Defending against Adversarial Patches with Robust Self-Attention

Norman Mu [1 2] **David Wagner** [1]

Abstract

We introduce a new defense against adversarial patch attacks based on our proposed Robust Self-Attention (RSA) layer. Robust Self-Attention replaces the outlier-sensitive weighted mean operation used by standard Self-Attention with a robust aggregation mechanism that detects and masks outlier tokens. Vision Transformer (ViT) models using our RSA layer achieve promising robust classification accuracy, outperforming patch adversarial training as well as a prior provable defense, all with zero additional parameters or training. Additionally, we provide further evidence for the strength of simple patch adversarial training as a baseline defense.

1. Introduction

Seeking a more realistic threat model than that of standard adversarial examples (Gilmer et al., 2018), the robustness community has explored many variations of patch adversarial examples in prior work.

Prior work has suggested that patch adversarial examples operate by inducing abnormally large class logit scores which overwhelm the contribution of benign image regions (Rao et al., 2020). Building off this insight, we design and evaluate a new patch adversarial defense.

In this work we assume a common patch adversarial threat model, where the attacker has whitebox access to the model and computes image- and location-specific perturbations. These perturbations consist of a single square patch of fixed size which is known to the defender.

2. Related Work

2.1. Patch Adversarial Examples

Brown et al. (2017) explore universal adversarial patches that can be applied at a variety of sizes and angles in the real world to fool image classifiers, while Karmon et al. (2018) introduce the threat model explored in this work, in addition to several related threat models. As in Rao et al. (2020), we focus on untargeted attacks for their ease of use in adversarial training.

2.2. Certifiable Patch Defenses

Following the success of certified robustness methods for standard adversarial attacks (Gowal et al., 2018; Raghunathan et al., 2018; Cohen et al., 2019), various works have demonstrated varying degrees of certifiable robustness against patch adversaries by applying interval bound propagation (Chiang et al., 2020) and randomized smoothing (Levine & Feizi, 2020).

Other methods use networks with limited receptive fields (Brendel & Bethge, 2019) to provide robustness against patch attacks, by robustly aggregating predictions from each region of the image (Zhang et al., 2020; McCoyd et al., 2020; Xiang et al., 2020; Xiang & Mittal, 2021).

Our method is inspired by this same intuition and takes advantage of the unique patch-based processing of the vision transformer architecture to insert a robust aggregation mechanism into every Self-Attention layer. We do not explore certified robustness in this work.

3. Robust Self-Attention

3.1. Self-Attention

We briefly review the Self-Attention layer (Vaswani et al., 2017) at the center of the vision transformer (ViT) architecture (Dosovitskiy et al., 2020). Given a set of N token vectors x_i, query, key, and value vectors q_i, k_i, v_i are computed as an affine function of x_i. N pair-wise attention weights are then computed for each token x_i via a scaled dot product and normalized with softmax:

$$\alpha_{i,j} = \text{softmax}_j \frac{q_i \cdot k_j}{\sqrt{D}}$$

where D is the dimension of all vectors. Finally, the updated value for each token is computed by aggregating the value vectors according to the pairwise attention weights specific to each token:

[1] UC Berkeley [2] Facebook AI Research. Correspondence to: Norman Mu <thenorm@berkeley.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

$$z_i = \sum_{i=1}^{N} \alpha_{i,j} v_j$$

The Self-Attention layer can be applied multiple times per Transformer layer in parallel with different weights to form *Multi-headed* Self-Attention, which is analogous to applying multiple convolutional filters per layer of a convolutional neural network.

3.2. Vision Transformer

The Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020) adapts transformer networks (Vaswani et al., 2017) to images. To perform image classification, the input image is first split into a set of contiguous, non-overlapping square patches (generally 16×16px). Each patch is treated as a token for the attention mechanism.

Each internal layer of the vision transformer consists of a Multi-headed Self-Attention layer applied to these tokens, then a two-layer neural network is used to update the value of each token. We apply a final classification layer to the value of a particular token after the last Transformer layer to compute the class logits. We refer the reader to the original ViT paper (Dosovitskiy et al., 2020) for further details. The intuition is that each token contains information about a spatially localized region of the image, which we use in our defense against patch attacks.

3.3. Robust Aggregation

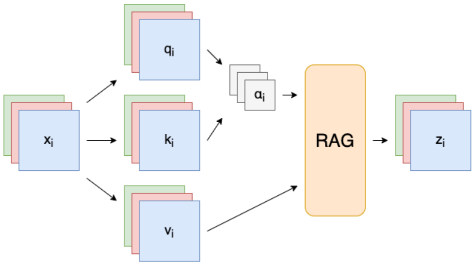

Figure 1. An illustration of the proposed Robust Self-Attention (RSA) layer, which replaces the weighted mean operation in Self-Attention with a robust aggregation mechanism (RAG).

