Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-features

Zhanyuan Zhang, Benson Yuan, Michael McCoyd, David Wagner
University of California, Berkeley
zhang_zhanyuan@berkeley.edu, yuanbenson@berkeley.edu, mmccoyd@cs.berkeley.edu, daw@cs.berkeley.edu

***Abstract*—Many works have demonstrated that neural networks are vulnerable to adversarial examples. We examine the adversarial sticker attack, where the attacker places a sticker somewhere on an image to induce it to be misclassified. We take a first step towards defending against such attacks using clipped BagNet, which bounds the influence that any limited-size sticker can have on the final classification. We evaluate our scheme on ImageNet and show that it provides strong security against targeted PGD attacks and gradient-free attacks, and yields certified security for a 95% of images against a targeted 20×20 pixel attack.**

***Index Terms*—adversarial examples, adversarial machine learning, adversarial robustness**

I. Introduction

Despite achieving superhuman performance on various computer vision, natural language processing, and game playing tasks, deep learning has been shown to be vulnerable to adversarial examples, which are inputs designed by adversaries to produce erroneous predictions by the model [1]. Previous work has mostly focused on a threat model where the attacker is allowed to perturb all pixels of the input, so long as the L_p norm of the perturbation does not exceed some threshold. However, it is not clear that this corresponds to any physically realizable attack [2].

In this paper we tackle adversarial sticker attacks [3], in which an attacker is restricted to placing a sticker somewhere in the image. These attacks correspond to a realistic threat: attackers could plausibly place a small sticker on a stop sign or hold it in a scene. Researchers have shown that these attacks fool existing classifiers, and can be mounted in the physical world despite viewpoint and brightness variation [4]. It is currently not known how to build classifiers that are robust against sticker attacks.

We develop a scheme, Clipped BagNet (CBN), that is robust by design against sticker attacks, and we show that it achieves close to state-of-the-art performance on ImageNet. We build on the BagNet-33 classifier, recently introduced by Brendel et al. [5]. BagNet shows that one can achieve approximately state-of-the-art accuracy on ImageNet with a classifier artificially restricted to a 33×33 receptive field. BagNet examines each 33×33 patch of the image separately. For each patch it makes a prediction for the class of the image based solely on that patch. It then aggregates these votes to make a final prediction. In Clipped BagNet, we modify the aggregation step to limit the influence of any one patch on the final prediction. This allows us to bound the effect of a sticker and provide certified security guarantees for many images.

We evaluate the security of CBN using both a white-box attack (PGD) and a black-box attack (SPSA). CBN achieves 83.6% top-5 clean accuracy on ImageNet (compare to ResNet-50: 93.4% top-5 accuracy). In our experiments, CBN achieves 65.2% robust top-5 accuracy against untargeted PGD attack with a 20×20 sticker (ResNet-50: 30.8% robust top-5 accuracy) and 99.8% robust top-1 accuracy against targeted PGD attack with a 20×20 sticker (ResNet-50: 53% robust top-1 accuracy). Due to the design of CBN, we can derive certified security results: 95.0% of images are certified secure against targeted attack with a 20×20 sticker, and 20.4% against untargeted attack. As far as we are aware, these are the strongest security against sticker attacks on an ImageNet-scale dataset to date.

II. Background

A. Adversarial Sticker Attack

The adversarial sticker attack allows the attacker to choose a region of their choice in the image and replace that portion of the image with any content of their choice. Brown et al. [3] introduced adversarial stickers by demonstrating a universal, robust, targeted physical adversarial sticker capable of fooling image classifiers when added to a real-world scene. Their work launched a line of research exploring these attacks. Researchers have demonstrated that a malicious sticker on a stop sign could cause it to be mis-classified as a 45 mph speed sign [4], printing a malicious image around the frames of eyeglasses or placing a malicious sticker on one's hat can fool face recognition [6], [7], holding a print-out of a malicious image can fool person detection by surveillance camera [8], and wearing an adversarial T-shirt can evade a real-time person detection system [9]. These highlight the need for effective safeguards against such attacks.

