Self-Progressing Robust Training

Minhao Cheng,[1,2] Pin-Yu Chen,[2] Sijia Liu,[3] Shiyu Chang,[2] Cho-Jui Hsieh,[1] Payel Das[2]

[1] Department of Computer Science, UCLA [2] IBM Research
[3] Department of Computer Science and Engineering, Michigan State University
{mhcheng,chohsieh}@cs.ucla.edu, {pin-yu.chen,shiyu.chang,daspa}@ibm.com, liusiji5@msu.edu

Abstract

Enhancing model robustness under new and even adversarial environments is a crucial milestone toward building trustworthy machine learning systems. Current robust training methods such as adversarial training explicitly uses an "attack" (e.g., L-inf-norm bounded perturbation) to generate adversarial examples during model training for improving adversarial robustness. In this paper, we take a different perspective and propose a new framework called SPROUT, self-progressing robust training. During model training, SPROUT progressively adjusts training label distribution via our proposed parametrized label smoothing technique, making training free of attack generation and more scalable. We also motivate SPROUT using a general formulation based on vicinity risk minimization, which includes many robust training methods as special cases. Compared with state-of-the-art adversarial training methods (PGD-L-inf and TRADES) under L-inf-norm bounded attacks and various invariance tests, SPROUT consistently attains superior performance and is more scalable to large neural networks. Our results shed new light on scalable, effective and attack-independent robust training methods.

Introduction

While deep neural networks (DNNs) have achieved unprecedented performance on a variety of datasets and across domains, developing better training algorithms that are capable of strengthening model robustness is the next crucial milestone toward trustworthy and reliable machine learning systems. In recent years, DNNs trained by standard algorithms (i.e., the natural models) are shown to be vulnerable to adversarial input perturbations (Biggio et al. 2013; Szegedy et al. 2014). Adversarial examples crafted by designed input perturbations can easily cause erroneous decision making of natural models (Goodfellow, Shlens, and Szegedy 2015) and thus intensify the demand for robust training methods.

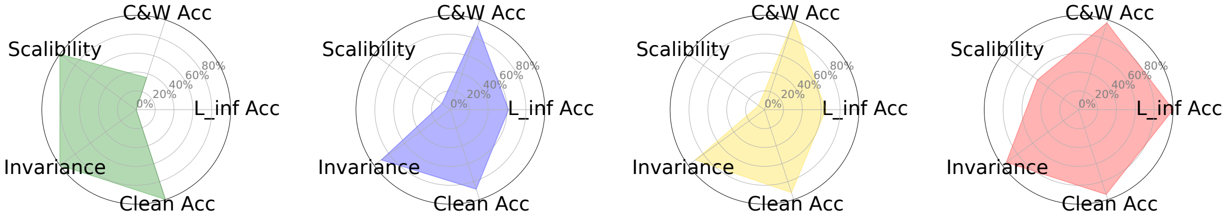
State-of-the-art robust training algorithms are primarily based on the methodology of adversarial training (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018), which calls specific attack algorithms to generate adversarial examples during model training for learning robust models. Albeit effective, these methods have the following limitations: (i) *poor scalability* – the process of generating adversarial examples incurs considerable computation overhead. For instance, our experiments show that, with the same computation resources, standard adversarial training (with 7 attack iterations per sample in every minibatch) of Wide ResNet on CIFAR-10 consumes 10 times more clock time per training epoch when compared with standard training; (ii) *attack specificity* – adversarially trained models are usually most effective against the same attack they trained on, and the robustness may not generalize well to other types of attacks (Tramèr and Boneh 2019; Kang et al. 2019); (iii) *preference toward wider network* – adversarial training is more effective when the networks have sufficient capacity (e.g., having more neurons in network layers) (Madry et al. 2018).

To address the aforementioned limitations, in this paper we propose a new robust training method named SPROUT, which is short for self-progressing robust training. We motivate SPROUT by introducing a general framework that formulates robust training objectives via vicinity risk minimization (VRM), which includes many robust training methods as special cases. It is worth noting that the robust training methodology of SPROUT is fundamentally different from adversarial training, as SPROUT features self-adjusted label distribution during training instead of attack generation. In addition to our proposed parametrized label smoothing technique for progressive adjustment of training label distribution, SPROUT also adopts Gaussian augmentation and Mixup (Zhang et al. 2018) to further enhance robustness. We show that they offer a complementary gain in robustness. In contrast to adversarial training, SPROUT spares the need for attack generation and thus makes its training scalable by a significant factor, while attaining better or comparable robustness performance on a variety of experiments. We also show exclusive features of SPROUT in terms of the novel findings that it can find robust models from either randomly initialized models or pretrained models, and its robustness performance is less sensitive to network width. Our implementation is publicly available. [1].

