UNDERSTANDING CERTIFICATION UNDER DISTRIBUTIONAL SHIFT

Andy Lapastora, Kevin Tran, Carina Zhang
Department of Computer Science
Stanford University
Stanford, CA, 94305 USA
{awlapas,ktran23,carinaz}@stanford.edu

ABSTRACT

Certified models have shown great promise as an adversarial defense because they are provably robust against attacks of a particular family. However, despite certified models' impressive performance when they are exposed to adversarial data, it remains unclear whether they are still robust when exposed to distributional shifts. In this work, we expose certified robust models from current literature to distributional shifts on their original training datasets. Despite being proven to be robust against L_p-bounded attacks, these certified models see significant decreases in certified test accuracy when tested on data from a shifted distribution. We analyzed specific observations from our experiment results in an attempt to understand why these models failed.

1 INTRODUCTION

Current state-of-the-art classifiers can fail catastrophically when exposed to small adversarial perturbations that are imperceptible to humans (Cohen et al., 2019). Researchers have begun to develop certified defenses, sometimes referred to as verified defenses, which are defenses that are provably robust against all attacks of a particular family (such as L_p bounded attacks) (Raghunathan et al., 2018). A rigorous definition of certification will be introduced in Section 3 of this paper.

These certified defenses have seen positive results in defending against specific attacks, however, it remains unclear whether these results will hold under distributional shift, when the test data is sampled from a different distribution than that of the training data (Cohen et al., 2019; Raghunathan et al., 2018; Moosavi-Dezfooli et al., 2017).

We expose three defense models that have been proven certified by different techniques to data distributions that they have not seen in training to test their ability to correctly classify data outside of their respective training sets. By capturing how these models perform in unknown situations, we can better evaluate a model's robustness. In the context of image classification, distributional shift in our work refers to both common image corruptions and changes in categorical distribution. Image corruptions are usually done by adding noise, while changes in categorical distribution requires the use of a new test set that matches the categories of the training set (Recht et al., 2018; Mu & Gilmer, 2019).

We tested three existing certified defense techniques - Randomized Smoothing, CROWN-IBP, and Dual Convex Relaxation (Cohen et al., 2019; Wong et al., 2018; Weng et al., 2018). All three defense schemes have been proven robust under L_p-bounded attacks. First, we used pre-trained models and tested the models on the original test datasets to verify the original results of the authors. Then, we applied attacks to the original test sets to see how the models perform. Finally, we tested each of the models on a shifted version of the original dataset. To evaluate the performance of these models, we used certified test accuracy. We compared our results with the results without distributional shift to determine whether these verification models remain generalizable under distributional shift. However, for these three models, we observe that significant drops in certified test accuracy occurred when tested with distributional shift.

Understanding whether certified defenses are robust under distributional shifts is essential in high-stake application scenarios, such as security verification and autonomous driving, where the failure of classification causes severe consequences. At the same time, there are many applications of image classification where complete consistency between the training data distribution and the test data distribution is nearly impossible. As an example, an autonomous vehicle trained in California should be able to drive just as well in Alaska. It is costly to train autonomous vehicle in every single climate and location; thus, if a robust model generalizes well among similar datasets, the cost of training will go down significantly.

2 Prior Work

Prior work has shown that it is important that neural networks are robust; a network should be able to correctly classify examples that are created specifically to fool it. Current work shows what defenses are currently available, what attacks have been formulated to overcome those defenses, what evaluation strategies are appropriate to define robustness, and what work has been done to expose robust models to shifted distributions.

Attacking and Defending a Neural Network. Attacks against Neural Networks were conceived of in order to test how well a given network performs against data it has not seen. Attacks are carried out via Adversarial Examples, or samples that are generated in order to fool the targeted classifier. Most often these samples are versions of the data that the model was trained on, just slightly perturbed in ways that make the classifier most likely to misclassify them (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2017; Schott et al., 2018). Adversarial examples generated in this way are called "white-box" attacks. Another type of attack, a "black-box" attack, is an attack where the attacker doesn't have access to the model beforehand (Papernot et al., 2017; 2016a). Both types of attacks have been found to be effective in finding vulnerabilities in networks.

As a response to adversarial attacks, many researchers have directed their focus onto developing defenses to improve the robustness of existing Neural Network models. Early research focused on designing defense mechanisms that would directly target the underlying vulnerabilities exposed by existing attacks, but as researchers developed defenses to address these concerns, attackers came up with even stronger attacks to circumvent these newly developed defenses, causing an "arms-race" (Carlini & Wagner, 2017).