B. The BagNet Classifier

We build on the Bagnet classifier, which has been shown to achieve close to state-of-the-art accuracy on ImageNet [5]. BagNet breaks the image into overlapping patches (BagNet-9 uses 9×9 patches, BagNet-33 uses 33×33 patches). For each patch, it feeds the patch into a ResNet-based classifier that outputs a 1000-dimensional vector of logits, which we can think of as inducing a probability distribution on the 1000 ImageNet classes. We obtain a 1000-dimension vector of logits for each patch; BagNet then averages these vectors to obtain a global vector of logits, which are fed to softmax to obtain

DOI 10.1109/SPW50608.2020.00026

final class probabilities for the input image. Figure 1 provides a high-level overview of BagNet.

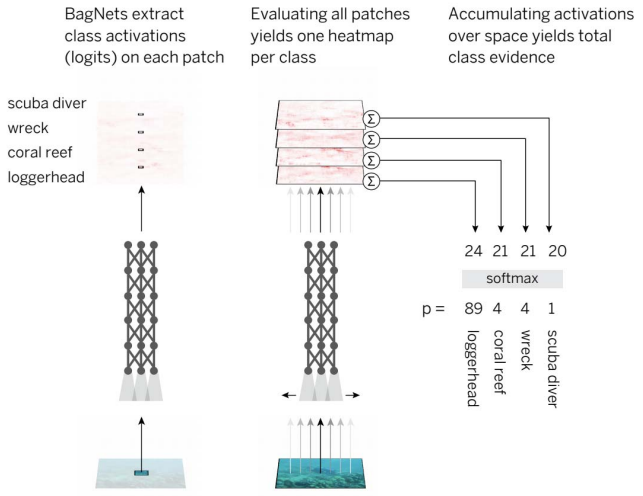

Fig. 1. BagNet extracts features from small image patches which are fed into a linear classifier yielding one logit heatmap per class. These heatmaps are spatially averaged and passed through a softmax layer to get the final class probabilities.

We can build a heatmap for each class, which represents the locations in the image that voted for that class. Each cell in the heatmap corresponds to a particular patch in the image, and its color reflects the logit for that class produced by the classifier. In Figure 2 we show one image and the heatmaps produced by BagNet-9 for 6 classes. The heatmaps show which regions in the image vote most strongly for that class. The model's top prediction is "broccoli", since it has the largest average. We can also see that BagNet-9 is slightly confused between bird and greens in this case, perhaps because the image contains two distinct objects. Consequently, as is standard in the literature, we measure top-5 accuracy where possible: when measuring accuracy, we credit the model as correct if the true label appears anywhere among the top 5 classes predicted by the model.

Figure 3 illustrates how BagNet is vulnerable to attack. It shows an image with an untargeted adversarial sticker in the upper-right and the heat maps BagNet produces. The true class is "monitor", but the sticker pushes this out of the top 5. We can see from the heatmaps that the sticker fools the model by assigning small negative scores to the true class (blue), and by assigning very large positive scores to the next 5 classes (bright red). Even though the sticker affects only 9 values in the heatmap, changing those to have very large positive scores drives up the average for those 5 classes by a significant amount. In effect, the problem with the average is that it is not robust to large changes to a few values.

BagNet is not robust against sticker attacks, but provides a starting point for our defense. In the rest of the paper we show how addressing this problem makes the resulting classifier more robust to sticker attacks.