Contributions

Multi-dimensional performance enhancement. To illustrate the advantage of SPROUT over adversarial training

[1] Code available at https://github.com/IBM/SPROUT



(a) Natural

(b) Adversarial training

(c) TRADES

(d) SPROUT (ours)

Figure 1: Multi-dimensional performance comparison of four training methods using VGG-16 network and CIFAR-10 dataset. All dimensions are separately normalized by the best-performance method. The average score of each method is 0.6718 for natural (standard training), 0.6900 for PGD-ℓ_∞ based adversarial training (Madry et al. 2018), 0.7107 for PGD-ℓ_∞ based TRADES (Zhang et al. 2019), and 0.8798 for SPROUT (ours). The exact numbers are reported in Appendix.

and its variations, Figure 1 compares the model performance of different training methods with the following five dimensions summarized from our experimental results: (i) *Clean Acc* – standard test accuracy, (ii) *L_inf Acc* – accuracy under ℓ_∞-norm projected gradient descent (PGD) attack (Madry et al. 2018), (iii) *C&W Acc* – accuracy under ℓ_2-norm Carlini-Wagner (C&W) attack, (iv) *scalability* – per epoch clock run-time, and (v) *invariance* – invariant transformation tests including rotation, brightness, contrast and gray images. Comparing to PGD-ℓ_∞ based adversarial training (Madry et al. 2018) and TRADES (Zhang et al. 2019), SPROUT attains at least 20% better L_inf Acc, 2% better Clean Acc, 5× faster run-time (scalability), 2% better invariance, while maintaining C&W Acc, suggesting a new robust training paradigm that is scalable and comprehensive.

We further summarize our main contributions as follows:

- We propose SPROUT, a self-progressing robust training method composed of three modules that are efficient and free of attack generation: parametrized label smoothing, Gaussian augmentation, and Mixup. They altogether attain the state-of-the-art robustness performance and are scalable to large-scale networks. We will show that these modules are complementary to enhancing robustness. We also perform an ablation study to demonstrate that our proposed parametrized label smoothing technique contributes to the major gain in boosting robustness.
- To provide technical explanations for SPROUT, we motivate its training methodology based on the framework of vicinity risk minimization (VRM). We show that many robust training methods, including attack-specific and attack-independent approaches, can be characterized as a specific form of VRM. The superior performance of SPROUT provides new insights on developing efficient robust training methods and theoretical analysis via VRM.
- We evaluate the multi-dimensional performance of different training methods on (wide) ResNet and VGG networks using CIFAR-10 and ImageNet datasets. Notably, although SPROUT is attack-independent during training, we find that SPROUT significantly outperforms two major adversarial training methods, PGD-ℓ_∞ adversarial training (Madry et al. 2018) and TRADES (Zhang et al. 2019), against the same type of attacks they used during training. Moreover, SPROUT is more scalable and runs at least 5× faster than adversarial training methods. It also attains higher clean accuracy, generalizes better to various invariance tests, and is less sensitive to network width.

Related Work

Attack-specific robust training. The seminal work of adversarial training with a first-order attack algorithm for generating adversarial examples (Madry et al. 2018) has greatly improved adversarial robustness under the same threat model (e.g., ℓ_∞-norm bounded perturbations) as the attack algorithm. It has since inspired many advanced adversarial training algorithms with improved robustness. For instance, TRADES (Zhang et al. 2019) is designed to minimize a theoretically-driven upper bound on prediction error ofadversarial examples. (Liu and Hsieh 2019) combined adversarial training with GAN to further improve robustness. Bilateral adversarial training (Wang and Zhang 2019) finds robust models by adversarially perturbing the data samples and as well as the associated data labels. A feature-scattering based adversarial training method is proposed in (Zhang and Wang 2019). Different from attack-specific robust training methods, our proposed SPROUT is free of attack generation, yet it can outperform attack-specific methods. Another line of recent works uses an adversarially trained model along with additional unlabeled data (Carmon et al. 2019; Stanforth et al. 2019) or self-supervised learning with adversarial examples (Hendrycks et al. 2019) to improve robustness, which in principle can also be used in SPROUT but is beyond the scope of this paper.

Attack-independent robust training. Here we discuss related works on Gaussian data augmentation, Mixup and label smoothing. Gaussian data augmentation during training is a commonly used baseline method to improve model robustness (Zantedeschi, Nicolae, and Rawat 2017). (Liu et al. 2018a,b, 2019) demonstrated that additive Gaussian noise at both input and intermediate layers can improve robustness. (Cohen, Rosenfeld, and Kolter 2019) showed that Gaussian augmentation at the input layer can lead to certified robustness, which can also be incorporated in the training objective (Zhai et al. 2020). Mixup (Zhang et al. 2018) and its variants (Verma et al. 2018; Thulasidasan et al. 2019) are

Methods	$g(\cdot)$	$h(\cdot)$	$\tilde{\boldsymbol{x}}$	$\tilde{\boldsymbol{y}}$
Natural	$\mathcal{I}$	$\mathcal{I}$	$\boldsymbol{x}$	$\boldsymbol{y}$
GA	$\mathcal{I}$	$\mathcal{I}$	$\mathcal{N}(\boldsymbol{x}, \Delta^2)$	$\boldsymbol{y}$
LS	$\mathcal{I}$	$(1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{u}$	$\boldsymbol{x}$	$\boldsymbol{y}$
Adv Train	$\mathcal{I}$	$\mathcal{I}$	$\text{PGD}_\epsilon(\boldsymbol{x})$	$\boldsymbol{y}$
TRADES	$\mathcal{I}$	$(1-\alpha)\boldsymbol{y} + \alpha f(\tilde{\boldsymbol{x}})$	$\text{PGD}_\epsilon(\boldsymbol{x})$	$\boldsymbol{y}$
ST	$f(\boldsymbol{x}) \circ f(\tilde{\boldsymbol{x}})$	$\mathcal{I}$	$\mathcal{N}(\boldsymbol{x}, \Delta^2)$	$\boldsymbol{y}$
Mixup	$\mathcal{I}$	$\mathcal{I}$	$(1-\lambda)\boldsymbol{x}_i + \lambda\boldsymbol{x}_j$	$(1-\lambda)\boldsymbol{y}_i + \lambda\boldsymbol{y}_j$
LS+GA	$\mathcal{I}$	$(1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{u}$	$\mathcal{N}(\boldsymbol{x}, \Delta^2)$	$\boldsymbol{y}$
BAT	$\mathcal{I}$	$\mathcal{I}$	$\text{PGD}_\epsilon(\boldsymbol{x})$ (one or two step)	$(1-\alpha)\boldsymbol{y}_i + \alpha\text{PGD}_{\epsilon'}(\boldsymbol{y})$
SPROUT	$\mathcal{I}$	Dirichlet$((1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{\beta})$	$(1-\lambda)\mathcal{N}(\boldsymbol{x}_i, \Delta^2) + \lambda\mathcal{N}(\boldsymbol{x}_j, \Delta^2)$	$(1-\lambda)\boldsymbol{y}_i + \lambda\boldsymbol{y}_j$

