ImageNet-Patch: A Dataset for Benchmarking Machine Learning Robustness against Adversarial Patches

Maura Pintor [1 2] Daniele Angioni [1] Angelo Sotgiu [1 2] Luca Demetrio [1 2]
Ambra Demontis [1] Battista Biggio [1 2] Fabio Roli [3 2]

Abstract

Adversarial patches are optimized contiguous pixel blocks in an input image that cause a machine-learning model to misclassify it. However, their optimization is computationally demanding, and requires careful hyperparameter tuning. To overcome these issues, we propose ImageNet-Patch, a dataset to benchmark machine-learning models against adversarial patches. It consists of a set of patches, optimized to generalize across different models, and applied to ImageNet data after preprocessing them with affine transformations. This process enables an approximate yet faster robustness evaluation, leveraging the transferability of adversarial perturbations.

1. Introduction

Understanding the security of machine-learning models is of paramount importance nowadays, as these algorithms are used in a large variety of settings, including security-related and mission-critical applications, to extract actionable knowledge from vast amounts of data. Nevertheless, such data-driven algorithms are not robust against attacks, as malicious attackers can easily alter the behavior of state-of-the-art models by carefully manipulating their input data (Biggio et al., 2013; Szegedy et al., 2014; Carlini & Wagner, 2017; Madry et al., 2018). In particular, attackers can hinder the performance of classification algorithms by means of *adversarial patches* (Brown et al., 2017), i.e., contiguous chunks of pixels which can be applied to any input image to cause the target model to output an attacker-chosen class. When embedded into input images, adversarial patches produce out-of-distribution samples. The reason is that the injected patch induces a spurious correlation with the target label, which is likely to shift the input sample off the manifold of natural images. Adversarial patches can be printed as stickers and physically placed on real objects, like stop signs that are then recognized as speed limits (Eykholt et al., 2018), and accessories that camouflage the identity of a person, hiding their real identity (Sharif et al., 2016). Therefore, the evaluation of the robustness against these attacks is of the uttermost importance, as they can critically impact real-world applications with physical consequences. This process is costly, as adversarial patches should also be effective under different transformations, including affine transformations like translation, rotation and scale changes, to be effective in the physical world, and should successfully transfer across different models, given that, in practical scenarios, it is most likely that complete access to the target model (i.e., access to its gradients) is not provided.

To overcome these issues, in this work we propose ImageNet-Patch, a dataset of pre-optimized adversarial patches that can be used to benchmark machine-learning models with small computational overhead. This dataset is constructed with a subset of the ImageNet validation set. It consists of 10 patches that target 10 different classes, applied on $5,000$ images each, for a total of $50,000$ samples. We create these patches by leveraging an ensemble of models, forcing the algorithm to propose patches that evade them all to improve transferability, also under different affine transformations (Sect. 2). Even though the resulting robustness evaluation will be approximate, evaluating on a pre-defined dataset is extremely fast, and it provides a first step to evaluate the robustness of models (Sect. 3). We test the efficacy of ImageNet-Patch by evaluating the successful generalization of the patches to unseen models (Sect. 4). We conclude by discussing related work (Sect. 5), as well as the limitations and future directions of our work (Sect. 6). Our dataset is available at `https://zenodo.org/record/6568778`.

2. Crafting Transferable Adversarial Patches

Attackers can compute adversarial patches by solving an optimization problem with gradient-descent algorithms (Brown et al., 2017). To be used in the real world,

[*] Equal contribution [1] University of Cagliari, Italy [2] Pluribus One, Italy [3] University of Genova, Italy. Correspondence to: Maura Pintor <maura.pintor@unica.it>.

ICML 2022 Shift Happens Workshop: Crowdsourcing metrics and test datasets beyond ImageNet., Baltimore, Maryland, USA, 2022.

the patches should be robust to affine transformations, like rotation, translation and scale, that are unavoidable when dealing with this scenario. Hence, the optimization process must include these perturbations as well, to force such invariance inside the resulting patches. Also, they can either generate a general misclassification, i.e. an *untargeted* attack, or force the model to predict a specific class, i.e. a *targeted* attack. Formally, targeted adversarial patches are computed by solving the optimization problem:

$$\min_{\boldsymbol{\delta}} \mathbb{E}_{\mathbf{A}\sim\mathcal{T}} \left[\sum_{j=1}^{J} \mathcal{L}(\boldsymbol{x}_j \oplus \mathbf{A}\boldsymbol{\delta}, y_t; \boldsymbol{\theta}) \right], \tag{1}$$