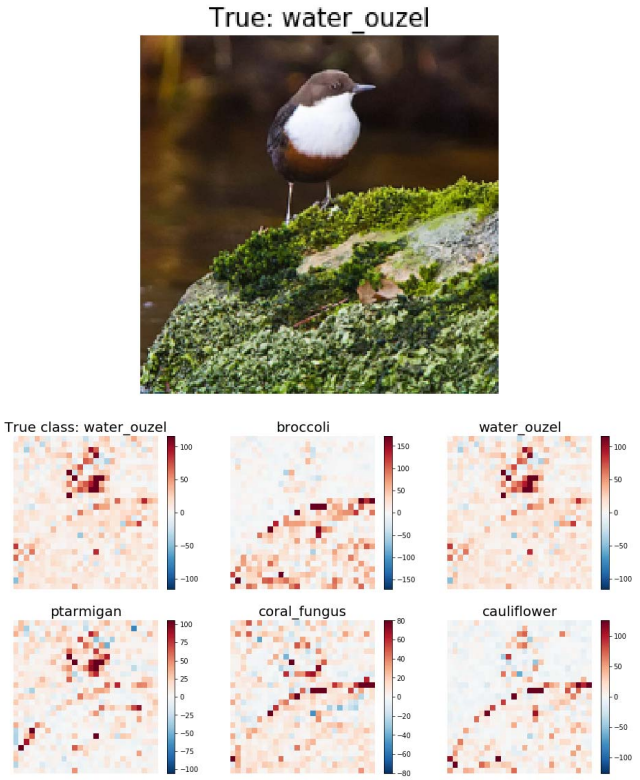

Fig. 2. An image from ImageNet and the corresponding heatmaps produced by BagNet-9. The first heatmap corresponds to the true label ("water_ouzel") while the rest are for the top 5 predictions generated from BagNet-9.

III. Problem and Approach

We study how to make object classification robust against sticker attacks in the digital domain. Thus, the attacker may choose a square region (of fixed size) in the image and control the exact value of all pixels within that region. This gives the attacker more power; there is no need to construct an attack that will be robust to variation in pose, viewpoint, brightness, or camera distortion. A defense that is robust against digital attacks will also necessarily be robust against physical attacks, making our results all the more meaningful.

A. Threat Model

We allow the attacker to freely choose the location where the digital sticker will be placed and replace the contents of that region with any valid values. In other words, given the initial image x_0 and a sticker size $s \times s$, the attacker constructs a location (i, j) in the image and a new image x' that agrees with x_0 everywhere except for locations $(i, j), \ldots, (i + s - 1, j + s - 1)$.

We evaluate security against white-box attacks, meaning that the adversary knows the model being used, its architecture, its weights, and the image being attacked. Then, guided by [10], we evaluate our CBN against black-box attack, where adversary has no access to the model and its parameters, but is still able to access to the image being attacked. We use the

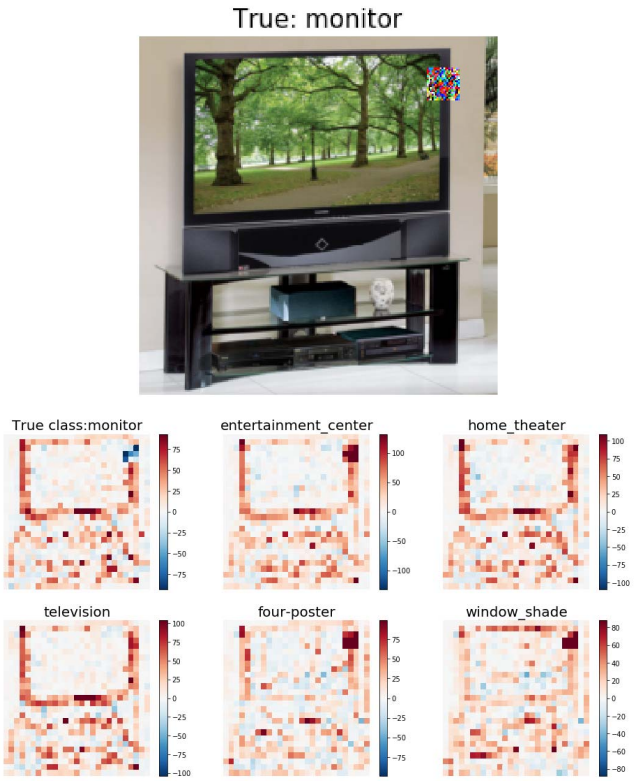

Fig. 3. An image from ImageNet with an adversarial sticker, and the heatmaps produced by BagNet-9. In this case the attack is successful: the true class is not among the top 5 predictions.

term "sticker" for the attack, as "patch" has a special meaning in the BagNet architecture [5].

In most of our experiments, we focus on a sticker with size of 20×20. We limit the sticker size to 20×20 as there are images from the ImageNet dataset that have objects which are smaller than that. We do not think it is reasonable to expect the classifier to accurately predict the class of an image when a well-placed sticker can completely occlude the true object.

B. Evaluation

We evaluate the security of a model in two ways.

First, we evaluate its security against standard algorithms for generating adversarial examples. We try to construct adversarial stickers, and report how often we are able to do so. If the adversary can construct an adversarial sticker for even a single location, we consider the attack to have succeeded against that image, and we count the fraction of images that remain correctly classified despite these attacks.