The final aggregation step in Self-Attention consists of a weighted mean across tokens, which is sensitive to outliers and therefore susceptible to adversarial manipulation. We propose the Robust Self-Attention layer, shown in Figure 1. RSA switches the weighted mean operation with a robust aggregation mechanism we call RAG, and can be used as a drop-in replacement for Self-Attention layers in the ViT network without any finetuning or additional modification.

Single suspicious token. First, we address the simplified setting in which only a single token lies within the window of the adversarial patch. Define the anomaly score of an image token as the L_2 distance between the token's value vector and the mean of all value vectors corresponding to image tokens (excluding the first classification token). The image token with the highest anomaly score is presumed untrustworthy and its value vector is replaced with the mean value vector. All attention weights associated with the untrustworthy token are replaced with a value of $1/N$, where N is the total number of ViT tokens.

Multiple suspicious tokens. In practice, it is possible for multiple ViT patches to fall within the window of a single adversarial patch if the adversarial patch straddles the border between two or more ViT patches. RAG accounts for this by leveraging the spatial contiguity of the adversarial patch, assuming knowledge of the maximum adversarial patch size. We define the anomaly score s of a window w as the average anomaly score of all tokens within the extent of w.

Now, the window w^* with the highest anomaly score is presumed untrustworthy and all vectors and weights corresponding to tokens within w^* are masked. In practice, we need only to consider the set of windows covering unique sets of ViT patches. Computing the window anomaly scores can be implemented efficiently as an average pooling operation on the 2D grid of token anomaly scores.

Pseudocode for a single-headed version of RSA is given in Algorithm 1. To accommodate multi-headed attention, we average the anomaly scores across attention heads and select the image token with the greatest overall anomaly score.

Adversarial patch sizes. We note here that Robust Self-Attention requires a single hyperparameter specifying the size of the adversarial patches, which our threat model assumes knowledge of. This in turn determines how many contiguous image tokens are masked out by the RAG mechanism. When evaluating on adversarial patches of size 0, i.e. clean images, RSA is equivalent to standard Self-Attention as no patches are masked out. If more image tokens than necessary are masked out, for instance when we assume a non-zero adversarial patch size on clean images, accuracy will suffer.

4. Methods

4.1. Dataset

We evaluate all methods on a random 100-class subset of ImageNet which we refer to as ImageNet-100. Training is performed on all 127k images, while accuracy is evaluated on 512 randomly sampled images from the 100 classes. Both training and evaluation are performed at the standard image size of 224px.

Algorithm 1 Pseudocode for single-headed Robust Self-Attention.

```
input tokens $\{x_i\}$, linear layers $Q, K, V$, the set of sliding
  windows $\mathcal{W}$, number of tokens $N$, and token dimension
  $D$
output tokens $\{z_i\}$
  function ROBUSTSELFATTENTION
    for each $1 \leq i \leq N$ do
      $q_i, k_i, v_i \leftarrow Q(x_i), K(x_i), V(x_i)$
    end for
    for each $1 \leq i, j \leq N$ do
      $\alpha_{i,j} \leftarrow \textsc{Softmax}_j \; q_i \cdot k_j / \sqrt{D}$
    end for
    return RAG($\{v_i\}, \{\alpha_{i,j}\}$)
  end function
input values $\{v_i\}$, attention weights $\{\alpha_{i,j}\}$
output tokens $\{z_i\}$
  function RAG
    $\mu \leftarrow \textsc{Mean}_{2 \leq i \leq N}(v_i)$
    for each $w \in \mathcal{W}$ do
      $s(w) \leftarrow \sum_{i \in w} ||v_i - \mu||_2$
    end for
    $w^* \leftarrow \textsc{Argmax}_{\mathcal{W}} \; s(w)$
    for all $i \in w^*$ do $v_i, \alpha_{*,i} \leftarrow \mu, \frac{1}{N}$
    for each $1 \leq i \leq N$ do
      $z_i \leftarrow \sum_j \alpha_{i,j} \cdot v_j$
    end for
    return $\{z_i\}$
  end function
```

4.2. Patch Adversaries

Existing untargeted L_∞ attacks can be easily adapted for patch attacks by fixing $\epsilon = 255/255$, scaling up the attack step size, and applying the adversarial perturbation δ to an image x with

$$x_{adv} = m \odot \delta + (1 - m) \odot x,$$

where $\odot$ indicates element-wise multiplication and m is a binary mask indicating where the patch is active (Rao et al., 2020).

