GSMOOTH: GENERALIZED RANDOMIZED SMOOTHING FOR UNIVERSAL SEMANTIC TRANSFORMATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The vulnerability of deep learning models to adversarial examples and semantic perturbations has limited the applications in risk sensitive areas. The recent development of certified defense approaches like randomized smoothing provide a promising direction towards building reliable machine learning systems. However, current certified defense methods cannot certify complex semantic perturbations which are more common in practical applications. Here we propose a generalized randomized smoothing framework for certified defense on universal semantic transformations. We provide a theoretical framework and tight robustness guarantee on the generalized smoothed classifier against semantic transformations. Specifically, we use a proxy neural network to approximate a transformation and certify the transformation based on the proxy neural network which provides a powerful tool for studying the properties of complex semantic transformations. Experiments on certifying multiple semantic perturbations and corruptions on multiple datasets demonstrate the effectiveness of our approach.

1 INTRODUCTION

Though deep learning models have achieved remarkable success on various tasks (LeCun et al., 2015; He et al., 2016), it has been shown that they are easily cheated by adversarial attacks (Biggio et al., 2013; Szegedy et al., 2013; Goodfellow et al., 2014) or semantic perturbations (Hendrycks & Dietterich, 2019), which limit the applications of them on many important tasks. For example, the automatic driving systems become fragile under bad weather or light conditions and can be misleaded even by a small patch on the roadmark (Brown et al., 2017), thus they are not safe enough for practical applications.

To develop more reliable machine learning systems, researchers attempt to design defense techniques against adversarial attacks or semantic perturbations. Existing methods can be categorized into empirical defense and certified defense. Methods like adversarial training (Madry et al., 2017) or its variants (Wong et al., 2018; Zhang et al., 2019a) are the most effective empirical defense for l_p norm bounded attack on neural networks. Moreover, Calian et al. (2021) proposed adversarial augmentation to improve the empirical performance under semantic perturbations. However, focusing on a corrupted dataset, the accuracy of empirical defense is not theoretical guaranteed, and they are easily broken by new adaptive attacks methods (Tramer et al., 2020). The certified defense aims to theoretically provide a certified region where the model is theoretically safe under any attack or perturbation (Wong & Kolter, 2018; Cohen et al., 2019; Gowal et al., 2018; Zhang et al., 2019b).

Though certified defenses have achieved great success, most of them are limited to defend l_p norm bounded attacks. In practical scenarios, we often encounter semantic perturbations such as rotation, translation, blur, bad weather and so on, which do not damage on the semantic features of images and are easily recognized by humans. The l_p distance between corrupted and original images is large even the corruption is slight, so current certified defense methods are incapable of certifying such semantic perturbations. To solve the problem, several recent works (Fischer et al., 2020; Li et al., 2021; Mohapatra et al., 2020) attempt to extend the certified defenses to certifying several simple semantic corruptions like translation, rotation and Gaussian blur. Their methods make a great progress and inspire people to develop better smoothing schemes for certifying semantic perturbations. However, it is still highly challenging. Current works are limited to certifying simple cases of semantic perturbations with explicit expression and they are not scalable to certifying general

semantic peturbations. For example, the translation of images are additive and rotation and resizing both have a closed mathematical expression. But in real world settings, many semantic corruption does not closed form expression or they are black boxes and hard to be analyzed theoretically. Some deterministic certified defenses like (Mohapatra et al., 2020) that use convex relaxation for the activation function require to solve a complex optimization for computing bound which is expensive. Probabilistic approaches based on randomized smoothing also demands a handcraft Lipschitz bound which is intractable for complicated semantic transformations.

