Bamboo: Ball-Shape Data Augmentation Against Adversarial Attacks from All Directions

Huanrui Yang
Duke University
huanrui.yang@duke.edu

Jingchi Zhang
Duke University
jingchi.zhang@duke.edu

Hsin-Pai Cheng
Duke University
hc218@duke.edu

Wenhan Wang
Microsoft
wenhanw@microsoft.com

Yiran Chen
Duke University
yiran.chen@duke.edu

Hai Li
Duke University
hai.li@duke.edu

Abstract

The robustness of Deep neural networks (DNNs) has been recently challenged by adversarial attacks State-of-the-art defending algorithms improve DNNs' robustness by paying high computational costs. Moreover, these approaches are usually designed against one or a few known attacking techniques only. The effectiveness to defend other types of attacking methods cannot be guaranteed. In this work, we propose *Bamboo* – the first data augmentation method designed for improving the general robustness of DNN without any hypothesis on the attacking algorithms. Our experiments show that *Bamboo* substantially improve the general robustness against arbitrary types of attacks and noises, achieving better results comparing to previous adversarial training methods, robust optimization methods and other data augmentation methods with the same amount of data points.

Introduction

In recent years, deep neural network (DNN) models (e.g., CNNs) have been widely used in many real-world applications (LeCun et al. 1998; Simonyan and Zisserman 2014). However, they exposed a high sensitivity to input data samples and therefore are vulnerable to *adversarial attacks*. A "small" perturbation can be applied on input samples, which is visually indistinguishable by humans but can result in the misclassification of DNN models (Szegedy et al. 2013; Carlini and Wagner 2017; Madry et al. 2018), indicating a serious threat against the systems using DNN models.

Many approaches have also been proposed to defend against adversarial attacks. However, *adversarial training* methods (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2018) won't guarantee the performance against previously unseen attacks (Carlini and Wagner 2017). While solving the *min-max* problem used in *Optimization based methods* (Sinha, Namkoong, and Duchi 2017; Yan, Guo, and Zhang 2018) often generates a high computational load.

Generally speaking, defending against adversarial attacks can be considered as a special case of increasing the generalizability of DNN to unseen data points. Therefore *data augmentation method* may also be effective. Previous studies show that training with additional data sampled from a Gaussian distribution centered at the original training data can enhance the model robustness against natural noise (Chapelle et al. 2001). The recently proposed *Mixup* method (Zhang et al. 2017) surprisingly improved the DNN robustness against adversarial attacks. However, these data augmentation may not offer the most efficient way to enhance the adversarial robustness of DNN as they are not designed against adversarial attacks.

In this work, we propose *Bamboo*—a ball shape data augmentation technique aiming for improving the general robustness of DNN against adversarial attacks from *all directions*. Without requiring any prior knowledge of the attacking algorithm, *Bamboo* can effectively enhance the general robustness of the DNN models against the adversarial noise. *Bamboo* can offer a significantly enhanced model robustness comparing to previous robust optimization methods, without suffering from the high computational complexity. Comparing to other data augmentation method, *Bamboo* can also achieve further improvement of the model robustness using the same amount of augmented data. Most importantly, as our method makes no assumption on the distribution of adversarial examples, it can work against all kinds of noise.

Background

Measurement of DNN robustness

A metric for measuring the robustness of the DNN is necessary. (Szegedy et al. 2013) propose the fast gradient sign method (FGSM) noise, which is one of the most efficient and most commonly applied attacking method. FGSM generates an adversarial example x' using the sign of the local gradient of the loss function J at a data point x with label y as: $x' = x + \epsilon\,\mathrm{sign}(\nabla_x J(\theta, x, y))$, where ϵ controls the strength of FGSM attack. For its high efficiency in noise generation, the classification accuracy under the FGSM attack with certain ϵ has been taken as a metric of the model robustness.