Training. We apply adversarial training, with PGD patch attack (Kurakin et al., 2017; Madry et al., 2018). The attack, which we call PatchPGD, uses 10 steps of PGD and a fixed step size of 0.25, using basic SGD without momentum for optimization. A single patch location within the image extent is randomly sampled for each image during training, and a single patch size is randomly sampled from $U[10, 50]$ for each training batch.

Evaluation. We evaluate the robustness of all models using the more powerful AutoPGD (DLR) attack, taken from AutoAttack (Croce & Hein, 2020). This attack, which we call PatchAutoPGD, uses the standard AutoPGD parameters: 100 steps with an adaptive step size, optimized using SGD with momentum. PatchAutoPGD is evaluated at evenly-spaced patch locations along a grid with stride 20, following Zhang et al. (2020). Consequently, 10px and 20px patches are evaluated at 121 locations, 30px and 40px patches at 100 locations, and 50px patches at 81 locations.

4.3. Models

Undefended models. We use ResNet-50 (He et al., 2016) as a baseline convolutional network. Our vision transformer uses the DeiT-small model (Touvron et al., 2020), which we refer to as ViT-small in order to emphasize the architecture over the training details. Both models have a similar number of parameters: 23M params for ResNet-50 and 22M params for ViT-small.

We adapt a pre-trained network, trained on the full ImageNet dataset, to our 100-class subset by replacing the original 1000-way linear layer L with a constructed 100-way linear layer L'. The weight matrix of L' consists of the submatrix of L's weight matrix corresponding to the 100 classes of interest. The bias vector is similarly cut down to size, and all other weights are left unchanged.

Adversarially trained models. We evaluate the effectiveness of adversarial training at defending against patch attacks. Our adversarially trained models are initialized with weights from an undefended model and finetuned for 40 epochs with a batch size of 512 against PatchPGD. ResNet-50 is trained using momentum SGD with learning rate 0.1, momentum 0.9, and weight decay 5e-5. ViT-small is trained using AdamW with a learning rate of 5e-6 and weight decay of 5e-3. The learning rate is dropped by $10\times$ halfway through training, at 20 epochs.

Robust Self-Attention models. Our proposed method consists of directly replacing the Self-Attention layers in ViT with Robust Self-Attention layers without any additional modifications. This can be performed on both undefended models as well as models already trained against a patch adversary, and we evaluate both. As discussed above, RSA uses the known adversarial patch size to determine how many image tokens to mask out during inference.

Clipped BagNet. We also evaluate Clipped BagNet (Zhang et al., 2020), which clips per-patch BagNet-33 (Brendel & Bethge, 2019) predictions using $\tanh(0.01x - 1)$, before averaging into a single global prediction.

5. Results

As shown in Table 1, patch adversarially trained ViT-small exhibits much lower robust accuracy at all attack sizes compared to patch adversarially trained ResNet-50. The un-

		Robust accuracy				
Network	Clean	10px	20px	30px	40px	50px
Clipped BagNet	77.15	53.32	33.59	18.55	8.79	2.73
ResNet-50	90.63	40.23	9.18	1.17	0.00	0.00
+ PatchPGD adv. tr.	91.41	74.80	67.58	64.65	59.57	54.69
ViT-small	**93.16**	45.70	0.00	0.00	0.00	0.00
+ PatchPGD adv. tr.	91.60	68.16	45.90	31.45	25.78	14.84
+ RSA	Tbl. 2	83.20	72.46	68.55	56.25	43.16
+ PatchPGD adv. tr. + RSA	Tbl. 2	**84.57**	**80.86**	**80.08**	**69.73**	**61.13**

Table 1. Classification accuracy on clean and adversarial ImageNet-100 validation images with varying attack sizes. Clean accuracy of ViT-RSA models depends on setting of patch size hyperparameter and is presented in Table 2. Best results are in **bold**.

Network	0px	10px	20px	30px	40px	50px
ViT-small + RSA	93.16	92.58	90.43	90.43	88.67	83.98
ViT-small + PatchPGD adv. tr. + RSA	91.60	89.45	88.09	88.09	87.11	84.18

Table 2. Classification accuracy on clean ImageNet-100 images of ViT-RSA models, under varying settings of known patch attack size. 0px assumes no attack and is equivalent to the standard ViT-small architecture evaluated on clean images.

defended ViT-small model also reaches 0% accuracy with much smaller attacks (20px) than ResNet-50 (40px). However, our experiments indicate that simply replacing Self-Attention with Robust Self-Attention provides a substantial improvement to robust accuracy at all patch sizes.