Table 1: Summary of robust training methods using VRM formulation in (4). $\text{PGD}_\epsilon(\cdot)$ means (multi-step) PGD attack with perturbation budget ϵ. Dirichlet$(\mathbf{b})$ is the Dirichlet distribution parameterized by $\mathbf{b}$. GA/LS stands for Gaussian-Augmentation/Label-Smoothing.

a recently proposed approach to improve model robustness and generalization by training a model on convex combinations of data sample pairs and their labels. Label smoothing was originally proposed in (Szegedy et al. 2016) as a regularizer to stabilize model training. The main idea is to replace one-hot encoded labels by assigning non-zero (e.g., uniform) weights to every label other than the original training label. Although label smoothing is also shown to benefit model robustness (Shafahi et al. 2019; Goibert and Dohmatob 2019), its robustness gain is relatively marginal when compared to adversarial training. In contrast to currently used static (i.e., pre-defined) label smoothing functions, in SPROUT we propose a novel parametrized label smoothing scheme, which enables adaptive sampling of training labels from a parameterized distribution on the label simplex. The parameters of the label distribution are progressively adjusted according to the updates of model weights.

General Framework for Robust Training

The task of supervised learning is essentially learning a K-class classification function $f \in \mathcal{F}$ that has a desirable mapping between a data sample $\boldsymbol{x} \in \mathcal{X}$ and the corresponding label $\boldsymbol{y} \in \mathcal{Y}$. Consider a loss function L that penalizes the difference between the prediction $f(\boldsymbol{x})$ and the true label $\boldsymbol{y}$ from an unknown data distribution P, $(\boldsymbol{x}, \boldsymbol{y}) \sim P$. The population risk can be expressed as

$$R(f) = \int L(f(\boldsymbol{x}), \boldsymbol{y}) P(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \tag{1}$$

However, as the distribution P is unknown, in practice machine learning uses empirical risk minimization (ERM) with the empirical data distribution of n training data $\{x_i, y_i\}_{i=1}^n$

$$P_\delta(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\boldsymbol{x} = \boldsymbol{x}_i, \boldsymbol{y} = \boldsymbol{y}_i) \tag{2}$$

to approximate $P(\boldsymbol{x}, \boldsymbol{y})$, where δ is a Dirac mass. Notably, a more principled approach is to use Vicinity Risk Minimization (VRM) (Chapelle et al. 2001), defined as

$$P_\nu(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} | \boldsymbol{x}_i, \boldsymbol{y}_i) \tag{3}$$

where ν is a vicinity distribution that measures the probability of finding the virtual sample-label pair $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ in the vicinity of the training pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Therefore, ERM can be viewed as a special case of VRM when $\nu = \delta$. VRM has also been used to motivate Mixup training (Zhang et al. 2018). Based on VRM, we propose a general framework that encompasses the objectives of many robust training methods as the following generalized cross entropy loss:

$$H(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, f) = -\sum_{k=1}^{K} [\log g(f(\tilde{\boldsymbol{x}})_k)] h(\tilde{y}_k) \tag{4}$$

where $f(\tilde{\boldsymbol{x}})_k$ is the model's k-th class prediction probability on $\tilde{\boldsymbol{x}}$, $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a mapping adjusting the probability output, and $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is a mapping adjusting the training label distribution. When $\tilde{\boldsymbol{x}} = \boldsymbol{x}$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}$ and $g = h = \mathcal{I}$, where $\mathcal{I}$ denotes the identity mapping function, the loss in (4) degenerates to the conventional cross entropy loss.

Based on the general VRM loss formulation in (4), in Table 1 we summarize a large body of robust training methods in terms of different expressions of $g(\cdot)$, $h(\cdot)$ and $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$.