where $\boldsymbol{\delta}$ is the adversarial patch to be computed, $\boldsymbol{x}_j$ is one of J samples of the training data, y_t is the target label,[1] $\boldsymbol{\theta}$ is the targeted model, $\mathbf{A}$ is an affine transformation randomly sampled from a set of affine transformations $\mathcal{T}$, $\mathcal{L}$ is a loss function of choice, that quantifies the classification error between the target label and the predicted one and $\oplus$ is a function that applies the patch on the input images. The latter is defined as: $\boldsymbol{x} \oplus \boldsymbol{\delta} = (\mathbf{1} - \boldsymbol{\mu}) \odot \boldsymbol{x} + \boldsymbol{\mu} \odot \boldsymbol{\delta}$, where we introduce a mask $\boldsymbol{\mu}$ that is a tensor with the same size of the input data $\boldsymbol{x}$, and whose components are ones where the patch should be applied and zeros elsewhere (Karmon et al., 2018). To produce a dataset that can be used as a benchmark for robustness assessment, with adversarial patches effective regardless of the target model, we consider an ensemble of models inside the optimization process. This addition forces the optimization algorithm to find effective solutions against the ensemble, boosting the transferability of the produced adversarial patches, i.e., the ability of the adversarial patch optimized against a model (or a set of them) to be effective against different models. Hence, the loss function to be minimized can be written as:

$$\min_{\boldsymbol{\delta}} \mathbb{E}_{\mathbf{A}\sim\mathcal{T}} \left[\sum_{m=1}^{M} \sum_{j=1}^{J} \mathcal{L}(\boldsymbol{x}_j \oplus \mathbf{A}\boldsymbol{\delta}, y_t; \boldsymbol{\theta}_m) \right], \tag{2}$$

where we modified Equation 1 to minimize the loss $\mathcal{L}$ over M models, parameterized by $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M$.

The objective function defined in Equation 2 can be optimized through gradient-descent techniques, and thus we use Algorithm 1 for minimizing it. After having randomly initialized the patch (line 1), we loop through the number of intended epochs (line 1), and the samples of the training data (line 1). In each epoch, we sample a random affine transformation that will be applied to the patch (line 1). We iterate over all models of the ensemble (line 1) to calculate the loss by accumulating its gradient w.r.t. the patch (line 1), and using it to update the patch at the end of each epoch (line 1). After all the epochs have been consumed, the final adversarial patch is returned (line 1).

Algorithm 1 Optimization of adversarial patches

```
 1: Input: $\boldsymbol{x}$, the training dataset containing $J$ images; $y_t$,
    the target class; $\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_M$, the ensemble of models; $\gamma$,
    the learning rate; $N$, the number of epochs.
 2: Output: $\boldsymbol{\delta}$, the adversarial patch
 3: $\boldsymbol{\delta} \sim U(0, 1)$
 4: for $i \in [1, N]$ do
 5:   $\boldsymbol{g} \leftarrow 0$
 6:   for $j \in [1, J]$ do
 7:     $\mathbf{A} \leftarrow$ random-affine()
 8:     for $m \in [1, M]$ do
 9:       $\boldsymbol{g} \leftarrow \boldsymbol{g} + \frac{1}{MJ} \nabla_{\boldsymbol{\delta}} \mathcal{L}(\boldsymbol{x}_j \oplus \mathbf{A}\boldsymbol{\delta}, y_t; \boldsymbol{\theta}_m)$
10:     end for
11:   end for
12:   $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} - \gamma \boldsymbol{g}$
13: end for
14: return $\boldsymbol{\delta}$
```

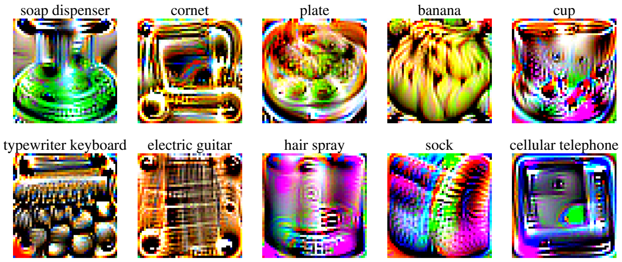

Figure 1. The 10 optimized adversarial patches.