Second, our scheme provides certified security: for some images, we can prove that no sticker (no matter where placed, and no matter its contents) of a certain size will be able to change the classification. We report the fraction of images that can be certified safe in this way.

ImageNet models are normally evaluated using top-5 accuracy, i.e., the fraction of images for which the true label appears as one of the top 5 labels predicted by the model. Accordingly, for untargeted attacks, we consider the attacker to succeed if they successfully remove the true label from the top 5 predictions. (Thus, a clean image where the true label does not appear among the top 5 predictions already counts as a success for the attacker, with no sticker needed.) In contrast, it is unclear what the right notion of security for targeted attacks is, in a top-5 setting[1]. Therefore, for targeted attacks, we evaluate using top-1 accuracy and consider the attack a success if the attacker causes the target label to be the model's top prediction.

C. Generating Adversarial Stickers

One can search for an adversarial sticker by exhaustively trying all possible locations for a 20×20 sticker, and for each location, search for contents of the sticker that change the classification. However, searching all possible locations is very expensive, as there are $205^2 \approx 42,000$ possible locations, and each location would require tens or hundreds of evaluations of the model. For efficiency we test locations with a stride of 20, so we only need to search $11^2 = 121$ locations. At each location, we search for an adversarial example with unbounded perturbation within the region. In practice, we iterate the location with stride of 20 from left to right, top to bottom, and we terminate the search as soon as we find a single location where the attack succeeds.

D. Attack Algorithms

We use several methods to construct the contents of the sticker.

1) Untargeted Attacks: To evaluate security against untargeted attacks, we use Projected Gradient Descent (PGD), an iterative white-box attack that uses gradient information from the model to search for contents of the sticker that will maximize the loss of the model:

$$x^{t+1} = \Pi\left(x^t + \alpha \operatorname{sgn}\left(\nabla_x L(\theta, x^t, y)\right)\right),$$

where x^t denotes the image after t iterations, y the true class, θ the weights of the network, L the cross-entropy loss, and Π a projection to the space of valid images. Mądry et. al [11] show empirically and motivate theoretically that PGD is a universal first-order adversary, the strongest attack among those that use the local gradients.

2) Targeted Attacks: We evaluate security against targeted attacks using PGD with a different objective function [12]:

$$\Phi(x') = Z(x')_{y_0} - \max_{j \neq y_0} Z(x')_j,$$

where $Z(x')$ denotes the logits of the network on image x' and y_0 the target label. We then use PGD to maximize $\Phi(x')$. In our experiments, we choose a target label y_0 uniformly at random from among the labels other than true label.

[1]We could count the attack a success if the target label appears among the top 5 predictions; or alternatively, we could count it a success if the target label appears among the top 5 predictions and the true label is not among the top 5 predictions. It is not clear which definition is more appropriate; which is more suitable may depend on the application and how the model is used in a larger system.

3) Gradient-free Attack: To ensure that our defense is secure and not just masking the gradient, we also try SPSA, a black-box gradient-free attack [13].

IV. Clipped BagNet

A. BagNet

BagNet's structure as an average of per-patch logits forms a foundation for our defense. This structure reduces the influence of any single pixel, as each pixel can only influence the few patches that surround it, unlike a traditional feed forward neural network where a single pixel can have unlimited influence on the entire network. However, this alone is not enough for robustness against adversarial stickers, as a sticker can drive a single patch's logit arbitrarily high or low, thus making an unbounded change to the global average and potentially causing mispredictions. Indeed, we show later that unmodified BagNet is not robust against adversarial stickers.

B. Clipping Functions

The primary weakness of BagNet arises because the average of unbounded values is not robust: a large change to a single value can cause an unbounded change to the global average. To address this, we artificially clip the per-patch logits before averaging them, thereby limiting the extent to which our CBN can be influenced by an adversarial sticker. We experimented with multiple clipping functions, including sigmoid, tanh, and binarization, and found that tanh seemed to perform at least as well as any of the others. After a grid search, we selected the clipping function $f(x) = \tanh(ax + b)$ with hyperparameters set to $a = 0.05$ and $b = -1$. [2]

Mathematically, CBN works as follows. Let $Z(x)$ denote the vector of the global logits for input image x (so that $Z(x)_c$ denotes the global logit for class c), $Z_p(x)$ the vector of logits for patch p on image x, $\mathcal{P}$ the set of all patch locations, and f the clipping function. Then CBN calculates

$$Z(x) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} f(Z_p(x)).$$

We added the clipping function to a pre-trained BagNet model.[3]

C. Certified Security

Next, we show how CBN can be used to obtain certified security for some images. In particular, for some images, we can prove that no sticker would change the model's classification.