Combining ViT-RSA with patch adversarial training further improves robust accuracy, exceeding the robust accuracy of ResNet-50 by more than 15.4% against 30px attacks. Relative to patch adversarially trained ResNet-50, ViT-RSA with patch adversarial training trades off clean accuracy for robust accuracy against, as we can see in Table 2.

Notably and in contrast to standard adversarial training (Tsipras et al., 2019; Zhang et al., 2019), patch adversarial training actually improves clean accuracy— by 0.8% with ResNet-50. Patch adversarial training may be viewed as a form of cutout data augmentation (Devries & Taylor, 2017), which is well-known to improve model performance. Surprisingly, patch adversarial training hurts clean accuracy for ViT-small by 1.6%.

6. Discussion

Baselines. Simple patch adversarial training is a surprisingly strong baseline for defending convolutional neural networks against patch attacks in practice, even without location optimization (Rao et al., 2020). Further work is needed to evaluate how much location optimization helps.

ViT architecture. Our experiments suggest that the ViT architecture is more susceptible to adversarial patches than ResNet-50, both when undefended and when trained against patch adversaries. A significant possibility is that our hyperparameter tuning was insufficient for finding strong optimization hyperparameters for ViT patch adversarial training due to the difficulty of optimizing ViT models in general (Chen et al., 2021b;a).

Sub-optimal training also help explain why patch adversarial training improved clean accuracy for ResNet but hurt clean accuracy for ViT. Indeed, for many hyperparameter settings ResNet-50 achieves the highest robust validation accuracy after a single epoch of patch adversarial training, whereas ViT-small continues to slowly improve through training.

Robust aggregation mechanisms. We experimented with a variety of robust aggregation mechanisms inspired by the field of robust statistics. However, the simplest robust aggregation mechanisms such as clipping, trimming, and winsorization are not directly applicable to vector-valued variables. Robust aggregation mechanisms which do operate on vector-valued variables such as the weighted median and weighted medoid are either computationally infeasible (median) or yielded poor results (medoid).

Certified robustness. Our work raises the question of whether it is possible to develop robust aggregation mechanisms for ViTs that admit certified robustness proofs.

7. Conclusion

We introduced a new defense against patch adversarial examples which appears to outperform strong empirical baselines. While initial results are promising, further evaluation is required to confirm the robustness of our method.

8. Acknowledgements

This work was supported by Facebook through the FAIR-BAIR Commons, by the Berkeley Deep Drive project, and by Open Philanthropy.

References

Brendel, W. and Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *ArXiv*, abs/1904.00760, 2019.

Brown, T., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *ArXiv*, abs/1712.09665, 2017.

Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pretraining or strong data augmentations. 2021a.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. *ArXiv*, abs/2104.02057, 2021b.

Chiang, P.-Y., Ni, R., Abdelkader, A., Zhu, C., Studer, C., and Goldstein, T. Certified defenses for adversarial patches. *ArXiv*, abs/2003.06693, 2020.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Devries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

Gilmer, J., Adams, R. P., Goodfellow, I., Andersen, D. G., and Dahl, G. E. Motivating the rules of the game for adversarial example research. *ArXiv*, abs/1807.06732, 2018.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelović, R., Mann, T. A., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *ArXiv*, abs/1810.12715, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Karmon, D., Zoran, D., and Goldberg, Y. Lavan: Localized and visible adversarial noise. In *ICML*, 2018.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017.

Levine, A. and Feizi, S. (de)randomized smoothing for certifiable defense against patch attacks. *ArXiv*, abs/2002.10733, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

McCoyd, M., Park, W., Chen, S., Shah, N., Roggenkemper, R., Hwang, M., Liu, J., and Wagner, D. A. Minority reports defense: Defending against adversarial patches. In *ACNS Workshops*, 2020.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *ArXiv*, abs/1801.09344, 2018.

Rao, S., Stutz, D., and Schiele, B. Adversarial training against location-optimized adversarial patches. In *ECCV Workshops*, 2020.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and J'egou, H. Training data-efficient image transformers & distillation through attention. *ArXiv*, abs/2012.12877, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv: Machine Learning*, 2019.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Xiang, C. and Mittal, P. Patchguard++: Efficient provable attack detection against adversarial patches. *ArXiv*, abs/2104.12609, 2021.

Xiang, C., Bhagoji, A., Sehwag, V., and Mittal, P. Patchguard: Provable defense against adversarial patches using masks on small receptive fields. *ArXiv*, abs/2005.10884, 2020.

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

Zhang, Z., Yuan, B., McCoyd, M., and Wagner, D. A. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features. *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 55–61, 2020.