3. The ImageNet-Patch Dataset

We now illustrate how we apply our methodology to generate the ImageNet-Patch dataset that will be used to evaluate the robustness of classification models against patch attacks. We start from the validation set of the original ImageNet database, containing $1,281,167$ training images, $50,000$ validation images and $100,000$ test images, divided into $1,000$ object classes. Then, we select the test set of $5,000$ images used in RobustBench (Croce et al., 2020) for testing model robustness against adversarial attacks. We create the corpus of images used to optimize adversarial patches from the remaining part of the ImageNet validation set randomly sampling 20 images from different classes.

We now define the ImageNet-Patch dataset. To optimize the patches on an ensemble, we select three deep neural network architectures trained on the ImageNet dataset, namely AlexNet (Krizhevsky et al., 2012), ResNet18 (He et al., 2016) and SqueezeNet (Iandola et al., 2016). We leverage the pretrained models available inside the PyTorch TorchVi-

[1] The same formulation holds for crafting untargeted attacks, by simply substituting the target label y_t with the ground truth label of the samples y, and inverting the sign of the loss function.

sion zoo,[2] trained to classify RGB images of size 224×224. We run Algorithm 1 to create squared patches with a size of 50×50 pixels, with a learning rate of 1, 20 training samples, 5000 training epochs, and using the cross-entropy loss. We include rotation and translation as affine transformations during the optimization of the patch, constraining rotations up to $\pm\frac{\pi}{8}$ to mimic the setup applied by Brown et al. (2017), and translations to a shift of ± 68 pixels on both axes from the center of the image (to avoid the patches being too close to the borders of the image). We optimize 10 patches with these settings, targeting 10 different classes of the ImageNet dataset ("soap dispenser", "cornet", "plate", "banana", "cup", "typewriter keyboard", "electric guitar", "hair spray", "sock", "cellular phone"). The resulting patches are shown in Fig. 1. We apply such patches to each of the $5,000$ images in the test set, generating a dataset of $50,000$ perturbed images with adversarial patches. We depict some examples of the applied patches in Fig. 4.

4. Experimental Analysis

We evaluate the evasion performance of the ImageNet-Patch dataset by considering three metrics: (i) the *clean accuracy* C_k, which is the accuracy of the target model in absence of attacks; (ii) the *robust accuracy* R_k, which is the accuracy of the target model in presence of adversarial patches; and (iii) the *success rate* S_k of a patch, that measures the percentage of samples for which the patch successfully altered the prediction of the target model toward the intended class. We denote with k the results obtained with the top-k scores, i.e. by computing the metric in the set of k highest outputs of the classification model $\boldsymbol{\theta}$ when receiving the sample x as input. We evaluate these three metrics for $k \in \{1, 5\}$.

To evaluate the effectiveness of the patches, we test ImageNet-Patch against 127 deep neural networks trained on the ImageNet dataset. To facilitate the discussion, we group the models in 5 groups. We denote as ENSEMBLE the models in the ensemble, STANDARD a set of standard-trained models, ADV-ROBUST a set of robust-trained models, AUGMENTATION a set of models robust to image perturbations and corruptions, and MORE-DATA a set of models trained on datasets that utilize substantially more training data than the standard ImageNet training set. We take all models from RobustBench[3] (Croce et al., 2020) and from the ImageNet Testbed repository[4] (Taori et al., 2020).

4.1. Experimental Results

We now detail the effectiveness of our dataset against the groups we have defined, sharing the results in Fig. 2. The ENSEMBLE group of models is characterized by low robust accuracy and the highest success rate of the adversarial patch. Such a result is expected since we optimize our adversarial patches to specifically mislead these models, as they are part of the training ensemble. The STANDARD group is characterized by a modest decrement of the robust accuracy, highlighting errors caused by the patches. The success rate is lower compared to those exhibited by the ENSEMBLE group, since patches are not optimized on these models. The ADV-ROBUST group is characterized by a drop of robust accuracy similar to the STANDARD group, but with a low success rate for the adversarial patches. This implies that robust models are affected by adversarial patches in terms of untargeted attacks, but not by targeted ones. The AUGMENTATION group contains mixed results, shifting from a modest to a severe drop in terms of robust accuracy, associated with an increment of the success rate, which is slightly less than that achieved by the STANDARD group. This might imply that augmentation techniques help the model to score good results on regular images, but performance drops when dealing with adversarial noise. Lastly, the MORE-DATA group outperforms the others in clean and robust accuracy while the success rate of the patches is similar to the AUGMENTATION group results. To better highlight the efficacy of our adversarial patches, we also depict the difference in terms of accuracy of these target models scored by applying our pre-optimized patches and randomly-generated ones in Fig. 3. The top row shows the