This analysis relies on the fact that the output of *tanh* of clipping is in the range $[-1, +1]$. We use this to derive a lower and upper bound on the values of all global logits, if a sticker is applied at a particular location to the image. Instead of using the method described in section §III-C to search for sticker locations, we iterate through each possible location for a sticker and check whether any location could change the classification.

For untargeted attacks, if the lower bound for the true label's logit is greater than the 5^{th} largest of the logit upper bounds (other than the true label), then we can conclude that no sticker at that location could push the true label out of the top 5. If this holds at all sticker positions, then we certify the image as safe against untargeted sticker attacks. For targeted attacks, if the largest of the logit lower bounds is greater than the upper bound for the target label, then we certify the image as safe against targeted sticker attacks.

Our bounds use the following results. In the following, let s denote the position of a sticker, define $\mathcal{P}_s$ as the set of patch locations p that overlap with s, and define $\mathcal{P}_{\bar{s}} = \mathcal{P} \setminus \mathcal{P}_s$ as the patch locations that do not overlap with s.

Lemma 1. *Let x, x' be two images that differ only by a sticker in position s. Then $\alpha(x, s) \leq Z(x') \leq \beta(x, s)$ where*

$$\alpha(x, s) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\bar{s}}} f(Z_p(x)) - \frac{|\mathcal{P}_s|}{|\mathcal{P}|}$$

$$\beta(x, s) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\bar{s}}} f(Z_p(x)) + \frac{|\mathcal{P}_s|}{|\mathcal{P}|}.$$

Proof. If p does not overlap with s, then $Z_p(x) = Z_p(x')$, so

$$\begin{aligned} Z(x') &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\bar{s}}} f(Z_p(x')) + \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_s} f(Z_p(x')) \\ &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\bar{s}}} f(Z_p(x)) + \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_s} f(Z_p(x')) \\ &\leq \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\bar{s}}} f(Z_p(x)) + \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_s} 1 \\ &= \beta(x, s). \end{aligned}$$

The lower bound can be derived using similar reasoning. □

Theorem 1. *If for every sticker position s there exists a class c with $c \neq t$ and $\alpha(x, s)_c > \beta(x, s)_t$, then x can be certified safe against a targeted attack to class t: no image x' obtained by placing a sticker somewhere will be classified by CBN as class t.*

Proof. The conditions imply that, no matter where the sticker is placed, the largest logit is never class t. The softmax preserves the relative order of the classes, so the classifier's top prediction will never be class t. □

Theorem 2. *If for all sticker positions s, $\alpha(x, s)_y$ is greater than the 5th largest value in $\{\beta(x, s)_c : c \neq y\}$, then x can be certified safe against untargeted attacks: CBN's classification of every image x' obtained by placing a sticker somewhere will have y among the top 5 classes.*

Proof. The conditions imply that, no matter where the sticker is placed, y will always be among the 5 largest logits. Softmax preserves the order of the classes, so y will always be among the classifier's top 5 predictions. □

[2] We used the same clipping function and parameters for every patch. In future work it might be interesting to try different parameters for each patch. Other tools from robust statistics, such as differential privacy, may also be relevant.

[3] BagNet was trained without clipping. It is plausible that re-training the model end-to-end with clipping present might yield better results than we report in this paper.

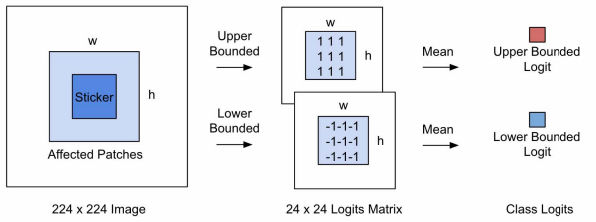

Fig. 4. Computing upper and lower bounds on logits, for certified security.