To address the challenge mentioned above, we propose a generalized randomized smoothing framework named GSmooth in this paper. At first we provide a unified generalized randomized smoothing framework for certifying universal semantic transformation. Then we categorize the transformations into resolvable transformations and non-resolvable transformations similar with (Li et al. (2021)). However, most non-resolvable transformations are complex and existing methods can not provide its certified radius. Noticing all these transformation, including resolvable and non-resolvable, can be approximated by using an image-to-image translation neural network. We show that investigating the properties like the certified bound or Lipschitz bound for the surrogate neural network is much easier than the original semantic transformations. We theoretically provide the certified radius for the proxy neural networks and use it for certifying the original transformations. Then we provide theoretical analysis and error bounding for the approximation. Finally, we verify our methods on several publicly available datasets. Extensive experimental results demonstrate that our methods are effective for certifying complex semantic transformations like different types of blur or image quality corruptions.

2 Related Work

2.1 Attack and Defense Based on Semantic Perturbation

Unlike l_p perturbation which adds small noise to every pixel of images, semantic attacks or physical attacks are usually unrestricted. Brown et al.; Song et al. uses an small patch added to the image to mislead the classifier or the object detector. Engstrom et al.; Engstrom et al.; Xiao et al. construct adversarial examples using spatial transformations like rotation or translation. Hendrycks & Dietterich shows that a wide variety of semantic perturbations degrade the performance for many deep learning models. Many works (Cubuk et al., 2019; Hendrycks et al., 2019; 2020) propose diverse data augmentation techniques to enhance robustness under semantic perturbations. Calian et al. (2021) propose adversarial data augmentation that can viewed as adversarial training for defending semantic perturbation. Beyond empirical defenses, several works (Mohapatra et al., 2020; Madry et al., 2017; Singh et al., 2019; Balunović et al., 2019) attempted to certify some simple geometric transformations. However, all of them belong to deterministic certification approaches and their performance on realistic datasets are unsatisfactory.

2.2 Randomized Smoothing

Randomized smoothing is a novel certification method originate from differential privacy (Lecuyer et al.). Cohen et al. then improved the certified bound and applied it to large scale deep neural networks and datasets. Yang et al. exhaustively analyzed the robust radius by using different noise distribution and norms. Hayes; Yang et al. pointed out that randomized smoothing suffers from curse of dimensionality for l_∞ norm. (Salman et al., 2019) adopted adversarial training to train smoothed classifiers to obtain better robustness guarantees. Li et al.; Fischer et al. extended randomized smoothing to certify some simple semantic transformations, e.g. image translation and rotation. It shows that randomized smoothing could be generalized to certify more diverse attacks or corruptions. However, their methods are limited to simple semantic transformations, which are easy to analyze their mathematical properties.

3 Proposed Method: Generalized Smoothing

In this section we present the framework and theoretical analysis of our Generalized Randomized Smoothing.

3.1 NOTATIONS

We first introduce the notations and the formulation of the problem.

Considering the p-classification situation, we denote the label space $\mathcal{Y} = \{1, 2, \dots p\}$ and our classifier $f(x) : \mathbb{R}^n \to [0,1]^p \subset \mathbb{R}^p$ which outputs probability for each class and we choose the label with the largest probability, which is defined as $\arg\max_i f(x)$. Let $\tau(\theta, x) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be a semantic transformation of raw input x with parameter $\theta \in \mathbb{R}^m$, the smoothed classifier is defined as

$$G(x) = \mathbb{E}_{\theta \sim g(\theta)}[f(\tau(\theta, x))], \tag{1}$$

which is the average score for the sample under a smoothing distribution $g(\theta)$ such as Gaussian Distribution. Suppose the noise distribution comes from the exponential family which means that $g(\theta) \propto \exp(-\psi(x))$. Let $||u|| = 1$ be any vector with unit norm and a random variable $\gamma_u = \langle u, \nabla\psi(\delta)\rangle$ where $\delta \sim g(\theta)$. The complementary CDF is $\varphi_u(c) = \mathbb{P}[\gamma_u > c]$ and inverse complementary CDF is $\varphi_u^{-1}(p) = \inf\{c|\mathbb{P}(\gamma_u > c) \leqslant p\}$. We define a function Φ as follows,

$$\Phi(p) = \max_{||u||=1} \mathbb{E}[\gamma_u \mathbb{I}\{\gamma_u > \varphi_u^{-1}(p)\}]. \tag{2}$$