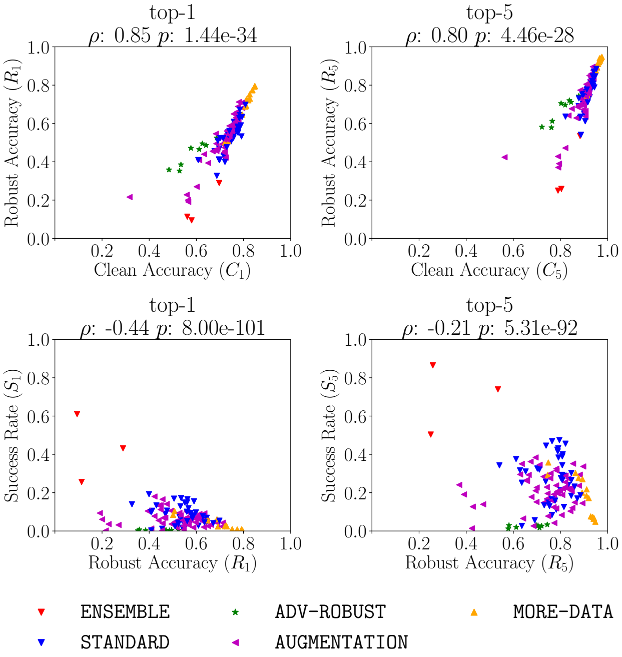

Figure 2. Results of our analysis on 127 models. *Top*: top-1 (left) and top-5 (right) clean accuracy vs robust accuracy. *Bottom*: top-1 (left) and top-5 (right) robust accuracy vs attack success rate.

[2] https://pytorch.org/vision

[3] https://robustbench.github.io/

[4] https://github.com/modestyachts/imagenet-testbed

results for the pre-optimized patches, while the bottom row focuses on the random ones, and each plot also shows a robust regression line, along with its 95% confidence interval. The regression analysis highlights meaningful observations we can extract from the benchmark. First, the robust accuracy of each model evaluated with random patches can be still computed as a linear function of clean accuracy, as shown by the plot of the second row of Fig. 3. Hence, the clean accuracy can be seen as an accurate estimator of the robust accuracy when using random patches, similarly to what has been found by Taori et al. (2020). However, when we evaluate the robustness with our pre-optimized patches, the relation between robust and clean accuracy slightly diverges from a linear regression model, as the distance of the points from the interpolating line increases. Among the many reasons behind this effect, we focus on the ADV-ROBUST group, as it lays outside the confidence level, and towards the bisector of the plot. Intuitively, models that are located above the regression line can be considered more robust when compared with the others, since their robust accuracy is closer to their clean accuracy, i.e. closer to the bisector line. However, even if their robust training is aiding their performances against patch attacks, their robustness is not as evident as the one obtained when considering their original threat model. Our dataset can help by providing additional analysis of robustness against patch attacks to assess for a more general and complete evaluation. Lastly, we notice that the MORE-DATA group seems to present a similar effect by distantiating from the regression line, but with a much lower magnitude.

5. Related Work

We now discuss relevant work related to the optimization of adversarial patches, and to the proposal of similar benchmark datasets. Brown et al. (2017) introduced the first universal physical patch attack. In this work, we leverage the same model-ensemble attack to create adversarial patches robust to affine transformations and applicable to different source images to target different models. From that, we publish a dataset that favors fast robustness evaluation to patch attacks without requiring costly optimization steps. Also, previous work proposed datasets for benchmarking adversarial robustness. The APRICOT dataset, proposed by Braunegg et al. (2020), contains $1,000$ annotated photographs of printed adversarial patches targeting object detection systems, i.e. producing targeted detections. ImageNet-C and ImageNet-P, proposed by Hendrycks & Dietterich (2018), are two datasets proposed to benchmark neural network robustness to image corruptions and perturbations, respectively. Differently from these works, we propose a dataset that can be used to benchmark the robustness of image classifiers to adversarial patch attacks.