Adversarial training has been widely implemented and proven to be effective for some specific corruptions in image detecting (Ford et al., 2019). For example, distillation and obfuscated gradients both demonstrated great potential in improving adversarial robustness in Neural Network models, but new attacks that exposed these two defenses were also developed shortly after these defense mechanisms themselves. (Papernot et al., 2016b; Athalye et al., 2018). Therefore, in recent years, the research community has been more focused on understanding the existence of adversarial examples in hopes of understanding what key properties factor into building an adversarial robust model. Recent research has shown that visual interpretability and larger model complexity are helpful to improving adversarial robustness, and work has been done to incorporate such properties into new defenses. (Schmidt et al., 2018; Madry et al., 2017; Anonymous, 2020). Researchers have also tried to train defense models with attacks that are unrelated to their models, but this approach has yet to yield impressive results (Kang et al., 2019).

Certification Certification (also referred to as verification) is the process of training a model whose prediction at a particular point doesn't change for some perturbation ϵ around that same point. This means that "a certain robustness certificate may guarantee that for a given example x, no perturbation δ with l_∞ norm less than some specified ϵ will change the class label that the network predicts for the perturbed example $x + \delta$" (Wong et al., 2018). Current research has made progress in certifying models against specific adversarial threat models when using test data from the same distribution as the training data. Figure 1 shows an example of a CIFAR-10 image perturbed with different levels of L_2 noise determined by ϵ (Cohen et al., 2019). The model provided by Cohen et al. (2019), for example, is certifiably robust against attacks of this nature up to a certain ϵ. This means that the prediction of a model will not change up to a certain epsilon, meaning that the classifier can guarantee that one of the perturbed images of the panda shown will still be classified correctly as a panda. While this work seems promising, little work has been done to see how these models perform under distributional shift.

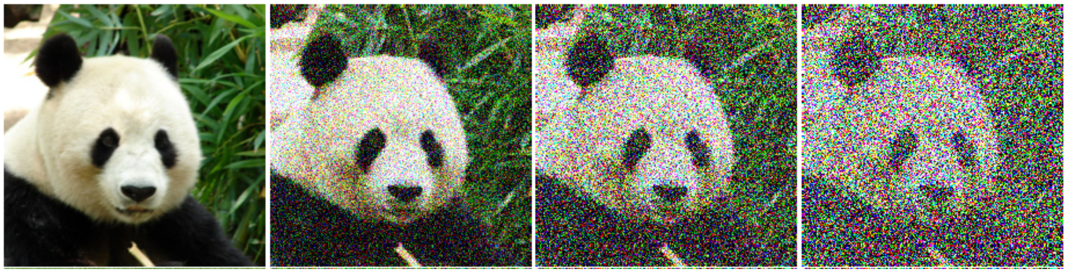

Figure 1: ImageNet image additively corrupted by varying levels of Gaussian noise with $\epsilon = 0.00, 0.25, 0.50, 1.00$ from left to right (Cohen et al., 2019).

Distributional Shift and Evaluating Robustness. Further, it has been proven that ”robust” models trained using adversarial training are sensitive to the input data distribution (Ding et al., 2019). This means that a robust model exposed to data from a new distribution (distributional shift) will not perform as well as expected. Little research has been done in this area to date. Adversarial robustness is difficult to measure, and it has been shown that there are some trade-offs with standard accuracy when considering robust accuracy (Tsipras et al., 2018). There is not one standard metric to measure robust accuracy in existing literature, but recent work has yielded new, complementary metrics that can measure the robustness of a model against unforeseen attacks(Kang et al., 2019; Carlini et al., 2019).

Our work seeks to explore shortcomings in the last two sections. We explore the effect of distributional shift on three certified robust models using certified test accuracy as a metric. In this way, we can understand how robust these models actually are. As will be discussed in depth later, it is important for real-world applications that robust models be able to handle data they haven’t seen previously.

3 Notations and Definitions

We define the necessary concepts used throughout our paper in this section, namely certification, distributional shift, and certified test accuracy.

3.1 Certification

Certification provides a guarantee that a classifier’s prediction will not change after perturbation on input data. In other words, it provides a radius within which a smoothed classifier produces the same prediction as its base classifier (Cohen et al., 2019). For a test sample (x, c), where c is the true label of x, the smoothed classifier will produce a prediction defined as $g(x+\delta) = \hat{c}_a$. We call a prediction $\hat{c}_a$ *correct* if $\hat{c}_a = c$ and *certified* if $\hat{c}_a = g(x)$, or the certified radius $r > \epsilon$. In our experiments, we investigate how the distribution of results shifted after the test set becomes a shifted test set.