For example, the vanilla adversarial training in (Madry et al. 2018) aims to minimize the loss of adversarial examples generated by the (multi-step) PGD attack with perturbation budget ϵ, denoted by $\text{PGD}_\epsilon(\cdot)$. Its training objective can be rewritten as $\tilde{\boldsymbol{x}} = \text{PGD}_\epsilon(\boldsymbol{x})$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}$ and $g = h = \mathcal{I}$. In addition to adversarial training only on perturbed samples of $\boldsymbol{x}$, Wang and Zhang (2019) designs adversarial label perturbation where it uses $\tilde{\boldsymbol{x}} = \text{PGD}_\epsilon(\boldsymbol{x})$, $\tilde{\boldsymbol{y}} = (1-\alpha)\boldsymbol{y} + \alpha\text{PGD}_\epsilon(\boldsymbol{y})$, and $\alpha \in [0, 1]$ is a mixing parameter. TRADES (Zhang et al. 2019) improves adversarial training with an additional regularization on the clean examples, which is equivalent to replacing the label mapping function $h(\cdot)$ from identity to $(1-\alpha)\boldsymbol{y} + \alpha f(\tilde{\boldsymbol{x}})$. Label smoothing (LS) alone is equivalent to the setup that $g = \mathcal{I}$, $\tilde{\boldsymbol{x}} = \boldsymbol{x}$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}$ and $h(\cdot) = (1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{u}$, where $\boldsymbol{u}$ is often set as a uniform vector with value $1/K$ for a K-class supervised learning task. Joint training with Gaussian augmentation (GA) and label smoothing (LS) as studied in (Shafahi et al. 2019) is equivalent to the case when $\tilde{\boldsymbol{x}} = \mathcal{N}(\boldsymbol{x}, \Delta^2)$, $\tilde{\boldsymbol{y}} = \boldsymbol{y}$, $g = \mathcal{I}$ and $h(\boldsymbol{y}) = (1-\alpha)\boldsymbol{y} + \alpha/K$. We defer the connection between

SPROUT and VRM to the next section.

SPROUT: Scalable Robust Training

In this section, we formally introduce SPROUT, a novel robust training method that automatically finds a better vicinal risk function during training in a self-progressing manner.

Self-Progressing Parametrized Label Smoothing

To stabilize training and improve model generalization, Szegedy et al. (2016) introduces label smoothing that converts one-hot label vectors into one-warm vectors representing low-confidence classification, in order to prevent a model from making over-confident predictions. Specifically, the one-hot encoded label $\boldsymbol{y}$ is smoothed using

$$\tilde{\boldsymbol{y}} = (1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{u} \tag{5}$$

where $\alpha \in [0,1]$ is the smoothing parameter. A common choice is the uniform distribution $\boldsymbol{u} = \frac{1}{K}$, where K is the number of classes. Later works (Wang and Zhang 2019; Goibert and Dohmatob 2019) use an attack-driven label smoothing function $\boldsymbol{u}$ to further improve adversarial robustness. However, both uniform and attack-driven label smoothing disregard the inherent correlation between labels. To address the label correlation, we propose to use the Dirichlet distribution parametrized by $\boldsymbol{\beta} \in \mathbb{R}_+^K$ for label smoothing. Our SPROUT learns to update $\boldsymbol{\beta}$ to find a training label distribution that is most uncertain to a given model θ, by solving

$$\max_{\boldsymbol{\beta}} L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \boldsymbol{\beta}; \theta) \tag{6}$$

where $\tilde{\boldsymbol{y}} = \text{Dirichlet}((1-\alpha)\boldsymbol{y} + \alpha\boldsymbol{\beta})$. Notably, instead of using a pre-defined or attack-driven function for $\boldsymbol{u}$ in label smoothing, our Dirichlet approach automatically finds a label simplex by optimizing $\boldsymbol{\beta}$. Dirichlet distribution indeed takes label correlation into consideration as its generated label $\boldsymbol{z} = [z_1, \ldots, z_K]$ has the statistical properties

$$\mathbb{E}[z_s] = \frac{\beta_s}{\beta_0},\ \text{Cov}[z_s, z_t] = \frac{-\beta_s\beta_t}{\beta_0^2(\beta_0+1)},\ \sum_{s=1}^{K} z_s = 1 \tag{7}$$

where $\beta_0 = \sum_{k=1}^K \beta_k$ and $s, t \in \{1, \ldots, K\}$, $s \neq t$. Moreover, one-hot label and uniform label smoothing are our special cases when $\boldsymbol{\beta} = \boldsymbol{y}$ and $\boldsymbol{\beta} = \boldsymbol{u}$, respectively. Our Dirichlet label smoothing co-trains with the update in model weights θ during training (see Algorithm 1).

Gaussian Data Augmentation and Mixup

Gaussian augmentation. Adding Gaussian noise to data samples during training is a common practice to improve model robustness. Its corresponding vicinal function is the Gaussian vicinity function $\nu(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{x}_i, \Delta^2)\delta(\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i)$, where Δ^2 is the variance of a standard normal random vector. However, the gain of Gaussian augmentation in robustness is marginal when compared with adversarial training (see our ablation study). Shafahi et al. (2019) finds that combining uniform or attack-driven label smoothing with Gaussian smoothing can further improve adversarial robustness. Therefore, we propose to incorporate Gaussian augmentaion with Dirichlet label smoothing. The joint vicinity function becomes $\nu(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{x}_i, \Delta^2)\delta(\tilde{\boldsymbol{y}}_i = \text{Dirichlet}((1-\alpha)\boldsymbol{y}_i + \alpha\boldsymbol{\beta}))$. Training with this vicinity function means drawing labels from the Dirichlet distribution for the original data sample $\boldsymbol{x}_i$ and its neighborhood characterized by Gaussian augmentation.