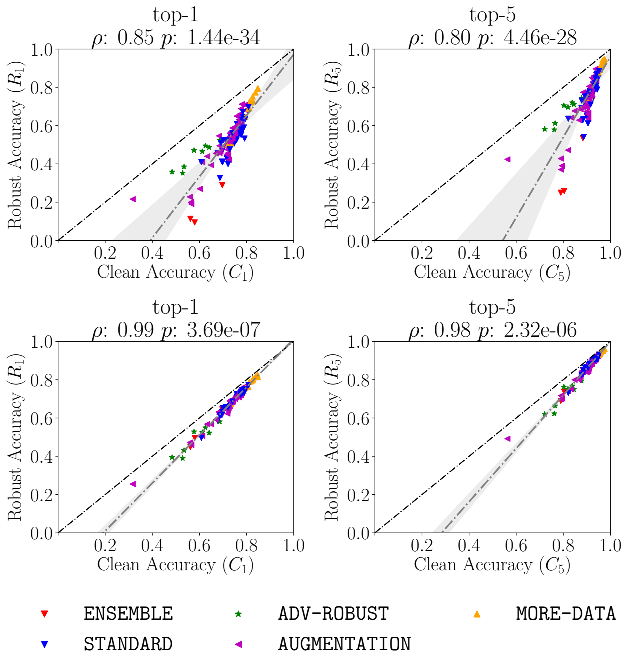

Figure 3. Clean vs robust accuracy for adversarial (*top*) and random (*bottom*) patches. The grey line and shaded area show a robust regression model fitted on the data along with the 95% confidence intervals. The results highlight the effectiveness of our pre-optimized strategy against choosing patches at random.

6. Conclusions, Limitations, and Future Work

We propose the ImageNet-Patch dataset, a collection of pre-optimized adversarial patches that can be used to compute an approximate-yet-fast robustness evaluation of machine-learning models against patch attacks. This dataset is constructed by optimizing squared blocks of contiguous pixels perturbed with affine transformations to mislead an ensemble of differentiable models, forcing the optimization algorithm to produce patches that can transfer across models, gaining cross-model effectiveness. Finally, these adversarial patches are attached to images sampled from the ImageNet dataset, composing a benchmark dataset of 50,000 images. While our methodology is efficient, it only provides an estimate of adversarial robustness, which can be computed more accurately by performing adversarial attacks directly against the target model. Hence, our analysis serves as a first preliminary robustness evaluation, to highlight the most promising defensive strategies. Moreover, we only release patches that target 10 different classes, and this number could be extended to target all the 1000 classes of the ImageNet dataset. We envision the use of our ImageNet-Patch dataset as a benchmark for machine-learning models. In addition, our methodology can generate adversarial patches for any kind of datasets of images, extending the achieved results on ImageNet to other data sources as well.

Acknowledgements

This work was partly supported by the PRIN 2017 project RexLearn, funded by the Italian Ministry of Education, University and Research (grant no. 2017TWNMH2); by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI.

References

Bai, T., Luo, J., and Zhao, J. Inconspicuous adversarial patches for fooling image recognition systems on mobile devices. *IEEE Internet of Things Journal*, 2021.

Benz, P., Zhang, C., Imtiaz, T., and Kweon, I. S. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F. (eds.), *Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part III*, volume 8190 of *LNCS*, pp. 387–402. Springer Berlin Heidelberg, 2013.

Braunegg, A., Chakraborty, A., Krumdick, M., Lape, N., Leary, S., Manville, K., Merkhofer, E., Strickhart, L., and Walmer, M. Apricot: A dataset of physical adversarial attacks on object detection. In *European Conference on Computer Vision*, pp. 35–50. Springer, 2020.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

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

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019a. URL `https://github.com/MadryLab/robustness`.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pp. 1802–1811, 2019b.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

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

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Howard, A. G., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Karmon, D., Zoran, D., and Goldberg, Y. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pp. 2507–2515. PMLR, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Lennon, M., Drenkow, N., and Burlina, P. Patch attack invariance: How sensitive are patch attacks to 3d pose? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–121, 2021.

Li, X. and Ji, S. Generative dynamic patch attack. *arXiv preprint arXiv:2111.04266*, 2021.

Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., and Tao, D. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1028–1035, 2019.

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

Mahajan, D. K., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/24357dd085d2c4b1a88a7e0692e60294-Abstract.html.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540. ACM, 2016.

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

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJx040EFvH.

Xiao, Z., Gao, X., Fu, C., Dong, Y., zhe Gao, W., Zhang, X., Zhou, J., and Zhu, J. Improving transferability of adversarial patches on face recognition with generative models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11840–11849, 2021.

Xie, C., Zhang, Z., Wang, J., Zhou, Y., Ren, Z., and Yuille, A. L. Improving transferability of adversarial examples with input diversity. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2725–2734, 2019.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification, 2019.

Yang, C., Kortylewski, A., Xie, C., Cao, Y., and Yuille, A. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pp. 681–698. Springer, 2020.