Let $y_A = \operatorname{argmax}_{i \in \mathcal{Y}} G(x)_i$ be the predicted label by our smoothed classifier $G(x)$ and $y_B = \operatorname{argmax}_{i \in \mathcal{Y} \setminus y_A} G(x)_i$ is the runner-up class. Without causing confusion, we use $G(x)_A, G(x)_B$ to represent the probability of the top and the runner-up class $G(x)_{y_A}, G(x)_{y_B}$.

3.2 CERTIFIED BOUND FOR CLOSED (RESOLVABLE?) SEMANTIC TRANSFORMATIONS

We first discuss a class of transformations that are called closed(resolvable?) which means that the composition of two transformations with parameters $\theta, \xi \in \mathbb{R}^m$ is still a transformation with a new parameter $\gamma(\theta, \xi) \in \mathbb{R}^m$ where $\gamma(\cdot, \cdot)$ is a function depends on these parameters.

Theorem 1 *Let $f(x)$ be any classifier and $G(x)$ be the smoothed classifier defined in Eq.(1), if there exists a function $M(\cdot)$, and the transformation $\tau(\cdot, \cdot)$ satisfies that*

$$\frac{\partial \gamma(\theta, \xi)}{\partial \xi} = \frac{\partial \gamma(\theta, \xi)}{\partial \theta} M(\theta, \xi),$$

and there exists two constants $\underline{p_A}$, $\overline{p_B}$ that satisfies that

$$G(x)_A \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant G(x)_B,$$

then $y_A = \operatorname{argmax}_{i \in \mathcal{Y}} G(\tau(\xi, x))_i$ holds for any $\|\xi\| \leqslant R$ where

$$R = \frac{\sigma}{2M^*} \left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{3}$$

here $M^* = max_{\xi,\theta}||M(\xi, \theta)||$.

This theorem states that for a vast majority of resolvable semantic transformations, the smoothed classifier is certifiably robust for semantic transformation within a certain range. Specifically, we show two examples of the theorem which are additive transformations and commutable transformations.

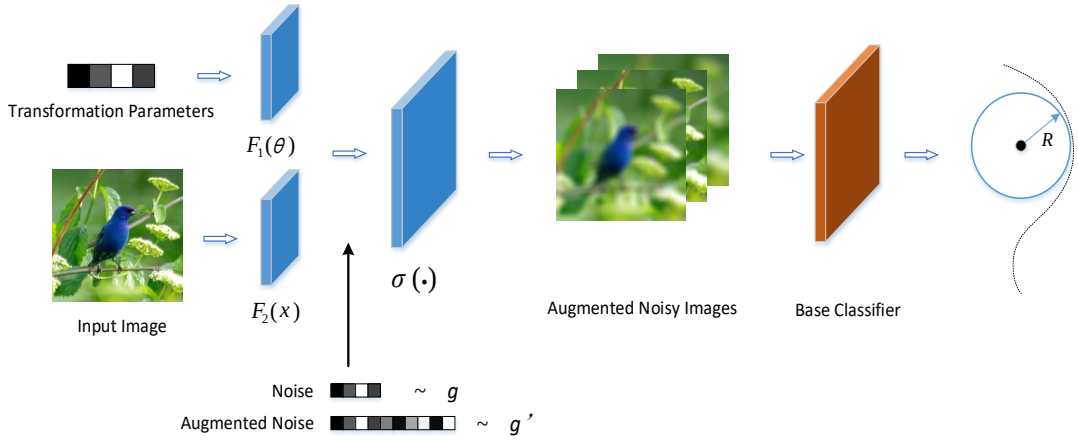
Additive Transformations. A semantic transformation is additive if it satisfies $\tau(\theta, \tau(\xi, x)) = \tau(\xi + \theta, x)$ for any $\theta, \xi \in \mathbb{R}^m$.