Algorithm 1 SPROUT algorithm

```
Input: Training dataset $(X, Y)$, Mixup parameter $\lambda$, Gaussian augmentation variance $\Delta^2$, model learning rate $\gamma_\theta$, Dirichlet label smoothing learning rate $\gamma_{\boldsymbol{\beta}}$ and parameter $\alpha$, generalized cross entropy loss $L$
Initial model $\theta$: random initialization (train from scratch) or pre-trained model checkpoint
Initial $\boldsymbol{\beta}$: random initialization
for epoch=$1, \ldots, N$ do
  for minibatch $X_B \subset X, Y_B \subset Y$ do
    $X_B \leftarrow \mathcal{N}(X_B, \Delta^2)$
    $X_{mix}, Y_{mix} \leftarrow$ Mixup$(X_B, Y_B, \lambda)$
    $Y_{mix} \leftarrow$ Dirichlet$(\alpha Y_{mix} + (1-\alpha)\boldsymbol{\beta})$
    $g_\theta \leftarrow \nabla_\theta L(X_{mix}, Y_{mix}, \theta)$
    $g_{\boldsymbol{\beta}} \leftarrow \nabla_{\boldsymbol{\beta}} L(X_{mix}, Y_{mix}, \theta)$
    $\theta \leftarrow \theta - \gamma_\theta g_\theta$
    $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \gamma_{\boldsymbol{\beta}} g_{\boldsymbol{\beta}}$
  end for
end for
return $\theta$
```

Mixup. To further improve model generalization, SPROUT also integrates Mixup (Zhang et al. 2018) that performs convex combination on pairs of training data samples (in a minibatch) and their labels during training. The vicinity function of Mixup is $\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}|\boldsymbol{x}_i, \boldsymbol{y}_i) = \delta(\tilde{\boldsymbol{x}} = (1-\lambda)\boldsymbol{x}_i + \lambda\boldsymbol{x}_j, \tilde{\boldsymbol{y}} = (1-\lambda)\boldsymbol{y}_i + \lambda\boldsymbol{y}_j)$, where $\lambda \sim \text{Beta}(a, a)$ is the mixing parameter drawn from the Beta distribution and $a > 0$ is the shape parameter. The Mixup vicinity function can be generalized to multiple data sample pairs. Unlike Gaussian augmentation which is irrespective of the label (i.e., only adding noise to $\boldsymbol{x}_i$), Mixup aims to augment data samples on the line segments of training data pairs and assign them convexly combined labels during training.

Vicinity function of SPROUT. With the aforementioned techniques integrated in SPROUT, the overall vicinity function of SPROUT can be summarized as $\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\beta}) = \delta(\tilde{\boldsymbol{x}} = \lambda\mathcal{N}(\boldsymbol{x}_i, \Delta^2) + (1-\lambda)\mathcal{N}(\boldsymbol{x}_j, \Delta^2),\ \tilde{\boldsymbol{y}} = \text{Dirichlet}((1-\alpha)((1-\lambda)\boldsymbol{y}_i + \lambda\boldsymbol{y}_j) + \alpha\boldsymbol{\beta})$.

In the experiment, we will show that Dirichlet label smoothing, Gaussian augmentation and Mixup are complementary to enhancing robustness by showing their diversity in input gradients.

SPROUT Algorithm

Using the VRM framework, the training objective of SPROUT is

$$\min_{\theta} \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} L(\nu(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\beta}); \theta), \tag{8}$$

where θ denotes the model weights, n is the number of training data, L is the generalized cross entropy loss defined in (4) and $\nu(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\beta})$ is the vicinity function of SPROUT. Our SPROUT algorithm co-trains θ and $\boldsymbol{\beta}$ via stochastic gradient descent/ascent to solve the outer minimization problem on θ and the inner maximization problem on $\boldsymbol{\beta}$. In particular, for calculating the gradient $g_{\boldsymbol{\beta}}$ of the parameter $\boldsymbol{\beta}$, we use the Pytorch implementation based on (Figurnov, Mohamed, and Mnih 2018). SPROUT can either train a model from scratch with randomly initialized θ or strengthen a pre-trained model. We find that training from either randomly initialized or pre-trained natural models using SPROUT can yield substantially robust models that are resilient to large perturbations (see Appendix). The training steps of SPROUT are summarized in Algorithm 1.

We also note that our min-max training methodology is different from the min-max formulation in adversarial training (Madry et al. 2018), which is $\min_\theta \sum_{i=1}^{n} \max_{\boldsymbol{\delta}_i : \|\boldsymbol{\delta}_i\|_p \leq \epsilon} L(\boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{y}_i; \theta)$, where $\|\boldsymbol{\delta}_i\|_p$ denotes the ℓ_p norm of the adversarial perturbation $\boldsymbol{\delta}_i$. While the outer minimization step for optimizing θ can be identical, the inner maximization of adversarial training requires running multi-step PGD attack to find adversarial perturbations $\{\boldsymbol{\delta}_i\}$ for each data sample in every minibatch (iteration) and epoch, which is attack-specific and time-consuming (see our scalability analysis in Table 6). On the other hand, our inner maximization is upon the Dirichlet parameter $\boldsymbol{\beta}$, which is attack-independent and only requires single-step stochastic gradient ascent with a minibatch to update $\boldsymbol{\beta}$. We have explored multi-step stochastic gradient ascent on $\boldsymbol{\beta}$ and found no significant performance enhancement but increased computation time.

Advantages of SPROUT. Comparing to adversarial training, the training of SPROUT is more efficient and scalable, as it only requires one additional back propagation to update $\boldsymbol{\beta}$ in each iteration (see Table 6 for a run-time analysis). As highlighted in Figure 1, SPROUT is also more comprehensive as it automatically improves robustness in multiple dimensions owing to its self-progressing training methodology. Moreover, we find that SPROUT significantly outperforms adversarial training and attains larger gain in robustness as network width increases (see Figure 3), which makes SPROUT a promising approach to support robust training for a much larger set of network architectures.

