs) have achieved great success in many research areas, such as computer vision, natural language and recommended system. However, recent works show that deep neural networks are extremely vulnerable to adversarial attacks [1, 2]. Adversarial attacks can be implemented by adding small but human-imperceptible noises to the test data examples and cause the accuracy of the test data to drop to an unacceptable level. This vulnerability of DNNs becomes a major threat to real-world deployments. On the other side, a deep model retrained by its adversarial examples has been proved to be more robust to adversarial attacks [3, 4]. Therefore, effective adversarial example generation algorithms are useful to study the vulnerability and to remedy the defects of DNNs models.

Existing studies on the vulnurability of DNNs mainly concentrate on supervised learning models [5, 6, 7]. The robustness of other types of DNNs, say semi-supervised learning (SSL) deep models [8, 9, 10], against adversarial learning has not been exploited yet. In this paper, we devote to studying the properties of semi-supervised deep models against different levels of adversarial attacks and developing a more effective attacking algorithm. We found that the responses of semi-supervised models are different from that of supervised models. More specifically, under strong level attacks, semi-supervised models tend to be more robust when using less labeled samples to train. This is significantly different from supervised learning, for which more training data lead to a more robust model [11]. In addition, we propose a new adversarial attack algorithm to enhance attacking success rate, which would be beneficial for developing more advanced defense strategies.

In summary, our contributions are as follows:

- We evaluate recent semi-supervised deep models against adversarial attacks and found different properties from supervised learning reported in existing works.
- We propose a new adversarial attack algorithm which achieves higher attacking success rate for both supervised and semi-supervised deep model.
- We construct a unifying software package to evaluate different attacking algorithms on different semi-supervised learning algorithms, and conduct extensive experiments using the package to validate our methodology.

2. METHODOLOGY

In this section, we first introduce the approach to evaluate the robustness of semi-supervised deep models under adversarial attack. Then, we present our new iterative attacking algorithm.

2.1. Attack semi-supervised model using FGSM

Suppose there is a target classifier $f : \mathcal{X} \subset \mathcal{R}^d \longrightarrow \mathcal{Y}$, which correctly classifies a sample $x \in \mathcal{X}$ to its true class $y \in \mathcal{Y}$ after training, i.e. $f(x) = y$. $\mathcal{Y}$ is the ground truth label set of $\mathcal{X}$ and its elements are from $\{1, 2, ..., C\}$. The aim of an adversarial attack is to manipulate the sample x slightly (i.e., by adding a small but customized perturbation to x) to produce x^* so that $f(x^*) \neq y$. Here $f(x^*)$ could be other class numbers. In this case, the classifier is misled by the adversarial sample x^*. Usually a norm constraint $\|x - x^*\|_n$ is applied in order to make sure that x^* is not too far from x, where the n could be 1, 2, ∞. Mathematically, suppose the

loss function of f is $J(f(x), y)$, adversarial attack generating x^* by solving the following optimization problem[1]

$$\begin{aligned} &\arg\max_{x^*} \quad J(f(x^*), y). \\ &\text{s.t. } \|x^* - x\|_\infty \leq \epsilon \end{aligned} \tag{1}$$

In order to generate an adversarial example from a real example x, fast gradient sign (FGSM) [2] use a one-step generation algorithm in equation Eq. 2.

$$x^* = x + \epsilon \cdot sign(\bigtriangledown_x J(f(x), y)). \tag{2}$$

ϵ is the step size. FGSM assumes that the decision boundary around the data point is linear and a small number of accumulated movements away from the class will bring the sample across the border.

In a semi-supervised scenario, we first use a small number of labeled samples, denoted as $\mathcal{X}_L$, and a big number of unlabeled samples, denoted as $\mathcal{X}_U$, to train the classifier f. When the classification accuracy of f on a separate testing set $\mathcal{X}_T$ reach to a satisfactory level (i.e. each classifier attains at least 96% reported accuracy in the original paper), we fix all the parameters of f and apply the FGSM attack approach to it. More specifically, we perform a one-step perturbation using equation Eq. 2 to all the testing samples in $\mathcal{X}_T$ to obtain a manipulated new testing set $\mathcal{X}_T^*$, and then evaluate the classification accuracy of f on $\mathcal{X}_T^*$.

Four representative and widely used semi-supervised learning algorithms are selected as the attack targets.

1. Mean Teacher [8]: The overall architecture of mean teacher (MT) consists of two parts: student model and teacher model. The network parameters of the teacher model are obtained by the moving average of the network parameters of the student model.