Commutable Transformations. A semantic transformation is commutable if it satisfies $\tau(\theta, \tau(\xi, x)) = \tau(\xi, \tau(\theta, x))$ for any $\theta, \xi \in \mathbb{R}^m$.

For these two types of transformations, it's trivial to verify that they satisfy the property proposed in theorem 1 and thus we might simply apply theorem 1 and get the certified radius

$$R = \frac{\sigma}{2} \left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right) \tag{4}$$

These two kinds of transformations include image translation and gaussian blur, which are basic semantic transformations and widely discussed in previous works ?. The certification of these simple



[]

Figure 1: A graphical illustration of our GSmooth. We use a surrogate image-to-image translation network to accurately fit a semantic transformation. Then we add a new augmented noise into the surrogate model and construct the generalized smoothed classifier. The augmented noise are sampled to fill up the residual non-resolvable part in the semantic space. We theoretically calculate the certification bound for the surrogate model to certify the original semantic transformation.

transformations only requires applying translation or gaussian blur to the sample and gets the average classification score under the noise distribution.

3.3 Certified bound for general semantic transformations

Translation and gaussian blur are only two specific cases of semantic transformation. In practice, most semantic transformations are not commutable or even not resolvable, so we need to develop better methods for certifying more types of semantic transformations. However, existing methods like Semanify-NN(Mohapatra et al., 2020) based on convex relaxation and TSS based on randomized smoothing(Li et al., 2021) require to develop a specific algorithm or manually analyzed the Lipschitz bound for each individual semantic transformation. This is not scalable and might be infeasible for more complicated transformation that might not have a explicit mathematical form. To handle the challenge and inspired by (Zhu et al., 2017) that neural networks are able to learn the style transfer which is indeed a complex and unknown semantic transformation, we propose to use a surrogate image-to-image translation model to accurately fit the semantic transformation. Our key insight is that the surrogate neural network is much easier to analyzed and they can be certified by introducing an additional noise inside its layers. By certifying the transformation based on the surrogate model, we are able to certify the original transformation if the approximation error is within an acceptable region. Our method is flexible and scalable since the surrogate neural networks have a uniform mathematical form for theoretical analysis and they are trained automatically. Then we discuss the details of our generalized smoothing framework. Specifically, we define the surrogate model as the following form that we found they have a simple certification bound,

$$\tau(\theta, x) = \sigma(F_1(\theta) + F_2(x)), \tag{5}$$

where $F_1(\cdot) : \mathbb{R}^m \to \mathbb{R}^d, F_2(\cdot) : \mathbb{R}^n \to \mathbb{R}^d, \sigma(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$ are three individual neural networks. Note that this form is equivalent to concatenate two input variables after being processed by two subnetworks. Due to the strong representation capacity of neural networks, the surrogate model could approximate the target transformation with high precision. However, even if we represent a transformation by a surrogate neural network, it does not turn a non-resolvable transformation into

a resolvable one. To handle the challenge, we propose a novel smoothed strategy by augmenting the dimension of both the data and the transformation parameters. Specifically, we introduce the augmented data $\tilde{x}$ and the augmented parameter $\tilde{\theta}$ as

$$\tilde{x} = \begin{pmatrix} x' \\ x \end{pmatrix}, \tilde{\theta} = \begin{pmatrix} \theta \\ \theta' \end{pmatrix}, \tag{6}$$

the additional parameters $\theta' \in \mathbb{R}^n$ are sampled from $g'(\theta')$, and the joint distribution of θ' and θ is $\tilde{\theta} \sim \tilde{g}(\tilde{\theta})$. The augmented distribution is represented as $\tilde{g}(\tilde{\theta}) = g'(\theta')g(\theta)$. We define the generalized smoothed classifier as

