Image recognition time for humans predicts adversarial vulnerability for models

David Mayo* CSAIL & CBMM MIT

Jesse Cummings* CSAIL & CBMM MIT

Xinyu Lin* CSAIL & CBMM MIT

Boris Katz CSAIL & CBMM MIT

Andrei Barbu
CSAIL & CBMM
MIT

Abstract

The success of adversarial attacks and the performance tradeoffs made by adversarial defense methods have both traditionally been evaluated on image test sets constructed from a randomly sampled held out portion of a training set. Mayo 2022 et al. [1] measured the difficulty of the ImageNet and ObjectNet test sets by measuring the minimum viewing time required for an object to be recognized on average by a human, finding that these test sets are heavily skewed towards containing mostly easy, quickly recognized images. While difficult images that require longer viewing times to be recognized are uncommon in test sets, they are both common and critically important to the real world performance of vision models. In this work, we investigated the relationship between adversarial robustness and viewing time difficulty. Measuring the AUC of accuracy vs attack strength (epsilon), we find that easy, quickly recognized, images are more robust to adversarial attacks than difficult images, which require several seconds of viewing time to recognize. Additionally, adversarial defense methods improve models robustness to adversarial attacks on easy images significantly more than on hard images. We propose that the distribution of image difficulties should be carefully considered and controlled for when measuring both the effectiveness of adversarial attacks and when analyzing the clean accuracy vs robustness tradeoff made by adversarial defense methods.

1 Introduction

The adversarial robustness of models today is typically evaluated by performing an adversarial attack across a distribution of images and then computing the effectiveness of the attack on average. Unfortunately, current image test sets are highly skewed towards containing more easy easy examples than hard ones [1]. This biases current robustness measurements to reflect model robustness to easy examples rather than long-tailed difficult examples. An example of this in terms of accuracy of a defense trained model compared to a standard model can be found in fig. 2.

Misalignment between the difficulty distributions of current benchmarks and real world environments poses significant safety concerns for the deployment of defense trained models in the real world. Practitioners may be making accuracy/robustness tradeoffs for the deployment of models based on

*Equal contribution. Website `https://objectnet.dev/flash` Corresponding author `dmayo2@mit.edu`

Preprint. Under review.

benchmarks that underestimate the long-tailed, difficult image accuracy sacrifice made by robustness trained models, while also overestimating the robustness benefits.

In this work we use the difficulty metric introduced in [1], the minimum viewing time required for a human to correctly recognize and image on average. This metric is an objective measure that serves as a proxy for how difficult it is for a particular image to be recognized. Minimum viewing time was measured by performing a psychophysics experiment in which ImageNet and ObjectNet images were show to participants for one of 17ms, 50ms, 150ms, and 10 seconds followed by a 1 of 50 classification forced choice response task. For each presentation time each image was seen by 7 participants for a total of 133,588 judgments across 5,000 images. In this study we focus on only ImageNet images. As in [1] we consider images whose objects can be recognized in 17ms as "easy" and images which require seconds of viewing time to be "hard".

We perform two sets of experiments exploring the relationship between minimum viewing time difficulty and 1) adversarial attack method and strength, and 2) adversarial defense strength.

Main contributions

1. We demonstrate that on average larger epsilon values are needed to fool quickly recognized easy images compared to hard images (see fig. 1).
2. We find that the drop in clean (unattacked) accuracy between defense trained models and standard models is disproportionately caused by a decrease in accuracy on hard images (see fig. 2).
3. For both standard models and adversarially robust models, easy images have a larger area under the attack failure rate vs. epsilon curve than hard images (see fig. 3).
4. Compared to standard models, robustness defense models improve their robustness to attacks on easy images significantly more than their robustness to attacks on hard images (see fig. 4).

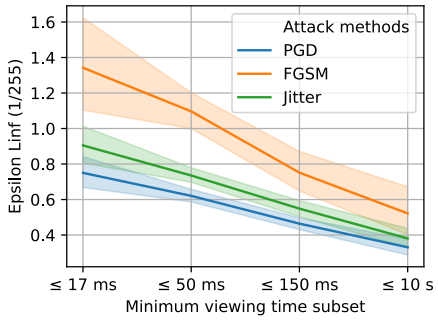


Figure 1: The average L_∞ epsilon (attack strength) required to perturb an ImageNet validation set image to be classified incorrectly decreases with increasing human recognition difficulty as measured by the minimum viewing time metric from [1]. All attacks were performed against a standard ImageNet trained ResNet-50 on ImageNet validation set images that were classified correctly when not attacked. Three attack methods, PGD, FGSM, and jitter, are shown. Many additional attack methods were tested and found to follow the same trend, a full table can be found in the appendix.