TABLE I
CLEAN ACCURACY AND ROBUSTNESS AGAINST 20×20 STICKER OF TESTED ARCHITECTURES

Scheme	Clean Accuracy Top-1	Clean Accuracy Top-5	Attack Robustness Targeted	Attack Robustness Untargeted
ResNet-50	78.6%	93.4%	53.0%	30.8%
ResNet-101	**79.4%**	94.2%	43.8%	28.2%
DenseNet	**79.4%**	**94.8%**	38.6%	18.2%
BagNet-33	69.0%	87.2%	54.2%	18.8%
CBN (Ours)	62.0%	83.6%	**99.8%**	**65.2%**

V. EXPERIMENTS

To evaluate CBN's robustness, we test it against what we believe are the strongest existing attacks, including adaptive attacks tailored to CBN. We only consider square stickers, but our analysis can be generalized to any other shape, since a sticker can only affect those patches that cover it. In other words, the square sticker in the Figure 4 could be replaced with any other shape as long as its affected patches remain the same. We evaluate against white-box untargeted attack (§V-A1), black-box untargeted attack (§V-A2), white-box targeted attack (§V-B), and report certified security (§V-C).

Guided by recommendations on defense evaluation in [10], we evaluate CBN against several state-of-the-art attacks: projected gradient descent (PGD), a white-box attack where the adversary has access to the model's parameters and thus its gradient, and SPSA [14], a black-box attack that does not require access to the gradients. The purpose of SPSA is to test whether the tanh clipping function introduces masked gradients [13] that prevent PGD from finding adversarial examples. Hyperparameters for the attacks are selected to generate the strongest attacks possible given our computational capacity. We verify the convergence of attack algorithms by doubling the number of iterations; the attack success rate was not appreciably affected. We evaluate on ImageNet and compare our CBN defense against three baseline neural networks: ResNet-50, ResNet-101, and DenseNet, which were the most robust of the conventional models that we have tested.

Table I summarizes the results of our experiments. CBN has reduced clean accuracy but approximately double robustness to attack. For efficiency, all evaluations are done on the same 500 randomly sampled images from ImageNet dataset. Under untargeted PGD attack, we record the top-5 accuracy after applying the adversarial sticker. Under targeted PGD attack, we record the percentage of images where the attack fails (i.e., is unable to cause the model to output the target class as its top-1 prediction). The latter number can be larger than the accuracy of the model, as misclassifying the image to something other than the target label is still counted as a failure of the attack.

A. Untargeted Attack

1) White Box Attack: PGD: We use a white-box PGD attack, restricted to a 20×20 sticker and with $\epsilon = 1$, so that there are no limits on how much each pixel within the sticker can be changed. We found that the Adam optimizer with learning rate 0.1, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ performed better than vanilla gradient descent. We report results for 80 iterations; we observed a similar success rate against CBN with 40 iterations, suggesting the attack has converged. Overall, CBN is significantly more robust against this attack than the other undefended models we evaluated.

2) Black Box Attack: SPSA: The *tanh* clipping function saturates at very large and very small values. It is conceivable that this could cause gradient to vanish, creating challenges for gradient-based attacks [13]. Therefore, we evaluate our scheme using a non-gradient based black-box attack as well, to verify that our defense is not merely masking the gradient. Three state-of-the-art black-box attacks in the literature are SPSA, NES [15] and boundary attack [16]. We chose SPSA because it achieves state-of-the-art results competitive with NES, and because the boundary attack is not applicable in our threat model.[4] We compare the effectiveness of PGD vs SPSA at driving the true label out of the top-5.

The original SPSA paper [14] introduces a black-box untargeted attack using the objective

$$\begin{array}{ll} \min_x & Z(x)_{y_0} - \max_{j \neq y_0} Z(x)_j \\ \text{s.t.} & \|x - x_0\|_\infty < \epsilon \end{array}$$

Here, ϵ is the maximal perturbation on a pixel, x_0 is the original image, y_0 is the ground true class of x_0, $Z(x)_{y_0}$ is the logit assigned to the true class, and $Z(x)_j$ is the logit for class j. SPSA uses the Adam optimizer to minimize this loss.