3.2 Distributional shift

Denote the input data as x and the output data as y. A distributional shift refers to the change from training to testing in the joint probability distribution of (x, y). Since the exact distribution of test data is unknown, it is hard to precisely characterize a specific distributional shift. We will showcase one specific type of distributional shift used in our paper by an example. Figure 2 shows the images under the class ”airplane” from CIFAR-10 and CINIC-10 respectively. CINIC-10 contains the same 10 classes as CIFAR-10 does, but CINIC-10 is sampled from from both CIFAR and ImageNet, a much larger population than CIFAR-10 (Darlow et al., 2018). Since CIFAR-10 and CINIC-10 are drawn from two different populations collected at different periods of time, CINIC-10 is considered a shifted version of CIFAR-10.

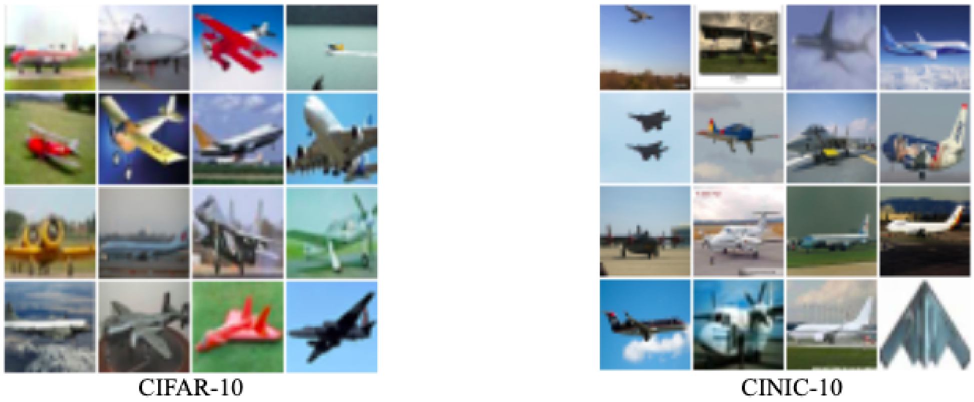

Figure 2: Images under the class ”airplane” from CIFAR-10 and CINIC-10 (Darlow et al., 2018).

Table 1: Summary of parameters

Certification	Training Set	Test Set	Classifier	Attacks
Randomized Smoothing	CIFAR-10	CINIC-10	Resnet110	$L_2, \epsilon \in (0, 1)$
CROWN-IBP	MNIST	MNIST-C	Robust Convnet	$L_\infty, \epsilon \in (0, 0.2)$
Dual Convex Relaxation	MNIST	MNIST-C	Robust Convnet	$L_\infty, \epsilon \in (0, 0.2)$
Dual Convex Relaxation	CIFAR-10	CINIC-10	Robust Convnet	$L_\infty, \epsilon \in (0, 0.1)$

3.3 Certified test accuracy

Our performance metric of choice is certified test accuracy. This is defined as the test set accuracy that the classifier can provably attain under any adversarial attack within the bounds that the verified model is defined as able to defend. For example, in the context of randomized smoothing, attacks are limited to L_2-bounded attacks. Certified test error, defined as the upper-bound on the possible test set error that the classifier can suffer under any norm-bounded attack, is another common metric used in the literature in this domain. However, we will use certified test accuracy as our evaluation metric in this paper for consistency (Wong et al., 2018).

Since prediction and certification are two independent processes, we want to understand the distribution of the results on top of certified test accuracy. A prediction is labeled as certified by the model when the model can guarantee that that prediction is within an upper bound on the amount of adversarial perturbation. If a prediction is correct and certified, this also means that the prediction didn’t change between exposure to the shifted data and the shifted data with attacks. In order to evaluate the performance of the models, we put each prediction the model makes into one of the four following categories:

1. correct and certified - C&C
2. correct and uncertified - C&U
3. incorrect and certified - I&C
4. incorrect and uncertified - I&U

4 Experimental Setup

We tested certified robust models on distributional shift. The three models we implemented use the certification methods of Randomized Smoothing, CROWN-IBP, and Dual Convex Relaxation respectively (Cohen et al., 2019; Zhang et al., 2019b; Wong et al., 2018). For each model, we ran the robust model on the original test set to reproduce the results obtained by the original authors.

Then, we tested each model on a shifted version of its dataset. We compared the predictions of the models on the original data to the predictions made on the shifted data. In this way, we were able to evaluate the models on how they performed on unseen data. After this, we applied attacks to the shifted data to further stress the models. The last two steps are the ones that we focus on, where the robust model is exposed to shifted data and then shifted data that has been attacked. Table-1 summarizes the parameters we used as a set-up of our experiments.