$$\tilde{G}(\tilde{x}) = \mathbb{E}_{\tilde{\theta}\sim\tilde{g}(\tilde{\theta})}[\tilde{f}(\tilde{\tau}(\tilde{\theta}, \tilde{x}))], \tag{7}$$

where $\tilde{f}$ is the "augmented target classifier" that equals the original classifier when constrained on the original input x. This can be achieved by setting the weights to additional dimensions to 0. Note that now all the functions are augmented for a $n + m$ dimensional input. We still use a surrogate neural network to represent the augmented transformation $\tilde{\tau} : R^{m+n} \rightarrow ?$,

$$\tilde{\tau}(\tilde{\theta}, \tilde{x}) = \tilde{\sigma}(\widetilde{F_1}(\tilde{\theta}) + \widetilde{F_2}(\tilde{x})), \tag{8}$$

where $\tilde{\sigma}(\cdot), \tilde{F_1}(\cdot), \tilde{F_2}(\cdot)$ are parts of the augmented surrogate model. By carefully designing the interaction between the augmented parameters and the original parameters, we could turn the transformation to a resolvable one and it does not change the original surrogate model when constraining to the original input x and θ. Specifically, we designed the function $\widetilde{F_1}$ and $\widetilde{F_2}$ as follows,

$$\widetilde{F_1}(\tilde{\theta}) = \begin{pmatrix} \theta \\ F_1(\theta) + \theta' \end{pmatrix}, \widetilde{F_2}(\tilde{x}) = \begin{pmatrix} x' \\ F_2(x) \end{pmatrix}. \tag{9}$$

Then we theoretically proved that the generalized smoothed classifier is certifiably robust within a given range. For simplicity, we choose two Gaussian distribution to sample transformation parameters, i.e. $\theta \sim \mathcal{N}(0, \sigma_1^2 I), \theta' \sim \mathcal{N}(0, \sigma_2^2 I)$. Before stating our main theorem, we introduce several notations $\tilde{z}_\xi = \tilde{F_1}(\tilde{\xi}) + \tilde{F_2}(\tilde{x}), \tilde{z}_\theta = \widetilde{F_1}(\tilde{\theta}) + \widetilde{F_2}(\tilde{x})$, $\tilde{y}_\xi = (y'_\xi, y_\xi)^T = \tilde{\sigma}(\widetilde{F_1}(\tilde{\xi}) + \tilde{F_2}(\tilde{x}))$ and $\tilde{y}_\theta = (y'_\theta, y_\theta)^T = \tilde{\sigma}(\widetilde{F_1}(\tilde{\theta}) + \tilde{F_2}(\tilde{x}))$.

Theorem 2 *Suppose $f(x)$ is any classifier and $\tilde{G}(\tilde{x})$ is the smoothed classifier defined in Eq.(7), if there exists $\underline{p_A}$, $\overline{p_B}$ that satisfies that*

$$\tilde{G}(\tilde{x})_A \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant \tilde{G}(\tilde{x})_B,$$

then $y_A = \operatorname{argmax}_{i\in\mathcal{Y}} \tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x}))_i$ for any $\|\xi\| \leqslant R$ where

$$R = \frac{1}{2M^*}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{10}$$

and the coefficient M^* is defined as

$$M^* = \max_{\xi,\theta} \sqrt{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\left\|\frac{\partial F_2(y_\xi)}{\partial \xi} - \frac{\partial F_1(\theta)}{\partial \theta}\right\|_2^2}. \tag{11}$$

This theorem is the main result of our generalized randomized smoothing. We have several observations about it. First, we see that the certified radius is similar to the result in theorem 2. Second, compared with resolvable transformations, we need to add a new type of noise when constructing the generalized smoothed classifier. This isotropic noise has the same dimension as the data and is added to the intermediate layers of surrogate neural networks. Third, we observe that the coefficient M^* depends on the norm of the difference of two Jacobian matrices, later we will discuss the meaning of this term in detail.