2. Smooth Neighborhood [13]: To take pair-wise relationship of data points into account, smooth neighbors on teacher graphs (SNTG) forces the predictions of neighboring points to be the same. For non-neighboring points, it pulls them farther than a predefined distance m.

3. MixMatch [9]: The key of MixMatch is to effectively fuse the ideas of pseudo label [14], consistency [15] and mixup [16] together to make a "holistic" approach.

4. FixMatch [10]: FixMatch further boosts MixMatch by aligning the output of a strong perturbed version to a weak perturbed version of an unlabeled sample x.

[1]One could also generate an adversarial which is quite different from x and visually belong to another class,but the classifier output remains unchanged, see [12].

2.2. A new adversarial attack method

In practice, a class boundary is often highly nonlinear in high-dimensional space, so FGSM may have limited attack ability. Instead, the iterative FGSM (I-FGSM) [17] continuously moves the adversarial example along the direction of the gradient, as shown in the equation Eq. 3. Although it can achieve a highly successful attack rate after several iterations, the adversarial example can easily fall into a bad local maximum and "overfit" the model, which is less likely to migrate across models, hence low transferability.

In this subsection, we propose a regularized update descent [18] based algorithm to generate adversarial examples to further improve the performance of FGSM and I-FGSM.

Suppose we adopt the gradient ascendant algorithm to solve the optimization problem in equation Eq. 1. It results in the following update

$$x_{t+1}^* = x_t^* + \alpha v_t \tag{3}$$

where $v_t = \bigtriangledown_x J(f(x_t^*), y)$ is the gradient of loss function J at x_t^*. Noting that the gradient vanishes for optimal solution (i.e. $v_t = 0$), the main idea of our adversarial attack algorithm is to optimize simultaneously the target variable x^* and the gradient update v_t. To be more specific, at t^{th} iteration, the original optimization problem in equation (1) is reformulated as

$$J(x_t^*, v_t) \equiv J(x_t^* + v_t) + \gamma_t \frac{v_t^2}{2} \tag{4}$$

where the last term $\frac{\gamma_t v_t^2}{2}$ is a regularization term to avoid over-confident updates (γ_t is the regularization coefficient).

When an update v_t at the current input x_t^* is small enough, using the second-order approximation we can get the following equation:

$$J(x_t^* + v_t) = J(x_t^*) + v_t J'\{x_t^*\} + \frac{1}{2} v_t^2 J''\{x_t^*\} + O\{v_t^3\} \tag{5}$$

Therefore, given the adversarial attack target classifier $f(x)$, its loss function J and an input sample image x, the update rule of our novel attach algorithm is given by

$$\begin{aligned} \tilde{v}_{t+1} &= \mu \cdot \tilde{v}_t + \frac{\tilde{v}_t}{||\tilde{v}_t||_1} \\ x_{t+1}^* &= x_t^* + \alpha \cdot sign(\tilde{v}_{t+1}) \end{aligned} \tag{6}$$

where $\tilde{v}_t = \bigtriangledown_x \tilde{J}(x_t^*, v_t)$. α_t is the learning rate and μ_t is the decay rate at iteration t. The first equation is the accumulated normalized momentum to stabilize gradient update. The overall adversarial learning algorithm is in algorithm 1. More theoretical analysis can be found in [19] and [20].

3. EXPERIMENTAL RESULTS

We first conduct experiments to evaluate the FGSM attack on the selected semi-supervised learning neural networks. Then

Algorithm 1 RUD-FGSM

Require:
A classifier f with loss function J,
A sample-label pair (x, y),
Perturbation parameter ϵ,
Iteration number T,
Decay factor μ.

Ensure:
An adversarial example x^* with $\|x^* - x\|_\infty \leq \epsilon$;

$\alpha = \frac{\epsilon}{T}, x_0^* = x$;
for each t=0 to $T-1$ **do**
  Input x_t^* to f and compute the gradient $\tilde{v}_t$;
  Update according to equation 6
end for
return $x^* = x_T^*$;

we make comparisons of our proposed new attack algorithm with other attack algorithms.