We adapt the SPSA attack to our setting of adversarial sticker. In particular, we seek an adversarial sticker where the attacker is limited to the sticker region with no limit on the size of the perturbation (as long as perturbed pixels remain valid), in contrast to the original paper which considered a small perturbation to the entire image. Also, we focus on top-5 accuracy rather than top-1 accuracy. Accordingly, we use an objective function that minimizes the true class's logit and maximizes the 5th largest of the other logits:

$$\min_x \quad Z(x)_{y_0} - \text{sort}\{Z(x)_j \mid j \neq y_0\}[5]$$

To find an optimal hyperparameter combination of SPSA and get a sense on how SPSA performs within our threat model, we apply SPSA on two sets of images, vulnerable images (10

[4]The boundary attack must be initialized with a sticker that causes the image to be misclassified, which is exactly the goal of the attack in the first place.

randomly selected images that were successfully attacked by PGD) and robust images (10 randomly selected images that survived PGD). With the best combination of hyperparameters (500 iterations, step size of 0.1, and randomly initializing stickers), 10 out of 10 vulnerable images and 1 out of 10 robust images are successfully attacked by SPSA. This suggests that SPSA is not significantly more powerful than PGD and is not suffering from masked gradients.

Then, we further verify that PGD was not subject to gradient masking by empirically verifying that PGD and SPSA converge to similar values. We first randomly sampled 133 images from the ImageNet validation set and for each image, randomly picked a location. Then, we run both SPSA and PGD on these sampled images at their sampled locations for 500 iterations. Figure 5 shows a scatter plot of the loss after the 500th iterations of PGD vs SPSA.

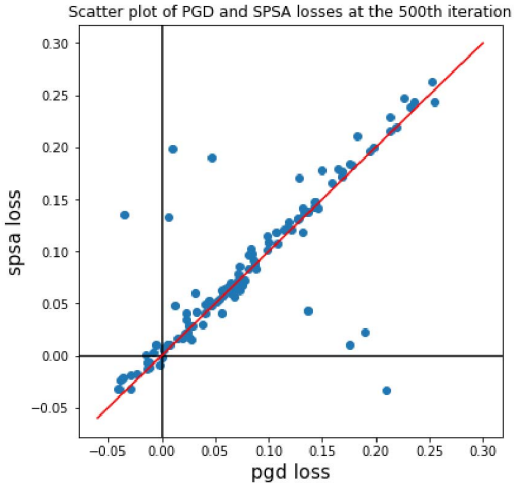

Fig. 5. Scatter plot of loss from PGD vs SPSA.

We can see from the scatter plot that PGD and SPSA are of approximately equal power. The red diagonal line shows where both attacks achieve the same loss. Points below the red line are cases where SPSA was more effective than PGD (e.g., due to gradient masking). We can see that in most cases both attacks achieve very similar results, and it is rare for SPSA to find a significantly better attack than PGD. This provides further evidence that PGD is an effective attack that accurately reflects the true robustness of CBN.

B. Targeted Attack

We use a white-box PGD attack with the margin-based loss from section §III-D2, random sticker initialization, and 80 iterations of the Adam optimizer. We pick a random target class (excluding the true class), then attack each location with stride 20, effectively measuring the average-case security against targeted attacks. Overall, CBN is significantly more robust against this attack than the undefended models we studied.

C. Certified Security

Evaluating a defense against state-of-the-art attack algorithms upper bounds the effectiveness of the defense, but leaves open the possibility that there might be even better attacks we haven't found yet. Our Clipped BagNet defense also allows us to compute a lower bound on its effectiveness. Specifically, we evaluate on the same 500 sampled images, and count the fraction that can be certified safe in both the untargeted and targeted setting using the methods from section §IV-C. Table II shows the results.

Note that all sticker sizes in the range $(8k-7)\times(8k-7)$ to $8k\times 8k$ share the same certified security. This occurs because patches in CBN have a stride of 8, and because our certified safety bounds make the worst-case assumption that once a sticker overlaps with a patch, the attacker can completely control the logit given by that patch.