As FGSM attack leverages only the local gradient for perturbing the input, if is found that even a DNN model achieves high accuracy under FGSM attack, it may still be vulnerable to other attacking methods (Papernot et al. 2016). (Madry et al. 2018) propose projected gradient descent (PGD), which attacks the input with multi-step variant FGSM that is projected into certain space $x + \mathcal{S}$ at the vicinity of data point x for each step. A single step of the PGD noise generation can be formulated as: $x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \epsilon\,\mathrm{sign}(\nabla_x J(\theta, x, y)))$. Their work shows that comparing to FGSM, adversarial training using PGD adver-

sarial is more likely to lead to a universally robust model. Therefore the classification accuracy under the PGD attack would also be an effective metric of the model robustness.

Besides these gradient based methods, the generation of adversarial examples can also be viewed as an optimization process. (Szegedy et al. 2013) describe the general objective of *untargeted* attacks as:$minimize_\delta\ D(x, x+\delta), s.t.\ C(x+\delta) \neq C(x)$. Where D is the distance measurement, which we use L_2 distance here; C is the classification result of the DNN; and $x' = x + \delta$ is the adversarial example to be found. CW attack (Carlini and Wagner 2017) defines an objective function f such that $C(x+\delta) \neq C(x)$ if and only if $f(x+\delta) \leq 0$. With the use of f, the optimization can be formulated as: $minimize_\delta\ D(x, x+\delta) + c \cdot f(x+\delta)$. Such objective can lead to a higher chance of finding the optimal δ efficiently (Carlini and Wagner 2017). Since the objective of CW attack is to find the minimal possible perturbation strength of a successful attack, the average strength required for a successful CW attack can be considered as a reasonable measurement of the model robustness.

Previous works increasing network robustness

There are previous attempts to derive a bound of the DNN robustness theoretically (Peck et al. 2017; Hein and Andriushchenko 2017), but these obtained bounds are often too loose or too complicated to be used as a guideline for robust training. A more practical approach is *adversarial training*. For example, we can generate adversarial examples from the training data and then include their classification loss to the loss function (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2018). This method can be efficiently optimized for the limited types of *known* adversarial attacks. However, it may not promise the robustness against other attacking methods, especially those newly proposed ones. Alternatively, the defender may online generate the worst-case adversarial examples of the training data and minimize the loss of such adversarial examples by solving a *min-max* optimization problem during the training process. For instance, the distributional robustness method (Sinha, Namkoong, and Duchi 2017) use the objective $minimize_\theta\ F(\theta) := \mathbb{E}[sup_{x'}\{L(\theta; x') - \gamma D(x', x)\}]$ to train the weight θ of a DNN model that minimize the loss L of adversarial example x' which is near to original data point x but has supremum loss. This method can achieve some robustness improvement, but suffers from high computational cost for optimizing both the network weight and the potential adversarial example. Also, this work only focuses on small perturbation attacks, so the robustness guarantee may not hold on the improvement of robustness under large attacking strength (Sinha, Namkoong, and Duchi 2017).

Proposed Approach

Vicinity risk minimization for robustness

Most of the supervised machine learning algorithms follow the principle of *empirical risk minimization* (ERM), which is based on the hypothesis that the testing data has a similar distribution as the training data, so minimizing the loss on the training data would naturally lead to the minimum testing loss. However, the distribution of adversarial examples generated by attacking algorithms may be different from the original training data. Thus the DNN models trained with ERM would have unsatisfactory performance on adversarial examples (Goodfellow, Shlens, and Szegedy 2014).

Instead of ERM, the *vicinity risk minimization* (VRM) principle targets to minimize the *vicinity risk* $\hat{R}$ on the *virtual data pair* $(\hat{x}, \hat{y})$ sampled from a *vicinity distribution* $\hat{P}(\hat{x}, \hat{y}|x, y)$ generated from the original training set distribution $P(x, y)$ (Chapelle et al. 2001). Consequently, the optimization objective of the VRM-based training can be described as: $minimize_\theta\ \hat{R}(\theta) := \mathbb{E}_{(\hat{x},\hat{y})} L(f(\hat{x}, \theta), \hat{y})$.