2 Experiments

2.1 Experiment 1: Evaluating the robustness of a vanilla ResNet-50 to several adversarial attack methods across image viewing time difficulty levels

We started with the 5,000 images from [1], which are images with objected-centered crops evenly split between the ImageNet [2] validation set and ObjectNet [3], spanning 50 object classes. For the experiments in this paper we used only the 2,500 ImageNet images. These images were then filtered down to 1,214 images that were classified correctly by a vanilla ResNet-50 model trained on ImageNet from the torchvision repository. Using the torchattacks framework [4] we performed 14 different attack methods, sweeping 140 epsilon values per attack method (from 0 to 0.005 at increments of 0.0001 and from 0.005 to 0.05 at increments of 0.0005). From the results of these attack strength sweeps we determined the minimum attack strength required to fool the ResNet-50

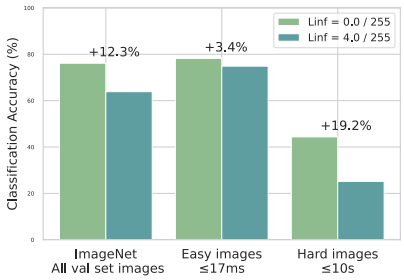
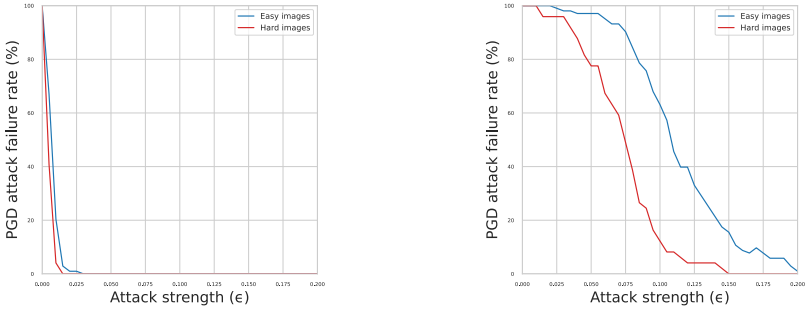


Figure 2: Comparison of the clean (unattacked) accuracy of a standard ResNet-50 and an adversarial defense trained ResNet-50 on 3 sets of images: the full ImageNet validation set, human easy images (ImageNet validation set images classified correctly with less than 17ms of viewing time on average), and human hard images (ImageNet validation set images classified correctly on average only after 10 seconds of viewing time).

for each individual image per attack. We then averaged these minimum epsilon values grouped by difficulty subset to measure the relationship between attack strength and viewing time difficulty. The results of this experiment for PGD, FGSM, and jitter attacks can be found in fig. 1. Results for all attack methods can be found in the appendix.



a) Standard ResNet-50 b) Adversarially trained ResNet-50, Linf = 4.0/255

Figure 3: We compare the PGD attack strength vs PGD attack failure rate (how frequently the attack fails to fool the model) for both easy and hard images. For both standard and defense trained models, easy images have a higher attack failure rate at any given attack strength. Defense trained models increase adversarial robustness overall, but they increase robustness to attacks on easy images significantly more than hard images.

2.2 Experiment 2: Evaluating the robustness of a defense trained model with increasing defense strength on easy and hard images

Using the PGD adversarial attack method and the same attack-strength sweep from experiment 1, we attack ResNet-50 models that were adversarially trained with several different levels of defense strength using the Salman et al. 2020 method [5] . We then investigate how attack failure rate—that is, how frequently the attack fails to fool the model—varies as the strength of the attack increases. We analyze this by plotting attack failure rate vs epsilon curves (fig. 3) and computing the area under the curve (fig. 4) similarly to [6]. Models that maintain performance over increasing epsilon will have higher AUC than the standard ResNet-50 model.

3 Results

The results of experiment 1 are summarized in fig. 1. Across all attack types, images that could be recognized by humans given less than 17ms of viewing time required a stronger perturbation to fool

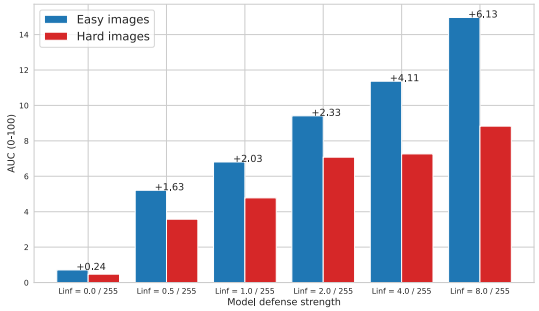


Figure 4: We find that with increasing levels of adversarial defense, robustness to PGD attacks—as measured by area under the curve AUC of line plots like fig. 3 (with a max attack strength cutoff of 0.5)—increases more when models are evaluated on hard images compared to easy images.

the ResNet-50 than images requiring longer viewing times to be recognized such as 10 seconds. This shows a clear relationship between what images humans find difficult and how vulnerable an image is to adversarial attack.