Before diving into our theoretical insight and proof of the theorem, we introduce a specific case of theorem2? which is more convenient for practical usage. We empirically found that taking a linear transformation as $F_1(\theta) = A_1\theta + b_1$ where $A_1 \in \mathbb{R}^{n\times m}$, $b_1 \in \mathbb{R}^n$ does not sacrifice the precision of the surrogate network. Then we have $\frac{\partial F_1(\theta)}{\partial \theta} = A_1$. After substituting the term in Eq.(11) we need to optimize ξ for calculating M^* and we make the bound tighter. Formally, we have the following corollary,

Corollary 1 *Suppose $f(x)$ is any classifier and $\tilde{G}(\tilde{x})$ is the smoothed classifier defined in Eq.(7), if the layers $F_1(\theta)$ in the surrogate neural network has the following form,*

$$F_1(\theta) = A_1\theta + b_1 \tag{12}$$

where $A_1 \in \mathbb{R}^{n\times m}$, $b_1 \in \mathbb{R}^n$ are the parameters. And if there exists $\underline{p_A}$, $\overline{p_B}$ that satisfies that

$$\tilde{G}(\tilde{x})_A \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant \tilde{G}(\tilde{x})_B,$$

then $y_A = \operatorname{argmax}_{i\in\mathcal{Y}} \tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x}))_i$ for any $\|\xi\| \leqslant R$ where

$$R = \frac{1}{2M^*}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{13}$$

and the coefficient M^* is defined as,

$$M^* = \max_{\xi} \sqrt{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\left\|\frac{\partial F_{22}(y_\xi)}{\partial \xi} - A_1\right\|_2^2} \tag{14}$$

4 Theoretical Analysis and Proof Sketch

In this section we briefly summarize the main idea for proving the theorem2 and the theoretical insight of our generalized randomized smoothing. The key idea is to prove the gradient of the smoothed classifier can be bounded by the a function of the classification confidence and the parameters of noise distribution. Formally, we calculate the gradient to the perturbation paramter ξ for our augmented smoothed classifier,

$$\nabla_{\tilde{\xi}}\tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x})) = \nabla_{\tilde{\xi}}\mathbb{E}_{\tilde{\theta}\sim\tilde{g}(\tilde{\theta})}[\tilde{f}(\tilde{\tau}(\tilde{\theta}, \tilde{\tau}(\tilde{\xi}, \tilde{x})))]. \tag{15}$$

We expand the expectation into integral and see that

$$\nabla_{\tilde{\xi}}\tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x})) = \int_{\mathbb{R}^{n+d}} \frac{\partial \tilde{F}(\tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi))}{\partial \tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi)} \cdot \frac{\partial \tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi)}{\partial \tilde{y}_\xi} \cdot \frac{\partial \tilde{y}_\xi}{\partial \tilde{\xi}} \tilde{g}(\tilde{\theta})\mathrm{d}\tilde{\theta}. \tag{16}$$

The key step is to eliminate the gradient of $\frac{\partial \tilde{F}(\tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi))}{\partial \tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi)}$ and replace it with $\frac{\partial \tilde{F}(\tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi))}{\partial \tilde{\theta}}$. Then we integral it by parts to get the following obejective,

$$\nabla_{\tilde{\xi}}\tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x})) = -\int_{\mathbb{R}^{n+d}} \tilde{F}(\tilde{\tau}(\tilde{\theta}, \tilde{y}_\xi))\frac{\partial}{\partial \tilde{\theta}}(\tilde{M}(\tilde{\xi}, \tilde{\theta})\tilde{g}(\tilde{\theta}))\mathrm{d}\tilde{\theta}. \tag{17}$$

After that, we could bound the gradient of the generalized smoothed classifier using the technique similar to randomized smoothing. More details of the proof can be found in Appendix?