For most of the attacking algorithms, there is a constraint on the strength of the perturbation, so the adversarial example $\hat{x}$ can be considered as within a r-radius ball around the original data x. Without any prior knowledge of the attacking algorithm, we can consider the adversarial examples as uniformly distributed within the r-radius ball: $\hat{x} \sim Uniform(||\hat{x} - x||_2 \leq r)$. However, directly sampling the virtual data point $\hat{x}$ within the ball may be data inefficient. Here we propose to further improve the data efficiency by utilizing the geometry analysis of DNN model. Previous research shows that the curvature of DNN's decision boundary near a training data point would most likely be very small (Goodfellow, Shlens, and Szegedy 2014; Fawzi, Moosavi-Dezfooli, and Frossard 2016). These observations show that minimizing the loss of data points sampled *within* the ball can be approximated by minimizing the loss of data points sampled *on* the edge of the ball. Formally, the vicinity distribution can be modified to:

$$\hat{P}(\hat{x}, \hat{y}|x, y) = Uniform(||\hat{x} - x||_2 = r) \cdot \delta(\hat{y}, y). \quad (1)$$

By optimizing the VRM objective with this vicinity distribution, we can improve the robustness of DNN against adversarial attacks with higher data efficiency in sampling the virtual data points for augmentation.

Bamboo and its intuitive explanation

We propose *Bamboo*, a ball-shape data augmentation scheme that augments the training set with N virtual data points uniformly sampled from a r-radius ball centered at each original training data point. Algorithm 1 provides a formal description of the proposed method.

Since the decision boundary of the DNN model tends to have small curvature around training data points (Fawzi, Moosavi-Dezfooli, and Frossard 2016), including the augmented data on the ball naturally pushes the decision boundary further away from the original training data, therefore increases the robustness of the learned model. Figure 1 shows the effect of *Bamboo* with a simple classification problem. Here we classify 100 data points sampled from the MNIST class of the digit "3" and digit "7" each using a multi-layer perceptron with one hidden layer. PCA is used for visualization. Figure 1a shows the decision boundary without data augmentation, where the decision boundary is more curvy and is overfitting to the training data. In Figure 1b, the decision boundary after applying our data augmentation becomes smoother and is further away from original training

Algorithm 1: *Bamboo*: Ball-shape data augmentation

Input : Augmentation ratio N, Ball radius r, Original training set (X, Y)

Output: Augmented training set $(\hat{X}, \hat{Y})$

$n :=$ length(X);

$\hat{X} := X, \hat{Y} := Y$; ▷ Initialization with training set data

$count := n$;

for $i = 1 : n$ **do**

  $x := X[i], y := Y[i]$;

  for $j = 1 : N$ **do**

    $count := count + 1$;

    Sample $\delta \sim \mathcal{N}(0, \mathcal{I})$; $\delta_r := \frac{\delta}{||\delta||_2} \cdot r$;

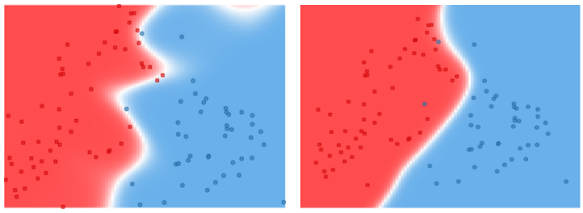
    $\hat{X}[count] := x + \delta_r$;

    $\hat{Y}[count] := y$; ▷ Augmenting the training set

  end

end

return $(\hat{X}, \hat{Y})$



(a) Without data augmentation (b) *Bamboo* data augmentation

Figure 1: *Bamboo*'s effect on the DNN decision boundary

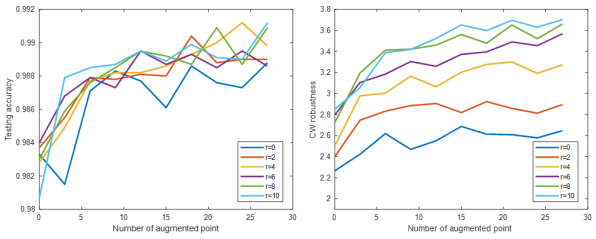
points, implying a more robust model with the training set augmented with our proposed *Bamboo* method.