Performance Evaluation

Experiment Setup

Dataset and network structure. We use CIFAR-10 and ImageNet (Deng et al. 2009) for performance evaluation. For CIFAR-10, we use both standard VGG-16 (Simonyan and Zisserman 2015) and Wide ResNet. The Wide ResNet models are pre-trained PGD-ℓ_∞ robust models given by adversarial training and TRADES. For VGG-16, we implement adversarial training with the standard hyper-parameters and train TRADES using the official implementation. For ImageNet, we use ResNet-152.

Implementation details. As suggested in Mixup (Zhang et al. 2018), we set the Beta distribution parameter $a = 0.2$ when sampling the mixing parameter λ. For Gaussian augmentation, we set $\Delta = 0.1$, which is within the suggested range in (Zantedeschi, Nicolae, and Rawat 2017). Also, we set the label smoothing parameter $\alpha = 0.01$. A parameter sensitivity analysis on λ and α is given in Appendix. Unless specified otherwise, for SPROUT we set the model initialization to be a natural model. An ablation study of model initialization is given in ablation study.

Adversarial Robustness under Various Attacks

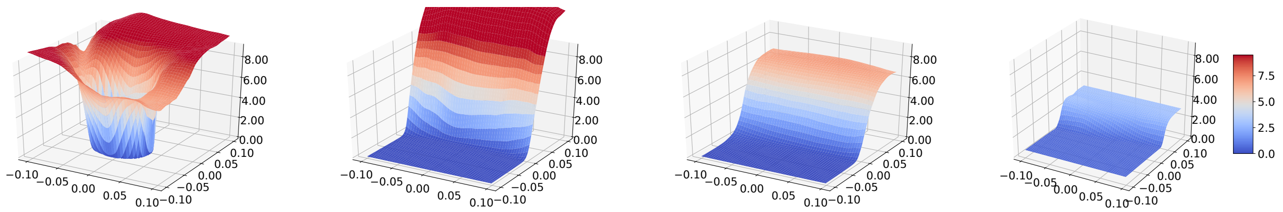
White-box attacks. On CIFAR-10, we compare the model accuracy under $\epsilon = 0.03 \approx 8/255$ strength of white-box ℓ_∞-norm bounded non-targeted PGD attack, which is considered as the strongest first-order adversary (Madry et al. 2018) with an ℓ_∞-norm constraint ϵ (normalized between 0 to 1). All PGD attacks are implemented with random starts and we run PGD attack with 20, 100 and 200 steps in our experiments. To be noted, we use both PGD^X (X-step PGD with step size $\epsilon/5$). As suggested, we test our model under different steps PGD and multiple random restarts. In Table 2, we find SPROUT achieves 62.24% and 66.23% robust accuracy on VGG16 and Wide ResNet respectively, while TRADES and adversarial training are 10-20% worse than SPROUT. Moreover, we report the results of C&W-ℓ_∞ attack (Carlini and Wagner 2017) in Appendix. Next, we compare against ℓ_2-norm based C&W attack by using the default attack setting with 10 binary search steps and 1000 iterations per step to find successful perturbations while minimizing their ℓ_2-norm. SPROUT can achieve 85.21% robust accuracy under ℓ_2 $\epsilon = 0.05$ constraint while Adv train and TRADES achieves 77.76% and 82.58% respectively. It verifies that SPROUT can improve ℓ_∞ robustness by a large margin without degrading ℓ_2 robustness. SPROUT's accuracy under C&W-ℓ_2 attack is similar to TRADES and is better than both natural and adversarial training. The results also suggest that the attack-independent and self-progressing training nature of SPROUT can prevent the drawback of failing to provide comprehensive robustness to multiple and simultaneous ℓ_p-norm attacks in adversarial training (Tramèr and Boneh 2019; Kang et al. 2019).

Transfer attack. We follow the criterion of evaluating transfer attacks in (Athalye, Carlini, and Wagner 2018) to inspect whether the models trained by SPROUT will cause the issue of obfuscated gradients and give a false sense of robustness. We generate 10,000 PGD-ℓ_∞ adversarial examples from CIFAR-10 natural models with $\epsilon = 0.03$ and evaluate their attack performance on the target model. Table 3 shows SPROUT achieves the best accuracy when compared with adversarial training and TRADES, suggesting the effectiveness of SPROUT in defending both white-box and transfer attacks.

ImageNet results. As many ImageNet class labels carry similar semantic meanings, to generate meaningful adversarial examples for robustness evaluation, here we follow the same setup as in (Athalye, Carlini, and Wagner 2018) that adopts PGD-ℓ_∞ attacks with randomly targeted labels. Table 4 compares the robust accuracy of natural and SPROUT models. SPROUT greatly improves the robust accuracy across different ϵ values. For example, when $\epsilon =$

	VGG-16					Wide-ResNet 20				
Methods	No attack	PGD^{20}	PGD^{100}	PGD^{200}	10 PGD^{100}	No attack	PGD^{20}	PGD^{100}	PGD^{200}	10 PGD^{100}
Nat train	**93.34%**	0.6%	0.1%	0.0%	0.0%	**95.93%**	0.0%	0.0%	0.0%	0.0%
Adv Train	80.32%	36.63%	36.29%	36.01%	36.8%	87.25%	45.91%	45.32%	45.02%	44.98%
TRADES	84.85%	38.81%	38.21%	37.95%	37.94%	84.92%	56.23%	56.13%	55.96%	56.01%
SPROUT	89.15%	**62.24%**	**58.93%**	**57.9%**	**58.08%**	90.56%	**66.23%**	**64.58%**	**64.30%**	**64.32%**

Table 2: The clean and robust accuracy of VGG-16 and Wide-ResNet 20 models trained by various defense methods. All robust accuracy results use $\epsilon = 0.03$ (ℓ_∞ perturbation). B PGD^A denotes an A-step PGD attack with B random restarts.