Ye, B., Yin, H., Yan, J., and Ge, W. Patch-based attack on traffic sign recognition. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 164–171. IEEE, 2021.

Zhang, R. Making convolutional networks shift-invariant again. In *ICML*, 2019.

A. Related Works on Patch Attacks

Aside from (Brown et al., 2017), there are other works that propose patch attacks, that are worth mentioning. The LaVAN attack, proposed by Karmon et al. (2018), attempts to achieve the same goal of Brown et al. by also reducing the patch size by placing it in regions of the target image where there are no other objects. The PS-GAN attack, proposed by Liu et al. (2019), addresses the problem of minimizing the perceptual sensitivity of the patches by enforcing visual fidelity while achieving the same misclassification objective. The DT-Patch attack, proposed by Benz et al. (2020), focuses on finding universal patches that only redirect the output of some given classes to different target labels, while retaining normal functioning of the model on the other classes. PatchAttack, proposed by Yang et al. (2020), leverages reinforcement learning for selecting the optimal patch position and texture to use for perturbing the input image for targeted or untargeted misclassification, in a black-box setting. The Inconspicuous Adversarial Patch Attack (IAPA), proposed by Bai et al. (2021), generates difficult-to-detect adversarial patches with one single image by using generators and discriminators. Lennon et al. (2021) analyze the robustness of adversarial patches and their invariance to 3D poses. Xiao et al. (2021) craft transferable patches using a generative model to fool black-box face recognition systems. They use the same transformations as (Xie et al., 2019), but unlike other attacks, they apply them to the input image with the patch attached, and not just on the patch. Ye et al. (2021) study the specific application of patch attacks on traffic sign recognition and use an ensemble of models to improve the attack success rate. The Generative Dynamic Patch Attack (GDPA), proposed by Li & Ji (2021), generates the patch pattern and location for each input image simultaneously, reducing the runtime of the attack and making it hence a good candidate to use for adversarial training.

We summarize in Table 1 these attacks, highlighting the main properties and comparing them with the attack we used to create the adversarial patches. In particular, in the *Cross-model* column we report the capability of an attack to be performed against multiple models (for black-box attacks we omit this information); in the *Transfer* column the proved transferability of patches, if reported in each work (thus it is still possible that an attack could produce transferable patches even if not tested on this setting); in *Targeted* and *Untargeted* columns the type of misclassification that patches can produce; in *Transformations* column the transformations applied to the patch during the optimization process (if any), which can increase the robustness of the patches with respect to them at test time.

Attack	Cross-model	Transfer	Targeted	Untargeted	Transformations
Sharif et al. (2016)	✗	✗	✓	✓	`rot`
Brown et al. (2017)	✓	✓	✓	✗	`loc`, `scl`, `rot`
LaVAN (Karmon et al., 2018)	✗	✗	✓	✗	`loc`
PS-GAN (Liu et al., 2019)	✗	✓	✗	✓	`loc`
DT-Patch (Benz et al., 2020)	✗	✗	✓	✗	✗
PatchAttack (Yang et al., 2020)	-	✓	✓	✓	`loc`, `scl`
IAPA (Bai et al., 2021)	✗	✓	✓	✓	✗
Lennon et al. (2021)	✗	✓	✓	✗	`loc`, `scl`, `rot`
Xiao et al. (2021)	-	✓	✓	✓	`various`
Ye et al. (2021)	✓	✓	✓	✗	`loc`, `scl`, `rot`
GDPA (Li & Ji, 2021)	✗	✗	✓	✓	`loc`
Ours	✓	✓	✓	✓	`loc`, `rot`

Table 1. Patch attacks, compared based on their main features. `loc` refers to the location of the patch in the image, `rot` refers to rotation, `scl` refers to scale variations, `various` include several image transformations (see (Xiao et al., 2021) for more details).

B. Additional Results

We briefly summarize here the results of our analysis, based on our ImageNet-Patch dataset to benchmark machine-learning models. We observe that data augmentation techniques do not generally improve robustness to adversarial patches. Moreover, we argue that real progress in robustness should be observed as a general property against different adversarial attacks, and not only against one specific perturbation model with a given budget (e.g., ℓ_∞-norm perturbations with maximum size of $8/255$). We are not claiming that work done on defenses for adversarial attacks so far is useless. Conversely, there has been great work and progress in this area, but it seems now that defenses are becoming too specific to current benchmarks and fail to generalize against slightly-different perturbation models. To overcome this issue, we suggest to test the proposed defenses on a wider set of robustness benchmarks, rather than over-specializing them on a specific scenario, and we believe that our ImageNet-Patch benchmark dataset provides a useful contribution in this direction.