Next We provide the theoretical insight for our augmentation scheme on transformation parameters and data. First, the key is to expand the transformation space by adding additional dimensions to form an closed space. In the augmented space, the Jacobian matrix of the semantic transformation became invertible which is crucial for our proof. Second, as we can see in Eq? that M^* is influenced by two factors. One is the standard deviation of two noise distribution. The other is the norm of the maximum matrix $\frac{\partial F_{22}(y_\xi)}{\partial \xi} - A_1$. In fact, it can be viewed as the residual of the non-resovable part of the transformation. Along this line, our method decompose the unknown semantic transformation into a resovable part and a residual part. The non-resovable residual part could be handled by introducing additional noise with standard deviation σ_2.

5 Experiments

5.1 Experimental Setup.

In this section conduct extensive experiments to show the effectiveness of our GSmooth on various types of semantic transformations. We use MNIST, Cifar10 and Cifar100 datasets? to verify our

	Type	Dataset	Attack Range	Certified Accuracy(%)						
				GSmooth(Ours)	TSS	DeepG	Interval	VeriVis	Semanify-NN	IndivSPT/distSPT
Gaussian Blur	Additive	MNIST	$\lvert\alpha\rvert^2 < 16$	67.4	90.6	—	—	—	—	—
		Cifar10			63.6	—	—	—	—	—
		Cifar100								
Translation	Additive	MNIST	$\lvert\alpha\rvert < 20$	82.2	99.6	—	—	98.8	98.8	—
		Cifar10			80.8	—	—	65.0	65.0	—
		Cifar100								
Brightness,Contrast	Resolvable	MNIST	$\lvert\alpha\rvert < 0.4$	82.5	97.6	$\leq$0.4	0.0	—	$\leq$74	—
		Cifar10			82.4	0.0	0.0	—	—	—
		Cifar100								
Rotation	Non-resolvable	MNIST		64.6	97.4	$\leq$85.8	$\leq$6.0	—	$\leq$92.48	$\leq$76
		Cifar10			70.6	62.5	20.2	—	$\leq$49.37	—
		Cifar100								
Scaling	Non-resolvable	MNIST	$\lvert\alpha\rvert < 0.3$		97.2	85.0	16.4	—	—	—
		Cifar10			58.8	0.0	0.0	—	—	—
		Cifar100								
rotational Blur	Non-resolvable	MNIST	$\lvert\alpha\rvert < 10$	39.7	—	—	—	—	—	—
		Cifar10			—	—	—	—	—	—
		Cifar100		16.6						
Defocus Blur	Non-resolvable	MNIST	$\lvert\alpha\rvert < 5$	>25.0	—	—	—	—	—	—
		Cifar10			—	—	—	—	—	
		Cifar100								
Motion Blur	Non-resolvable	MNIST	$\lvert\alpha\rvert < 5$		—	—	—	—	—	—
		Cifar10			—	—	—	—	—	—
		Cifar100								
Zoom Blur	Non-resolvable	MNIST	$\lvert\alpha\rvert < 0.5$	44.6	—	—	—	—	—	—
		Cifar10			—	—	—	—	—	—
		Cifar100								
pixelate	Non-resolvable	MNIST	$\lvert\alpha\rvert < 0.5$	> 45.3	—	—	—	—	—	—
		Cifar10			—	—	—	—	—	—
		Cifar100								

methods. We train a resnet with 110 layersHe et al. from scratch. Similar to prior works, we apply moderate data augmentationCohen et al. to improve the generalization of the classifier. The training process of classifiers and certification for semantic transformations are done on 2080Ti GPUs. For the surrogate image-to-image translation model for simulating semantic transformations, we adopt the U-Net architecture Ronneberger et al. for $\widetilde{\sigma}(\cdot)$ layers and several simple convolutional or linear layers for $\widetilde{F_1}(\cdot)$ and $\widetilde{F_2}(\cdot)$. All models are trained using Adam optimizer with initial learning rate 0.001 that decays every 50 epochs until convergence.