(a) Natural (b) Adv Train (c) TRADES (d) SPROUT

Figure 2: Loss landscape comparison of different training methods

Method	VGG 16	Wide ResNet
Adv Train	79.13%	85.84%
TRADES	83.53%	83.9%
SPROUT	86.28%	89.1%

Table 3: Robust accuracy of CIFAR-10 under transfer attack

Method	Clean Acc	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.015$
Natural	78.31%	37.13%	9.14%	2.12%
SPROUT	74.23%	65.24%	52.86%	35.04%

Table 4: Accuracy of ImageNet under PGD-ℓ_∞ attack

0.01, SPROUT boosts the robust accuracy of natural model by over 43%. When $\epsilon = 0.015 \approx 4/255$, a considerably large adversarial perturbation on ImageNet, SPROUT still attains about 35% robust accuracy while the natural model merely has about 2% robust accuracy. Moreover, comparing the clean accuracy, SPROUT is about 4% worse than the natural model but is substantially more robust. We omit the comparison to adversarial training methods as we are unaware of any public pre-trained robust ImageNet models of the same architecture (ResNet-152) prior to the time of our submission, and it is computationally demanding for us to train and fine-tune such large-scale networks with adversarial training. On our machine, training a natural model takes 31,158.7 seconds and training SPROUT takes 59,201.6 seconds. Comparing to the run-time analysis, SPROUT has a much better scalability than adversarial training and TRADES.

Loss Landscape Exploration

To further verify the superior robustness using SPROUT, we visualize the loss landscape of different training methods in Figure 2. Following the implementation in (Engstrom, Ilyas, and Athalye 2018), we vary the data input along a linear space defined by the sign of the input gradient and a random Rademacher vector, where the x- and y- axes represent the magnitude of the perturbation added in each direction and the z-axis represents the loss. One can observe that the loss surface of SPROUT is smoother. Furthermore, it attains smaller loss variation compared with other robust training methods. The results provide strong evidence for the capability of finding more robust models via SPROUT.

Invariance Test

In addition to ℓ_p-norm bounded adversarial attacks, here we also evaluate model robustness against different kinds of input transformations using CIFAR-10 and Wide ResNet. Specifically, we change rotation (with 10 degrees), brightness (increase the brightness factor to 1.5), contrast (increase the contrast factor to 2) and make inputs into grayscale (average all RGB pixel values). The model accuracy under these invariance tests is summarized in Table 5. The results show that SPROUT outperforms adversarial training and TRADES. Interestingly, natural model attains the best accuracy despite the fact that it lacks adversarial robustness, suggesting a potential trade-off between accuracy in these invariance tests and ℓ_p-norm based adversarial robustness.

Method	Rotation	Brightness	Contrast	Gray
Natural	88.21%	93.4%	91.88 %	91.95%
Adv Train	82.66%	83.64%	84.99%	81.08%
TRADES	80.81%	81.5 %	83.08%	79.27%
SPROUT	85.95%	88.26 %	86.98%	81.64%

Table 5: Accuracy under invariance tests

Methods	CIFAR-10	
	VGG 16	Wide ResNet
Natural	146.7	1449.6
Adv Train	1327.1	14246.1
TRADES	1932.5	22438.4
SPROUT	271.7	2727.8

Table 6: Training-time (seconds) for 10 epochs

Scalability

SPROUT enjoys great scalability over adversarial training based algorithms because its training requires much less number of back-propagations per iteration, which is a dominating factor that contributes to considerable run-time in adversarial training. Table 6 benchmarks the run-time of different training methods for 10 epochs. On CIFAR-10, the run-time of adverarial training and TRADES is about 5× more than SPROUT.

Ablation Study

Dissecting SPROUT. Here we perform an ablation study using VGG-16 and CIFAR-10 to investigate and factorize the robustness gain in SPROUT's three modules: Dirichlet label smoothing (Dirichlet), Gaussian augmentation (GA) and Mixup. We implement all combinations of these techniques and include uniform label smoothing (LS) (Szegedy et al. 2016) as another baseline. Their accuracies under PGD-ℓ_∞ 0.03 attack are shown in Table 7. We highlight some important findings as follows.

- Dirichlet outperforms uniform LS by a significant factor, suggesting the importance of our proposed self-progressing label smoothing in improving adversarial robustness.
- Comparing the performance of individual modules alone (GA, Mixup and Dirichlet), our proposed Dirichlet attains the best robust accuracy, suggesting its crucial role in training robust models.
- No other combinations can outperform SPROUT. Moreover, the robust gains from GA, Mixup and Dirichlet appear to be *complementary*, as SPROUT's accuracy is close to the sum of their individual accuracy. To justify their diversity in robustness, we compute the cosine similarity of their pairwise input gradients and find that they are indeed quite diverse and thus can promote robustness when used together. The details are given in Appendix.

Effect on network width. It was shown in (Madry et al. 2018) that adversarial training (Adv Train) will take effect when a network has sufficient capacity, which can be achieved by increasing network width. Figure 3 compares the robust accuracy of SPROUT and Adv Train with varying network width on Wide ResNet and CIFAR-10. When the network has width = 1 (i.e. a standard ResNet-34 network (He et al. 2016)), the robust accuracy of SPROUT and Adv Train are both relatively low (less than 47%). However, as the width increases, SPROUT soon attains significantly better robust accuracy than Adv Train by a large margin (roughly 15%). Since SPROUT is more effective in boosting robust accuracy as network width varies, the results also suggest that SPROUT can better support robust training for a broader range of network structures.