Finally, we report detailed results for 15 models taken from the different groups in Table 2. In particular, we consider the three models used for the ensemble, AlexNet (Krizhevsky et al., 2012), ResNet18 (He et al., 2016) and SqueezeNet (Iandola

		top-1			**top-5**			**top-10**		
	Model	C_1	R_1	S_1	C_5	R_5	S_5	C_{10}	R_{10}	S_{10}
ENSEMBLE	AlexNet (Krizhevsky et al., 2012)	0.562	0.113	0.256	0.789	0.250	0.504	0.849	0.327	0.613
	ResNet18 (He et al., 2016)	0.697	0.289	0.431	0.883	0.535	0.739	0.923	0.641	0.839
	SqueezeNet (Iandola et al., 2016)	0.580	0.094	0.610	0.804	0.259	0.865	0.865	0.355	0.926
STANDARD	GoogLeNet (Szegedy et al., 2015)	0.697	0.469	0.090	0.895	0.702	0.326	0.932	0.778	0.482
	MobileNet (Howard et al., 2019)	0.737	0.541	0.017	0.910	0.764	0.083	0.945	0.826	0.141
	Inception v3 (Szegedy et al., 2016)	0.696	0.412	0.106	0.883	0.628	0.317	0.921	0.703	0.426
ADV-ROBUST	Engstrom et al. (Engstrom et al., 2019a)	0.625	0.495	0.005	0.838	0.720	0.026	0.887	0.789	0.051
	Salman et al. (Salman et al., 2020)	0.641	0.486	0.003	0.845	0.711	0.017	0.894	0.780	0.034
	Wong et al. (Wong et al., 2020)	0.535	0.385	0.003	0.765	0.612	0.020	0.833	0.695	0.039
AUGM.	Zhang et al. (Zhang, 2019)	0.566	0.191	0.093	0.790	0.370	0.241	0.848	0.459	0.330
	Hendrycks et al (Hendrycks et al., 2021)	0.769	0.632	0.020	0.929	0.842	0.104	0.956	0.890	0.181
	Engstrom et al (Engstrom et al., 2019b)	0.684	0.495	0.036	0.886	0.729	0.148	0.928	0.800	0.232
MORE-DATA	Yalniz et al. (Yalniz et al., 2019)-a	0.813	0.726	0.029	0.958	0.911	0.217	0.976	0.943	0.328
	Yalniz et al. (Yalniz et al., 2019)-b	0.838	0.774	0.008	0.970	0.936	0.073	0.984	0.962	0.125
	Mahajan et al. (Mahajan et al., 2018)	0.735	0.507	0.104	0.914	0.748	0.357	0.949	0.826	0.491

Table 2. Evaluation of the ImageNet-Patch dataset using the chosen metrics. On the rows, we list 15 models used for testing, divided into the isolated groups. On the columns, we detail the clean accuracy, the robust accuracy and the success rate of the adversarial patch, repeated for top-1,5, and 10 accuracy.

et al., 2016), as the first group, ENSEMBLE. We consider for the second group, STANDARD, 3 standard-trained models, that are GoogLeNet (Szegedy et al., 2015), MobileNet (Howard et al., 2019) and Inception v3 (Szegedy et al., 2016), available in PyTorch Torchvision. We then consider 3 robust-trained models as the ADV-ROBUST available on RobustBench, specifically a ResNet-50 proposed by Salman et al. (2020), a ResNet-50 proposed by Engstrom et al. (2019a) and a ResNet-50 proposed by Wong et al. (2020). We also additionally consider a set of 6 models from the ImageNet Testbed repository[5] proposed by Taori et al. (2020), to analyze the effects of non-adversarial augmentation techniques and of training on bigger datasets. We select 3 models specifically trained for being robust to common image perturbations and corruptions, namely the models proposed by Zhang (2019), Hendrycks et al. (2021), and Engstrom et al. (2019b), that we group as AUGMENTATION group. We further select other 3 models, namely two of the ones proposed by Yalniz et al. (2019) and one proposed by Mahajan et al. (2018), that have been trained on datasets that utilize substantially more training data than the standard ImageNet training set. We group these last models as the MORE-DATA group.