4.1 Randomized Smoothing

Model Description. Randomized smoothing is a provable adversarial defense in L_2. It is a technique that constructs a new, smoothed classifier g from an arbitrary base classifier f. When queried at x, the smoothed classifier g returns whichever class that the base classifier f has the largest probability to return when x is perturbed by Gaussian noise (Cohen et al., 2019).

Experimental Method. We used pre-trained randomized smoothing models with different configurations to run experiments on different levels of L_2 perturbations specified in Table-1. In the context of randomized smoothing, it is impossible to precisely evaluate the smoothed classifier, g, because the probability distribution over the classes when the input images are perturbed by Gaussian noise. (Cohen et al., 2019). Since the radius in which g is provably robust is not exact, we use Monte Carlo method as a sampling algorithms to approximate both the prediction, $g(x)$, and the certification L_2 radius. As described thoroughly in the original paper, the prediction procedure ”samples n inputs from the test data and returns $\hat{c}$, whose count is largest.” The certification procedure ”collects n samples of $f(x + \epsilon)$, count how many times $f(x + \epsilon) = c$, and uses a Binomial confidence interval to obtain a lower bound on $P(f(x + \epsilon) = c)$” (Cohen et al., 2019). While the original paper uses $n = 100000$, smaller values of n does not result high probability of abstaining from prediction. Since certification is time consuming, we choose $n = 1000$ as Monte Carlo sample size to speed things up.

Note that in the randomized smoothing model, we can only approximate the certified test accurate rather than compute it exactly. By using a Monte Carlo method, our final results might fluctuate depending on the sample size. However, the original paper's experiment results with different n values have shown that the noise introduced by Monte Carlo method when using smaller n does not have any significant influence on the direction of our final results.

4.2 CROWN-IBP

Model Description. For certification, this model combines Interval Bound Propagation (IBP) training and CROWN, a linear based relaxation method, to produce lower bounds of margins between ground-truth class and all classes (including the ground-truth class itself). The model will be certifiably robust for a particular input example if and only if all margins between the ground-truth class and other classes (except the ground truth class) are positive. The model is trained to be certifiably robust for perturbations up to a certain epsilon.

Experimental Method. We used pre-trained models provided by Zhang et al. (2019b) to run the experiments. These models were trained on l_∞-normed perturbations. The first model was trained on the MNIST dataset to be certifiably robust against l_∞-normed attacks that perturb MNIST images up to a radius of ϵ=0.3. The second model was trained on the CIFAR-10 dataset to be certifiably robust against l_∞-normed attacks that perturb CINIC-10 images up to a radius of ϵ=0.007.

First, we tested each of these models on test data from the distribution they were trained on. Then, we tested each model on a shifted distribution. The model trained on MNIST was tested on MNIST-C and the model trained on CIFAR-10 was tested on CINIC-10. Finally, after testing the models on their appropriate shifted distributions, we applied Projected Gradient Descent(PGD) attacks on the test data from the original distribution and the test data from the shifted distribution and tested our models against this perturbed data. For the MNIST and MNIST-C datasets, PGD attacks of ϵ=0.3 and ϵ=0.4 were applied. For the CIFAR-10 and CINIC-10 datasets, PGD attacks of ϵ=0.007 and 0.03 were applied. These PGD attack parameters were suggested by Zhang et al. (2019b) in their paper.

Table 2: Randomized Smoothing Model Trained on CIFAR-10

(a) CIFAR-10 Test Set

Attack	C&C	C&U	I&C	I&U
None	**73.86**	0.00	26.14	0.00
L_2 $\epsilon = .12$	69.60	10.40	6.20	13.80
L_2 $\epsilon = .25$	56.80	16.00	5.60	21.60
L_2 $\epsilon = .50$	36.00	24.00	4.40	35.60
L_2 $\epsilon = 1.00$	20.00	23.20	4.00	52.80

(b) CINIC-10 Test Set

Attack	C&C	C&U	I&C	I&U
None	**35.00**	0.00	65.00	0.00
L_2 $\epsilon = .12$	34.09	3.18	49.09	13.63
L_2 $\epsilon = .25$	12.73	3.64	66.82	16.82
L_2 $\epsilon = .50$	11.82	4.55	57.73	25.91
L_2 $\epsilon = 1.00$	16.36	4.09	49.55	30.00

4.3 Dual Convex Relaxation

Model description. This model defines a convex outer bound on the adversarial polytype of a model. In other words, this is the set of all new images that can be generated after applying a certain amount of norm-bounded perturbation to the original input data (Wong & Kolter, 2017).