3.1. Attacking semi-supervised deep models using FGSM

We conduct experiments on the benchmark dataset CIFAR-10 [21], which consists of 60000 colour images equally distributed in 10 classes, and is divided into 50000 training images and 10000 test images. To train the semi-supervised networks, we only evaluate model accuracies for different number of labeled samples: 250, 500, 1000, 2000, 4000. We train the networks of each semi-supervised learning algorithm using the reported parameter settings and training strategies in their original papers. In addition, to evaluate the sensitivity of each semi-supervised model to different perturbation levels (hence the adversarial attack strength), we change the parameter ϵ from 0.01 to 1. The larger the parameter is, the stronger the sample is perturbed and the attack is. The experimental results are show in Table 1, which includes testing accuracies on perturbed testing data. The baseline accuracies are shown in Table 2, which includes testing accuracies on normal testing data. The perturbed testing set is a duplication of the normal testing set, except that each sample in it is poisoned by the attack algorithm FGSM.

From these results, we can see that when ϵ is small (0.01), the accuracy losses of all semi-supervised models are also small and roughly the same for a different number of labeled samples. As the value of ϵ goes large, the prediction accuracies decay very quickly. Particularly, for each, the more labeled samples are used for training, the quicker the accuracies decrease. Noting that this phenomenon of semi-supervised learning is very different from existing researches on supervised learning and, to the best of our knowledge, has not been reported. In [11], the authors found that with more labeled data, the robustness of a supervised learning model against adversarial attack increases.

Method	Labeled	ϵ=0.01	ϵ=0.1	ϵ=0.5	ϵ=1.0
MT	250	70.32	69.15	66.13	57.60
	500	75.32	68.26	59.07	50.31
	1000	83.01	79.96	53.13	29.24
	4000	89.73	80.21	45.52	25.71
SNTG	250	77.93	77.35	70.68	59.76
	500	81.16	80.47	69.50	54.79
	1000	84.45	83.43	56.43	30.62
	4000	90.07	87.19	44.38	27.16
MM	250	88.48	88.24	84.06	70.78
	500	89.30	88.11	77.84	61.07
	1000	89.91	88.55	61.73	31.44
	4000	92.73	91.38	62.41	29.38
FM	250	91.92	90.89	86.28	73.85
	500	92.37	91.45	79.41	55.88
	1000	93.21	92.14	67.69	33.78
	4000	94.73	93.21	60.19	30.99

Table 1. Testing accuracies (in percent) when parameter ϵ of FGSM attack algorithm changes. "Labeled" means the total number of labeled data for training.

Method / Labeled	**MT**	**SNTG**	**MM**	**FM**
250	70.33	77.94	88.49	91.93
500	75.34	81.17	89.32	92.40
1000	83.02	84.46	89.92	93.23
4000	89.76	90.09	92.74	94.75

Table 2. Testing accuracies (in percent) without adversarial attack.

3.2. Comparisons of adversarial attack algorithms

In this experiment, FGSM[2], MI-FGSM[17], PGD[3], CW[22] and our proposed adversarial attack algorithm RUD-FGSM1 are compared to attack the selected semi-supervised learning neural networks.

For FGSM, only one parameter ϵ, which means the perturbation's absolute value, needs to be considered. For MI-FGSM, the decay factor μ is set to be 1 as [17] suggests. For RUD-FGSM, we set the parameter learning rate to α to 0.8 and the decay factor μ to 1. The iterations parameter T is set to 20 after many attempts. We include two forms of CW, i.e. CW(L_2) and CW(L_∞). They differs in the norm of perturbation. The learning rate for both forms is 0.01. For PGD, the attack step size is set to be 0.1 and the maximum number of iterations is set to be 100, as suggested in the paper. We set the same parameter $\epsilon = 0.3$ for these algorithms. The results are shown in Table 3.

The results show that by using iterative multi-step perturbations for adversarial attack, MI-FGSM, RUF-FGSM, PGD, CW(L_2) and CW(L_∞) attain considerably higher accuracy decrease than the FGSM one-step perturbation. Note that the

SSL Method \ Attack Method	"Clean"	FGSM	MI-FGSM	PGD	CW(L_2)	CW(L_∞)	RUD-FGSM
MixMatch	89.32	82.07	80.23	81.75	80.73	80.04	**79.21**
ReMixMatch	90.94	85.38	84.09	84.11	84.66	82.71	**80.26**
FixMatch	92.40	86.75	85.74	84.93	83.57	**82.92**	85.13
MT	75.34	69.66	67.32	66.84	65.42	64.99	**64.71**
MT+SNTG	81.17	72.40	69.18	69.35	67.12	69.86	**66.55**

Table 3. Testing accuracies (in percent) for different attack algorithms with only 500 labeled data for training. "Clean" means the version of testing set with no adversarial attack.

proposed RUF-FGSM achieves more decrease in most cases. A reasonable explanation is that the algorithm uses higher-order information than the momentum algorithm to look further ahead and thus is able to converge faster and more steady than other algorithms . Besides, to prove our algorithm also works in fully supervised scenario, we conduct experiment on MNIST dataset[23] in fullly supervised paradigm. The result is shown in Table 4. The parameter is set to be the same in Table 3. It can be seen that our proposed algorithm outperforms other algorithms, proving its effectiveness under full supervised scenario.