C. Visualization of Images with the Patches

Finally, we provide in Fig. 1 some example of images with the applied patches, classified by SqueezeNet (Iandola et al., 2016). Note that for some images, e.g. the otter depicted in the fourth column, the patches seem to be less effective. For other images, as for the guenon in the first column, the patches work well even when not superimposed directly on the subject of the image.

[5] https://github.com/modestyachts/imagenet-testbed

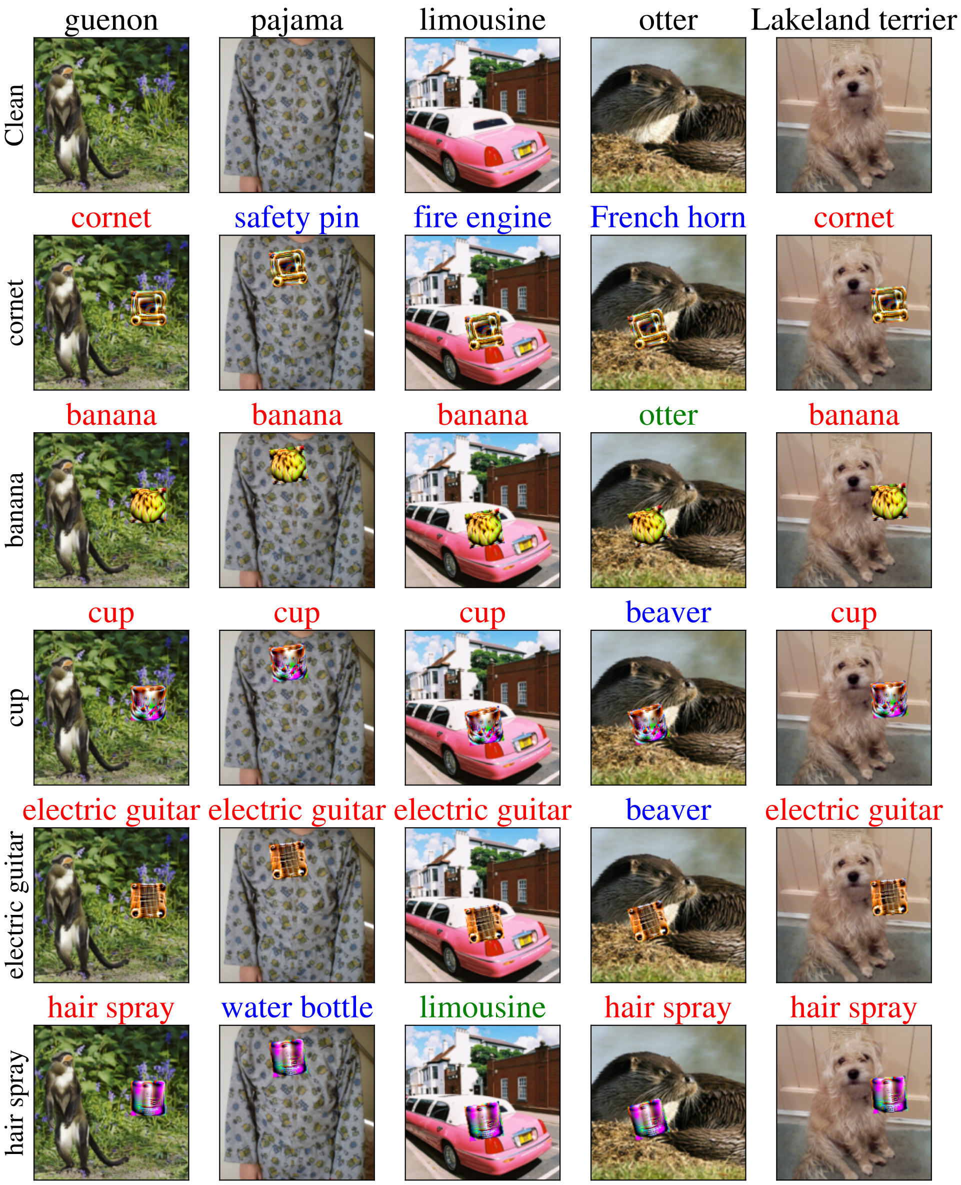

Figure 4. A batch of clean images initially predicted correctly by a SqueezeNet (Iandola et al., 2016) model, and its perturbation with 5 different adversarial patches. Each row contains the original image with a different patch, whose target is displayed in the left. The predictions are shown on top of each of the samples, in *green* for correct prediction, *blue* for misclassification, and in *red* for a prediction that ends up in the target class of the attack.