TABLE II
CERTIFIED SECURITY ON VARIOUS SIZES OF STICKERS

Sticker Size	Targeted	Untargeted
$1\times 1 \sim 8\times 8$	99%	50%
$9\times 9 \sim 16\times 16$	97%	32%
$17\times 17 \sim 24\times 24$	95%	20%
$25\times 25 \sim 32\times 32$	86%	10%
$33\times 33 \sim 40\times 40$	72%	5%
$41\times 41 \sim 48\times 48$	46%	3%
$49\times 49 \sim 56\times 56$	30%	0%
$57\times 57 \sim 64\times 64$	11%	0%
$65\times 65 \sim 72\times 72$	4%	0%

VI. RELATED WORK

Chou *et al.* [17] introduce SentiNet which is capable of detecting a wide range of adversarial attacks including physical sticker attack, data poisoning attacks, and trojaned models. For sticker attack, SentiNet focuses on *universal* adversarial stickers, such as [3] and [6], where the adversarial sticker works no matter where it is located in the image and no matter what image it is applied to. In classifying as an attack, they use the ability of a salient region to influence classification when pasted in other images. In contrast, CBN defends against adversarial stickers regardless of them being robust or specialized to one image. We also provide a certified robustness guarantee.

Concurrently, Wu *et al.* [18] also study defense against adversarial sticker attack. They focus on are face and traffic sign recognition, while we focus on the ImageNet dataset, which makes our results difficult to compare directly. They use standard adversarial training, which is expensive to apply to ImageNet dataset; our CBN defense does not involve any form of model retraining and can be applied to large-scale datasets like ImageNet. Also, our approach is able to provide certified security bounds. Chiang *et al.*

Chiang *et al.* [19] propose certified defenses for adversarial patches, but they haven't tested their defenses on ImageNet dataset; they study 2×2 and 5×5 stickers for MNIST and CIFAR.

VII. Conclusion

In this paper, we propose Clipped BagNet as our defense against an adversarial sticker attack, and comprehensively evaluate the defense. Compared with undefended baseline model, our scheme significantly improve models' robustness against various adaptive adversarial sticker attacks (white/black-box, and untargeted/targeted). We also are able to evaluate the lower bound of our defense. Since BagNet only utilizes local features, our defense may be generalizable to other local feature extractors, or could likely be extended against other attacks based on regional modification on images, such as backdoor attack [20]. We have not evaluated the effectiveness of these generalizations; we leave it to future works.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013. I

[2] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," 2018. I

[3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," *arXiv e-prints*, p. arXiv:1712.09665, Dec 2017. I, II-A, VI

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Models," *arXiv e-prints*, p. arXiv:1707.08945, Jul 2017. I, II-A

[5] W. Brendel and M. Bethge, "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet," *arXiv e-prints*, p. arXiv:1904.00760, Mar 2019. I, II-B, III-A

[6] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *CCS '16*, 2016. II-A, VI

[7] S. V. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," *ArXiv*, vol. abs/1908.08705, 2019. II-A

[8] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," 2019. II-A

[9] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. L. Lin, "Evading real-time person detectors by adversarial t-shirt," *ArXiv*, vol. abs/1910.11099, 2019. II-A

[10] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On Evaluating Adversarial Robustness," *arXiv e-prints*, p. arXiv:1902.06705, Feb 2019. III-A, V

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv e-prints*, p. arXiv:1706.06083, Jun 2017. III-D1

[12] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," *arXiv e-prints*, p. arXiv:1608.04644, Aug 2016. III-D2

[13] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," *arXiv e-prints*, p. arXiv:1802.00420, Feb 2018. III-D3, V, V-A2

[14] J. Uesato, B. O'Donoghue, A. van den Oord, and P. Kohli, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," *arXiv e-prints*, p. arXiv:1802.05666, Feb 2018. V, V-A2

[15] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Query-efficient black-box adversarial examples (superceded)," 2017. V-A2

[16] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," 2017. V-A2

[17] E. Chou, F. Tramèr, G. Pellegrino, and D. Boneh, "Sentinet: Detecting physical attacks against deep learning systems," 2018. VI

[18] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," in *International Conference on Learning Representations*, 2020. VI

[19] P. yeh Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studor, and T. Goldstein, "Certified defenses for adversarial patches," in *International Conference on Learning Representations*, 2020. VI

[20] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *CoRR*, vol. abs/1708.06733, 2017. VII