Experiment

Experiment setup

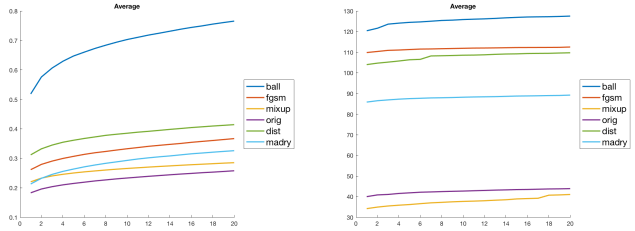
For evaluating the effect of parameter r and N on the performance of our model, we use the average strength of successful CW attack (Carlini and Wagner 2017) as the metric of robustness. When comparing with previous work, we use both CW attack strength (marked as *CW rob* in Table 1) and the testing accuracy under FGSM attack (Szegedy et al. 2013) with $\epsilon = 0.1, 0.3, 0.5$ respectively (marked as *FGSM1*, *FGSM3* and *FGSM5* in Table 1). The accuracy under 50 iterations of PGD attack (Madry et al. 2018) with $\epsilon = 0.3$ is also evaluated here (marked as *PGD3* in Table 1). We also test the accuracy under Gaussian noise with variance $\sigma = 0.5$ (marked as *GAU5* in Table 1), which demonstrates the robustness against attacks from all directions.

To visualize the effect on decision boundary, we follow the setting used in (He, Li, and Song 2018)'s work, where we use 784 random orthogonal directions for MNIST and 1000 random orthogonal directions for CIFAR-10 to linear search for decision boundary. For each testing data point, we compute the average of the top 20 smallest distance across all the testing data points, implying the overall effectiveness of different methods on increasing the robustness.



(a) Testing accuracy (b) CW robustness

Figure 2: Performance result on MNIST dataset



(a) MNIST (b) CIFAR-10

Figure 3: Decision boundary comparison

Parameter tuning

Bamboo augmentation has two hyper-parameters: the ball radius r and the ratio of the augmented data N. In figure 2a, when we fix the radius r, the testing accuracy increases as the number of augmented points grows up. Adjusting the radius has little impact on the testing accuracy. Figure 2b presents that when r is fixed, the robustness improves as N increases. The effectiveness of further increasing N becomes less as N gets larger. Under the same data amount, increasing the radius r can also enhance the robustness, while the effectiveness of increasing r saturates as r gets larger.

Boundary visualization

Figure 3 shows the top 20 smallest decision boundary on random orthogonal directions average across MNIST and CIFAR-10 testing points respectively. Comparing to previous methods, our *Bamboo* data augmentation can provide largest gain on robustness for the most vulnerable directions.

Performance comparison

Table 1 summarizes the performance of the DNN model trained with *Bamboo* comparing to other methods. *Bamboo* achieves the highest robustness under CW attack on both MNIST and CIFAR-10 experiments, and the lowest accuracy drop when facing Gaussian noise. *Bamboo* demonstrates a higher robustness against a wide range of attacking methods and the performance of our method is less sensitive to the change of the attacking strength. Also, the overall performance of *Bamboo* is better than Mixup with the same amount of data augmented. All these observations lead to the conclusion that our proposed *Bamboo* method can effectively improve the overall robustness of DNN models, no matter which kind of attack is applied or which direction of noise is added. The ImageNet experiment results showed in Table 2 show the same trend as well.