5.2 Experimental Results

5.2.1 Main Results

To demonstrate the effectiveness of our GSmooth on certifying complex semantic transformations, we measure the certified correct accuracy for different semantic transformations on different datasets in Table **??**. We compare the results of our GSmooth with several baselines. **TSS**(Li et al.) and **IndivSPT/distSPT** (Fischer et al.) and methods based on randomized smoothing for some simple semantic transformation. Our GSmooth is a natural and powerful extension of their methods. Other baselines are deterministic certification approaches. **DeepG**(Balunović et al.) uses linear relaxations similar to Wong & Kolter. **Interval**(Singh et al.) is based on interval bound propagation. **VeriVis**(Pei et al.) enumerates all possible outcomes for semantic transformations with finite values of parameters. **Semanify-NN**(Mohapatra et al.) proposed an new preprocessing layer to turn the problem into a l_p norm certification.

We have the following observations for the experimental results. First, only our method achieves non-zero accuracy on certifying some complex semantic transformations. This is a breakthrough that greatly extends the boundary of randomized smoothing based methods. Second, we see the performance of GSmooth is similar to the state-of-the-art randomized smoothing approaches like TSS on several simple semantic transformations like Gaussian blur, translation. This is a natural result since our method works similar for resovable transformations. For some simple non-resovable transformations like rotation, our accuracy is slightly lower but still comparable. The possible reason is that we uniformly bounded the norm of the non-resolvable residual part of the semantic transformation. In TSS(Li et al.) they derive more elaborate Lipschitz bound for rotation which is better than us. Third, xxx.

5.2.2 ABLATION STUDY

Ablation Study on the influence of noise distribution. The choice of noise distribution is an important for randomized smoothing based methods. Since our GSmooth could certify different types of semantic transformations that exhibit different properties. Understanding the influence of different smoothing distributions for different semantic transformations is necessary. We choose (folded) Gaussian, uniform and exponential distribution and compare the certified accuracy on XXX?. As shown in Fig.(?), We found that XXX?.

Ablation Study on the influence of noise variance for traning?certificaton.

Since our GSmooth contains two different variance for controlling the resolvable part and the residual part for a non-resolvable semantic transformation. Intuitively we see that higher the noise variance decrease the coefficient M^* but it might also degrade the clean accuracy. Here we investigate the effect of different noise variance on the certified accuracy. The results are shown in XXX?.

6 CONCLUSIONS

7 ACKNOWLEDGMENTS

REFERENCES

Mislav Balunović, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. *Advances in Neural Information Processing Systems 32*, 2019. 2.1, 5.2.1

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, 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*, pp. 387–402, 2013. 1

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1, 2.1

Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. *arXiv preprint arXiv:2104.01086*, 2021. 1, 2.1

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019. 1, 2.2, 5.1

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019. 2.1

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. 2018. 2.1

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pp. 1802–1811. PMLR, 2019. 2.1

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. *arXiv preprint arXiv:2002.12463*, 2020. 1, 2.2, 5.2.1

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018. 1

Jamie Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 786–787, 2020. 2.2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1, 5.1

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2.1

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 2.1

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 2.1

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019. 2.2

Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification, 2021. 1, 2.2, 3.3, 5.2.1

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. 1, 2.1

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–252, 2020. 1, 2.1, 3.3, 5.2.1

Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Towards practical verification of machine learning: The case of computer vision systems. *arXiv preprint arXiv:1712.01785*, 2017. 5.2.1

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015. 5.1

Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*, 2019. 2.2

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019. 2.1, 5.2.1

Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018. 2.1

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. 1

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020. 1

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018. 1, 5.2.1

Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018. 1

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 2.1

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020. 2.2

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019a. 1

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019b. 1

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017. 3.3

A PROOF OF THEOREMS

In this section, we will provide detailed proof of theorems in our paper.