Experimental method. We used four pre-trained robust scaled models provided by Wong et al. (2018) to run experiments. There are two sets of two models, a large and a small model trained on l_∞-normed perturbations. The small network has two convolutional layers of 16 and 32 filters with a fully connected layer of 100 units (Wong & Kolter, 2017). The large network is a larger version of the small, with four convolutional layers with 32, 32, 64, and 64 filters, and two fully connected layers of 512 units (Wong et al., 2018). The first set of models was trained on MNIST, and the second was trained on CIFAR-10 (LeCun & Cortes, 2010; Krizhevsky et al.).

First, we tested the models on their respective original test sets. For MNIST, this is set of 6000 samples, and for CIFAR-10 this is a set of 1000 samples. Then we test each model on a shifted distribution: MNIST-C for the model trained on MNIST and CINIC-10 for the model trained on CIFAR-10 (Mu & Gilmer, 2019; Darlow et al., 2018). MNIST-C is a corrupted version of the MNIST dataset that was formulated to test out-of-distribution robustness for computer vision models and consists of 16000 samples. Since the original MNIST dataset is only 10000 samples, we randomly select 1000 samples from MNIST-C to test on. After testing the model on the shifted data, we applied attacks to both the original and shifted data. Wong et al. (2018) provide both Projected Gradient Descent (PGD) and Fast Gradient Sign (FGS) attacks that we applied to the datasets with different values for ϵ, which is the parameter controlling the amount of perturbation the attack applies to the data. We used a PGD attack with $\epsilon = .2$ for the MNIST models and an FGS attack with $\epsilon = .1$ for the CIFAR-10 models.

5 Results on distributional shift

We report both the certified test accuracy and the certified vs. correct results matrix for each model. Our results show that each model's certified test accuracy significantly decreases after being exposed to a distributional shift. Figure 3 shows that the number of correct and certified predictions made by the classifiers dropped between 5% (CROWN-IBP) and 38% (Randomized Smoothing). Full results including adding adversarial attacks and models trained on other datasets are shown in each following section. These results follow a similar pattern. For context, since CIFAR-10 contains ten classes with equally weighted numbers of images in each class, 10% certified test accuracy is as good as random guessing.

5.1 L_2 certified results

Randomized Smoothing shows the most significant drop in certified test accuracy. From the results in Table 2, we can see that when ϵ stays the same, the percentage of correct and certified predictions made by the classifier drops 10% - 30%, which is a significant drop in performance. From the confusion matrix, we can see that the percentage of uncertified predictions increases as ϵ increases.

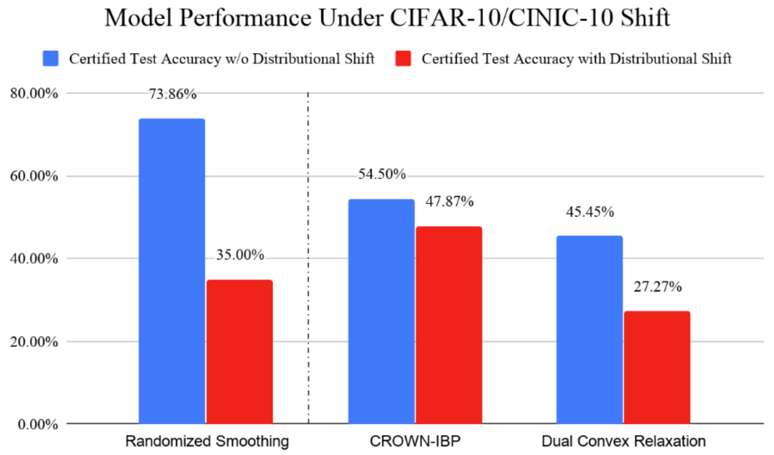

Figure 3: Certified test accuracy drops significantly when certified robust models are tested on distributional shift. Note that Randomized Smoothing is designed to defend against L_2 attacks while CROWN-IBP and Dual Convex Relaxation are designed to defend against L_{inf} attacks.

Table 3: Robust CROWN-IBP Model Trained on MNIST

(a) MNIST Test Set

Attack	C&C	C&U	I&C	I&U
None	**93.32**	4.88	0.00	1.80
PGD $\epsilon = .30$	93.32	3.31	0.00	3.37
PGD $\epsilon = .40$	87.54	8.22	0.00	4.24

(b) MNIST-C Test Set

Attack	C&C	C&U	I&C	I&U
None	**68.50**	0.00	31.50	0.00
PGD $\epsilon = .30$	66.60	0.00	33.40	0.00
PGD $\epsilon = .40$	66.07	0.00	33.93	0.00

When tested on CINIC-10, Randomized Smoothing's performance has a sharp drop to 12.73% when ϵ increases from 0.12 to 0.25. This means that when $\epsilon = 0.25$, the predictions made by the model is around the same level of random guessing. Another interesting observation is that the percentage of certified predictions increases when randomized smoothing is tested on CINIC-10. The percentage of predictions in the incorrect and certified category is extremely high when randomized smoothing is exposed to CINIC-10.