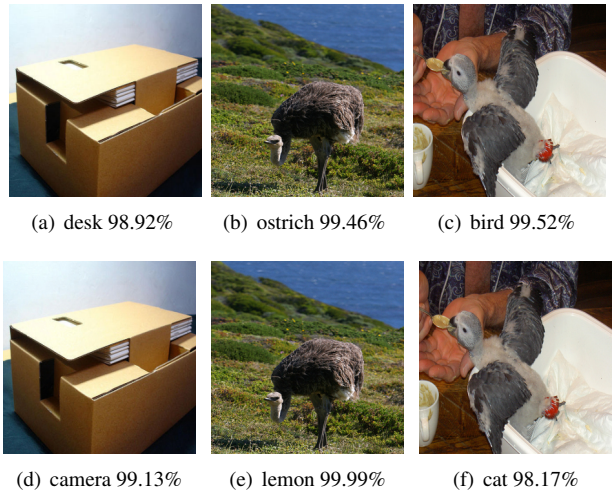
Attack Method	Accuracy
FGSM	57.28
MI-FGSM	44.60
PGD	49.55
CW(L_2)	46.98
CW(L_∞)	45.27
RUD-FGSM	42.13

Table 4. Testing accuracies (in percent) for different adversarial attack on mnist dataset. The data is calculated by averaging of 5 runs.

To further validate the capability of our new attack algorithm, figure 1 shows the network classification results of the original testing samples (a-c) and the attacked samples (d-f). From these examples, we can see that the proposed attack algorithm is able to generate highly cheating samples with very small data perturbation (almost imperceptible to human eyes). Thus, it is very helpful to study adversarial mechanisms and to develop effective defense measures.

4. CONCLUSION

In this paper, we study the behavior of semi-supervised learning deep models to adversarial attacks. Complementary to existing researches on supervised learning, we found that the number of labeled data is reverse proportional to the robustness of a semi-supervised learning model, especially for a strong attack. We hypothesize that less labeled samples may result in a less abrupt classification boundary so that the adversarial movements have fewer effects on the network decision. We leave this as our future work. We also propose a new adversarial attack algorithm and demonstrate its validity by comparing it with widely used FGSM and MI-FGSM.



(a) desk 98.92% (b) ostrich 99.46% (c) bird 99.52%

(d) camera 99.13% (e) lemon 99.99% (f) cat 98.17%

Fig. 1. The original images classification outputs (a-c) and their adversarial counterparts outputs (d-f) generated by the proposed RUD-FGSM attack algorithm. The percentages are network classification confidence.

Furthermore, based on the current evaluation of semi-supervised learning models and our proposed effective attack algorithm, it is imperative to develop more robust semi-supervised learning approaches and defense techniques [24]. We will work on this in future research.

5. REFERENCES

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

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig

Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[4] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[5] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[6] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[7] Battista Biggio and Fabio Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[8] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[9] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5049–5059.

[10] Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," 2020.

[11] Ke Sun, Zhanxing Zhu, and Zhouchen Lin, "Towards understanding adversarial examples systematically: Exploring data size, task and model factors," *arXiv preprint arXiv:1902.11019*, 2019.

[12] Sanli Tang, Xiaolin Huang, Mingjian Chen, Chengjin Sun, and Jie Yang, "Adversarial attack type i: Cheat classifiers by significant changes," *arXiv preprint arXiv:1809.00594*, 2018.

[13] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8896–8905.

[14] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3.

[15] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Advances in neural information processing systems*, 2016, pp. 1163–1171.

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[18] Aleksandar Botev, Guy Lever, and David Barber, "Nesterov's accelerated gradient and momentum as approximations to regularised update descent," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1899–1903.

[19] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein, "Are adversarial examples inevitable?," *arXiv preprint arXiv:1809.02104*, 2018.

[20] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif, "Theoretical evidence for adversarial robustness through randomization," in *Advances in Neural Information Processing Systems*, 2019, pp. 11838–11848.

[21] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[22] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.