Table 1: **Performance comparison**: bold type marks the best performance, and italics type marks the second from the best performance

MNIST	Original	FGSM $\epsilon = 0.5$	DIST $c = 0.01$	PGD $\epsilon = 0.3$, 50 iterations	Mixup $\alpha = 0.12$, 10×data	**Ours** $r = 8$, 10×data
CW rob	2.442	2.390	2.5010	2.343	*2.803*	**3.554**
Test acc	0.9818	0.9817	0.9873	0.9869	**0.9904**	**0.9904**
FGSM1 acc	0.5382	0.6375	*0.8542*	0.7511	0.8323	**0.9292**
FGSM3 acc	0.2606	**0.8963**	0.1169	*0.5840*	0.2623	0.5558
FGSM5 acc	0.1423	**0.9390**	0.0244	0.1340	0.1344	*0.2878*
PGD 3 acc	0.0126	0.0258	0.0065	**0.2534**	0.0180	*0.1281*
GAU 5 acc	*0.6358*	0.6316	0.5735	0.5886	0.5813	**0.9556**
CIFAR-10	Original	FGSM $\epsilon = 0.5$	DIST $c = 0.01$	PGD $\epsilon = 0.3$, 50 iterations	Mixup $\alpha = 0.12$, 16×data	**Ours** $r = 10$, 16×data
CW rob	38.010	38.210	*38.503*	38.108	37.648	**38.746**
Test acc	*0.8395*	0.7995	0.7935	0.7791	**0.8521**	0.8249
FGSM1 acc	0.4922	0.4927	0.3825	0.4588	**0.7483**	*0.6853*
FGSM3 acc	0.4463	0.6517	0.2241	0.3848	**0.7287**	*0.6806*
FGSM5 acc	0.4093	**0.7572**	0.1998	0.3405	*0.7192*	0.6721
PGD 3 acc	0.2987	0.2233	0.1871	**0.5291**	*0.5018*	0.4111
GAU 5 acc	0.3701	*0.6356*	0.6169	0.5390	0.3371	**0.6961**

Table 2: **Performance comparison on ImageNet**

	Original	Mixup	**Ours**
Top-1 acc	57.336	58.213	**60.520**
Top-5 acc	80.647	81.452	**83.216**
Top-1 FGSM	11.342	12.947	**14.062**
Top-5 FGSM	22.860	26.400	**26.562**

Conclusion and future work

In this work we propose *Bamboo*, the first data augmentation method that is specially designed for improving the overall robustness of DNNs. Without making any assumption on the distribution of adversarial examples, *Bamboo* is able to effectively improve the robustness of DNN models against different kinds of attacks, and can achieve stable performance on large DNN models or facing strong adversarial attacks.

In future work we will discuss the theoretical relationship between the resulted DNN robustness and the parameters in our method, and how will the change in the scale of the classification problem affect such relationship. We will also propose new training tricks better suited for training with augmented dataset.

References

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

[Chapelle et al. 2001] Chapelle, O.; Weston, J.; Bottou, L.; and Vapnik, V. 2001. Vicinal risk minimization. In *Advances in neural information processing systems*, 416–422.

[Fawzi, Moosavi-Dezfooli, and Frossard 2016] Fawzi, A.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2016. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, 1632–1640.

[Goodfellow, Shlens, and Szegedy 2014] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[He, Li, and Song 2018] He, W.; Li, B.; and Song, D. 2018. Decision boundary analysis of adversarial examples. In *International Conference on Learning Representations*.

[Hein and Andriushchenko 2017] Hein, M., and Andriushchenko, M. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, 2263–2273.

[LeCun et al. 1998] LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

[Madry et al. 2018] Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

[Papernot et al. 2016] Papernot, N.; McDaniel, P. D.; Goodfellow, I. J.; Jha, S.; Celik, Z. B.; and Swami, A. 2016. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR* abs/1602.02697.

[Peck et al. 2017] Peck, J.; Roels, J.; Goossens, B.; and Saeys, Y. 2017. Lower bounds on the robustness to adversarial perturbations. In *Advances in Neural Information Processing Systems*, 804–813.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Sinha, Namkoong, and Duchi 2017] Sinha, A.; Namkoong, H.; and Duchi, J. 2017. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

[Szegedy et al. 2013] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199.

[Yan, Guo, and Zhang 2018] Yan, Z.; Guo, Y.; and Zhang, C. 2018. Deepdefense: Training deep neural networks with improved robustness. *arXiv preprint arXiv:1803.00404*.

[Zhang et al. 2017] Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.