5.2 L_∞ CERTIFIED RESULTS

5.2.1 CROWN-IBP

The results of the experiments for CROWN-IBP models are reflected in Tables 3 and 4.

Results for MNIST and MNIST-C. Applying PGD attacks to MNIST and MNIST-C datasets generally lowered the certified test accuracy of the CROWN-IBP model trained on MNIST (except for the attack that perturbed the MNIST dataset by $\epsilon = 0.3$, but that attack nonetheless lowered the

Table 4: Robust CROWN-IBP Model Trained on CIFAR-10

(a) CIFAR-10 Test Set

Attack	C&C	C&U	I&C	I&U
None	**54.50**	16.32	0.00	29.18
PGD $\epsilon = .007$	54.50	15.18	0.00	30.32
PGD $\epsilon = .03$	00.08	65.83	0.00	34.09

(b) CINIC-10 Test Set

Attack	C&C	C&U	I&C	I&U
None	**47.81**	11.25	0.00	40.94
PGD $\epsilon = .007$	47.81	11.25	0.00	40.94
PGD $\epsilon = .03$	0.94	55.00	0.00	44.06

Table 5: Results matrix for Robust DCR Model Trained on MNIST, tested on MNIST and distributional shift (MNIST-C), and results for PGD attack on both test sets.

(a) MNIST Test Set

Model	C&C	C&U	I&C	I&U
Small	**97.27**	2.73	0.00	0.00
Large	**94.55**	3.64	1.82	0.00

(b) MNIST-C Test Set

Model	C&C	C&U	I&C	I&U
Small	**43.64**	0.00	56.36	0.00
Large	**32.73**	0.00	67.27	0.00

(c) MNIST Test Set, PGD $\epsilon = 0.2$

Model	C&C	C&U	I&C	I&U
Small	12.73	0.00	84.55	2.73
Large	11.82	0.91	83.64	3.64

(d) MNIST-C Test Set, PGD $\epsilon = 0.2$

Model	C&C	C&U	I&C	I&U
Small	43.64	0.00	56.36	0.00
Large	31.82	0.00	68.18	0.00

number of correct and uncertified predictions). However, the decrease in certified accuracy is not enormous, ranging from between 2.43% to 5.78%.

In addition, the certified accuracy when tested on MNIST-C (68.50%) is significantly lower than when testing on MNIST (93.32%). The model made no incorrect and certified predictions on the MNIST dataset but made between 31.50% to 33.93% incorrect and certified predictions on the MNIST-C dataset. Furthermore, when testing on the MNIST-C dataset, the model did not make any incorrect and uncertified predictions. All of its incorrect predictions were certified. These observations suggest that the CROWN-IBP model is not able to withstand data from a different distribution than the data it was trained on.

Results for CIFAR-10 and CINIC-10. Applying $\epsilon = 0.007$ PGD attacks to CIFAR-10 and CINIC-10 datasets did not change the certified test accuracy of the CROWN-IBP model trained on the CIFAR-10 dataset. However, there is a significant decrease in certified test accuracy when we increase the PGD attack radius to $\epsilon = 0.03$. The model had a certified accuracy of 54.50% when testing on the CIFAR-10 dataset with an applied PGD attack of $\epsilon = 0.007$, but had only a certified test accuracy of 0.08% when testing on the CIFAR-10 dataset with an applied PGD attack of $\epsilon = 0.03$. However, the number of correct and uncertified predictions significantly increased, suggesting that some of the model's correct and certified predictions for the CIFAR-10 dataset with applied PGD attack of $\epsilon = 0.007$ became correct and uncertified predictions for the CIFAR-10 dataset with applied PGD attack of $\epsilon = 0.03$. With this increase in correct and uncertified predictions for the CIFAR-10 dataset, the total number of correct predictions (correct and certified predictions + correct and uncertified predictions) made by the model across no attack, PGD attack of $\epsilon = 0.007$ and PGD attack of $\epsilon = 0.03$ remain similar (although slightly lower for PGD attack of $\epsilon = 0.03$).

In contrast to the CROWN-IBP model trained on the MNIST dataset, the CROWN-IBP model trained on the CIFAR-10 dataset doesn't make any incorrect and certified predictions. All of its incorrect predictions are uncertified. However, this model also experiences decreases in certified accuracy when testing on its shifted dataset, CINIC-10 (from 54.50% to 47.81%), suggesting that this model also is unable to withstand a distributional shift.