Methods	VGG 16	
	PGD^{20}	PGD^{100}
GA	12.44%	9.46%
Mixup	0.76%	0.08%
Dirichlet	29.64%	9.77%
GA+Mixup	41.88%	40.29%
Mixup+Dirichlet	17.53%	7.97%
GA+Dirichlet	55.27%	53.79%
Uniform LS	15.36%	5.12%
SPROUT	62.64%	58.93%

Table 7: Robust accuracy under ℓ_∞ 0.03 strength with different combinations of the modules in SPROUT.

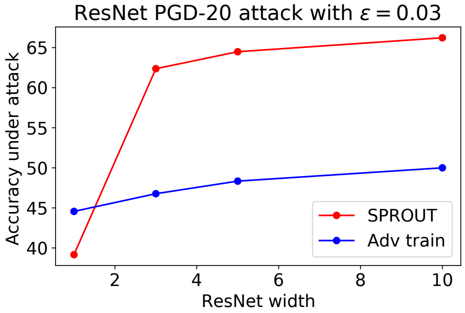


Figure 3: Effect of network width against PGD-ℓ_∞ attack on CIFAR-10 and ResNet-34.

Conclusion

This paper introduced SPROUT, a self-progressing robust training method motivated by vicinity risk minimization. When compared with state-of-the-art adversarial training based methods, our extensive experiments showed that the proposed self-progressing Dirichlet label smoothing technique in SPROUT plays a crucial role in finding substantially more robust models against ℓ_∞-norm bounded PGD attacks and simultaneously makes the corresponding model more generalizable to various invariance tests. We also find that SPROUT can strengthen a wider range of network structures as it is less sensitive to network width changes. Moreover, SPOURT's self-adjusted learning methodology not only makes its training free of attack generation but also becomes scalable solutions to large networks. Our results shed new insights on devising comprehensive and robust training methods that are attack-independent and scalable.

Acknowledgments

This work was done during Minhao Cheng's internship at IBM Research. Cho-Jui Hsieh and Minhao Cheng are partially supported by National Science Foundation (NSF) under IIS-1901527, IIS-2008173 and Army Research Lab under W911NF-20-2-0158.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Coference on International Conference on Machine Learning* .

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402.

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

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Liang, P.; and Duchi, J. C. 2019. Unlabeled data improves adversarial robustness. *Neural Information Processing Systems* .

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

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning* .

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

Engstrom, L.; Ilyas, A.; and Athalye, A. 2018. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272* .

Figurnov, M.; Mohamed, S.; and Mnih, A. 2018. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, 441–452.

Goibert, M.; and Dohmatob, E. 2019. Adversarial Robustness via Adversarial Label-Smoothing. *arXiv preprint arXiv:1906.11567* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. *International Conference on Learning Representations* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. *Neural Information Processing Systems* .

Kang, D.; Sun, Y.; Hendrycks, D.; Brown, T.; and Steinhardt, J. 2019. Testing Robustness Against Unforeseen Adversaries. *arXiv preprint arXiv:1908.08016* .

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018a. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 369–385.

Liu, X.; and Hsieh, C.-J. 2019. Rob-gan: Generator, discriminator, and adversarial attacker. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11234–11243.

Liu, X.; Li, Y.; Wu, C.; and Hsieh, C.-J. 2018b. Adv-bnn: Improved adversarial defense through robust bayesian neural network. *arXiv preprint arXiv:1810.01279* .

Liu, X.; Xiao, T.; Si, S.; Cao, Q.; Kumar, S.; and Hsieh, C.-J. 2019. Neural sde: Stabilizing neural ode networks with stochastic noise. *arXiv preprint arXiv:1906.02355* .

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

Shafahi, A.; Ghiasi, A.; Huang, F.; and Goldstein, T. 2019. Label Smoothing and Logit Squeezing: A Replacement for Adversarial Training? *arXiv preprint arXiv:1910.11585* .

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* .

Stanforth, R.; Fawzi, A.; Kohli, P.; et al. 2019. Are Labels Required for Improving Adversarial Robustness? *Neural Information Processing Systems* .

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

Thulasidasan, S.; Chennupati, G.; Bilmes, J.; Bhattacharya, T.; and Michalak, S. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. *Neural Information Processing Systems* .

Tramèr, F.; and Boneh, D. 2019. Adversarial Training and Robustness for Multiple Perturbations. *Neural Information Processing Systems* .

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Courville, A.; Lopez-Paz, D.; and Bengio, Y. 2018. Manifold mixup: Better representations by interpolating hidden states. *International Conference on Machine Learning* .

Wang, J.; and Zhang, H. 2019. Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks. *International Conference on Computer Vision* .

Zantedeschi, V.; Nicolae, M.-I.; and Rawat, A. 2017. Efficient defenses against adversarial attacks. In *ACM Workshop on Artificial Intelligence and Security*, 39–49.

Zhai, R.; Dan, C.; He, D.; Zhang, H.; Gong, B.; Ravikumar, P.; Hsieh, C.-J.; and Wang, L. 2020. Macer: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378* .

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations* .

Zhang, H.; and Wang, J. 2019. Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training. *Neural Information Processing Systems* .

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *International Conference on Machine Learning* .