5.2.2 Dual Convex Relaxation

The results of the experiments for the model using Dual Convex Relaxation are reflected in Tables Table 5 and Table 6.

Results for MNIST and MNIST-C. For the models trained on MNIST, we see a significant drop in certified test accuracy when testing on MNIST-C, from 97.27 to 43.64 for the small model and from 94.55 to 32.73 for the large model. For context, Wong et al. (2018)'s best reported test accuracy is 96.33 for MNIST on the small model. However, for a PGD attack with $\epsilon = 0.2$ on both datasets, while the certified test accuracy drops even more significantly for MNIST, the numbers remain similar for MNIST-C. This may suggest that attacks applied to MNIST-C don't affect the model's performance on that dataset, since the distribution is already shifted. We also see that predictions for both MNIST-C and MNIST-C under an attack are 100% certified. However, we see that under an attack on MNIST, the model performs just slightly better than random. Additionally,

Table 6: Results matrix for Robust DCR Model Trained on CIFAR-10, tested on CIFAR-10 and and distributional shift (CINIC-10), and FGS attack results for both test sets.

(a) CIFAR-10 Test Set

Model	C&C	C&U	I&C	I&U
Small	**45.45**	2.27	47.73	4.55
Large	**72.73**	18.18	9.09	0.00

(b) CINIC-10 Test Set

Model	C&C	C&U	I&C	I&U
Small	**27.27**	0.00	45.45	27.27
Large	**54.55**	18.18	0.00	27.27

(c) CIFAR-10 Test Set, FGS $\epsilon = 0.1$

Model	C&C	C&U	I&C	I&U
Small	36.36	0.00	56.82	6.82
Large	45.45	0.00	36.36	18.18

(d) CINIC-10 Test Set, FGS $\epsilon = 0.1$

Model	C&C	C&U	I&C	I&U
Small	18.18	0.00	54.54	27.27
Large	36.36	0.00	45.45	18.18

the uncertified correct predictions made on MNIST became incorrect predictions but remained uncertified on MNIST-C.

Results for CIFAR-10 and CINIC-10. Similarly, as seen in Table 6, we see a substantial drop in certified test accuracy when testing on CIFAR-10 as opposed to CINIC-10 for the models trained on CIFAR-10. Wong et al. (2018) were able to improve upon prior state-of-the-art on CIFAR-10 by roughly 10%, so the decrease from 72.73 to 54.55 in certified test accuracy when the large model was exposed to distributional shift is significant. Interestingly, for the large model, we see a relatively large number for predictions that are correct and uncertified (18.18), which remains the same on the CINIC-10 test set. Additionally, we see that all incorrect predictions are also uncertified on the CINIC-10 test set. When a Fast Gradient Sign attack with $\epsilon = 0.1$ is applied to CIFAR-10, we see that certified test accuracy drops and correct and uncertified predictions become incorrect and uncertified. The same phenomenon holds for the same attack applied to CINIC-10.

6 Discussion

We attempt to understand the unexpected phenomena and patterns we observed in our experiments. Some of the patterns, such as sub-class clustering, exist across models, and the others are model-specific observations.

Correct predictions cluster within certain classes. When the three certification methods are tested on CINIC-10, correct predictions seem to cluster around specific classes. For the model using Randomized Smoothing, the correct predictions cluster within the airplane, automobile, and bird classes. For the model using CROWN-IBP, correct predictions cluster around the airplane, automobile and cat classes. For the model using Dual Convex Relation, correct predictions solely cluster around the airplane class. Since we scaled down the CINIC-10 dataset by random sampling in order to speed up the experiment run time, we introduced some randomness during the sampling procedure. However, since the CINIC-10 test set is still evenly distributed amongst all ten classes, sampling is unlikely to be cause of this observation.

Pairs of labels frequently mistaken for one another. When the model using CROWN-IBP trained on MNIST is tested on MNIST-C, particular pairs labels are often mistaken for one another. Specifically, the label 4 was often mistaken to be 9 and the label 0 was mistaken to be 8. This cause of this behavior might be that 4 is similar to 9 while 0 is similar to 8. This observation is harder to decipher in other datasets. When CROWN-IBP trained on CIFAR-10 is tested on CINIC-10, images from the ship class are often misclassified as the airplane class. Similarly, when Randomized Smoothing trained on CIFAR-10 is tested on CINIC-10, images from the frog class and the truck class are often mistaken as automobiles. This misclassification may be attributed to the fact that the backgrounds of airplane and ship images are often both blue. Furthermore, the colors of ships and planes may often be quite similar.

Shifted test set produces more certified predictions for CROWN-IBP. When the model using CROWN-IBP trained on MNIST is tested on MNIST-C, the model makes only uncertified predictions, with between 31.50% to 33.93% of these uncertified predictions being incorrect. In contrast, when tested on the default MNIST dataset, the model makes relatively few uncertified predictions and didn't make any incorrect and uncertified predictions. The observation that the model certifies all examples in the MNIST-C dataset and that around 30% of these predictinos are wrong suggests that that model is vulnerable to distributional shift.

Shifted test set produces more uncertified predictions for Dual Convex Relaxation. The number of uncertified examples increased in the model using Dual Convex Relaxation when testing on CINIC-10. For the large model, we see a drop from 81.80% certified to 54.55%. This may suggest that a distributional shift is enough of a change in the input data to put predictions outside of the guaranteed certification boundary of the model.

Not as dramatic a decrease in certified test accuracy for CROWN-IBP. The model using CROWN-IBP trained on CIFAR-10 saw a drop in certified test accuracy from 54.50% to 47.81% when tested on CIFAR-10 and CINIC-10 respectively. While this drop is not as dramatic as the drop seen in other models, the resulting performance is comparable. The initial performance of 54.50% is already low and through further analysis, we saw that the model using CROWN-IBP predicts airplane, automobile and cat images particularly well (i.e. correct predictions for this model for both CINIC-10 and CIFAR-10 cluster around airplane, automobile, and cat images). Thus despite not having as steep of a certified accuracy drop as the other models, the model using CROWN-IBP trained on CIFAR-10 is still vulnerable to distributional shift.

7 Practical implications

Our paper shows that certified models do not perform well under distributional shift. We have seen that all three models that we explored can be vulnerable to distributional shifts. When exposed to shifted data, the certified test accuracy of three models decreases significantly. This vulnerability to distributional shift becomes problematic in safety-critical applications where a model that is expected to be provably safe may not actually perform as well as expected. Overall, our results suggest the need for further research in developing robust models that won't be vulnerable to distributional shifts.

Current literature in the field of machine learning suggests that there is an inherent trade-off between robustness and accuracy Zhang et al. (2019b). Increasing robustness requires more relaxation and thus may result in lower certified test accuracy for a particular set of perturbations. Since the three models studied in our paper are tuned to a specific fixed, known family of distortions, the generalizability of these models decreases. For example, Dual Convex Relaxation provides provable guarantees for adversarial robustness. However, this method ignores the performance of classifier on the non-adversarial examples (Zhang et al., 2019a).

The models we explored provide provable guarantees for various L_p bounded attacks, but they do not take into account other adversarial attacks such as distributional shift. This fact explains why all three models experienced a decrease in certified test accuracy when they were tested on shifted data. The authors of these models put strong emphasis on building certified models against a specific family of attacks. As a result, testing them against a different type of attack, such as shifted data, has a significant drop in performance.

8 Conclusion

Our work shows that the models using the three methods of certification that we examined, namely Randomized Smoothing, CROWN-IBP, and Dual Convex Relaxation, do not perform at a satisfactory level under distributional shift. All three models experienced a decrease in certified test accuracy when exposed to data from a different distribution. This suggests that the authors' claims of robustness make may not be completely well-founded, as the robustness certification guarantees do not take into account distributional shift. Distributional shift is necessary when testing how robust a model is because in real-life scenarios it is crucial that certified models are robust under all circumstances.

To conclude, we hope that our work provide a new perspective on robust certified models and promote further research in developing robust models that won't be vulnerable to distributional shifts.

REFERENCES

Anonymous. Visual interpretability alone helps adversarial robustness. In *Submitted to International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyes70EYDB. under review.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

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

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. *arXiv preprint arXiv:1902.08336*, 2, 2019.

Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

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

Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016*, 2019.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

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

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519. ACM, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2018.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks. *arXiv preprint arXiv:1804.09699*, 2018.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope, 2017.

Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019a.

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

A Code

Andy Lapastora, Dual Convex Relaxation: `https://colab.research.google.com/drive/1EvYAQOqFiCh9ys0E99jY-10jtW7M1Gxf`

Kevin Tran, CROWN-IBP: `https://github.com/kevtran23/CROWN-IBP` `https://colab.research.google.com/drive/1LqzMdZPaJcxxhsNTDoE14lodjKfmUL2z`

Carina Zhang, Randomized Smoothing: `https://colab.research.google.com/drive/1Zr34Wwuy0UfxTAafa_x3zd-hyyTt4jIx`

B Overleaf link

`https://www.overleaf.com/7319459322rychrdvsxmgz`