The results of experiment 2 are summarized in fig. 3 and fig. 4. In fig. 3, both standard ResNet-50 and the adversarially trained ResNet-50 achieve a higher accuracy under PGD attack on easy images compared to hard images for every epsilon value greater than zero. fig. 3 also shows that a defense trained ResNet-50 trained with $L_\infty = 4.0/255$ in the adversarial defense presented in [5] exhibits the same trend, but the easy image accuracy curve is shifted significantly further to the right, towards higher epsilon values, than the hard image curve. This means that while robustness training is improving adversarial robustness on both subsets of images, it is significantly more effective at improving robustness on easy images compared to hard images.

We summarize these findings by computing the area under the attack failure rate vs epsilon curve (AUC). fig. 4 compares this metric between model accuracy on easy and hard image subsets for an adversarial defense trained model [5] with increasing defense strengths.

4 Related Work

Ever since the publication of some of the earliest work on adversarial examples [7], much research has been conducted in improving both adversarial attacks [8, 9, 10, 11, 12, 13, 14, 9, 15, 16], and defenses against them [5, 13, 17]—although, at times with great difficulty [18]. The discovery of adversarial examples has naturally spawned investigation into the learning dynamics and the stimuli that give rise to the phenomenon with some researchers positing that adversarial examples are, indeed, features rather than bugs [19]. Along with understanding adversarial robustness in models, some researchers have investigated how the phenomenon relates to humans. Guo et al. report that biological neurons are susceptible to adversarial perturbations [20]. Modeling human-like foveated vision in models is shown to improve adversarial robustness [21] suggesting humans' robustness may lie in perceptual mechanisms. Tsipras et al. show that robust classifiers learn features that are qualitatively better aligned with human perception [22]. With particular relevance to our work, humans have been shown both to be able to decipher adversarial examples [23] and be fooled by them [24], thus suggesting that there may exist a link between human psychophysics and stimuli that fool models.

5 Conclusion

Evaluation methods are a critical component for both guiding future model development and making tradeoffs to mitigate risks during real world model deployments. We propose that the distribution of image difficulties should be carefully considered and controlled for when measuring both the effectiveness of adversarial attacks and when analyzing the clean accuracy vs robustness tradeoff made by adversarial defense methods.

Acknowledgments and Disclosure of Funding

We would like to thank Ko Kar, Jim DiCarlo, Dan Yamins, Martin Schrimpf, Gamal Elsayed, and Arturo Deza for their helpful feedback and discussion about our experiments. We would also like to thank David Lu for his contributions to our early experiments and directions.

This work was supported by the Center for Brains, Minds and Machines, NSF STC award 1231216, the NSF award 2124052, the MIT CSAIL Systems that Learn Initiative, the MIT CSAIL Machine Learning Applications Initiative, the CBMM-Siemens Graduate Fellowship, the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the United States Air Force Research Laboratory and United States Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0014, and the Office of Naval Research under award number N00014-20-1-2589 and award number N00014-20-1-2643. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Anonymous authors. How hard are computer vision datasets? calibrating dataset difficulty to viewing time, 2022. In review at ICLR. https://openreview.net/forum?id=zA7hVj3rR19.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015.

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[5] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020.

[6] Owen Kunhardt, Arturo Deza, and Tomaso Poggio. The effects of image distribution and task on adversarial robustness. *arXiv preprint arXiv:2102.10534*, 2021.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[8] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[9] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Making stochastic neural networks from deterministic ones, 2017.

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[11] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

[12] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. 2019.

[13] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-BNN: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, 2020.

[15] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern Eskofier. Exploring misclassifications of robust neural networks to enhance adversarial attacks, 2021.

[16] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

[17] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

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

[19] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[20] Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James Dicarlo. Adversarially trained neural representations are already as robust as biological neural representations. In *International Conference on Machine Learning*, pages 8072–8081. PMLR, 2022.

[21] Jonathan M Gant, Andrzej Banburski, and Arturo Deza. Evaluating the adversarial robustness of a foveated texture transform module in a cnn. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

[22] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[23] Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature communications*, 10(1):1–9, 2019.

[24] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

[25] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[26] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

A Appendix

Attack name	$\leq$ 17 ms	$\leq$ 50 ms	$\leq$ 150 ms	$\leq$ 10 s
PGD [10]	0.002586	0.002338	0.001863	0.001376
UPGD [4]	0.002165	0.001941	0.001520	0.001109
APGD [8]	0.002070	0.001860	0.001461	0.001069
TPGD [25]	0.004444	0.003494	0.002862	0.001907
EOTPGD [13]	0.003989	0.003438	0.002453	0.001729
FGSM [7]	0.005647	0.004593	0.003191	0.002073
BIM [9]	0.005016	0.004627	0.004055	0.003680
RFGSM [26]	0.011581	0.010714	0.007998	0.005990
FFGSM [11]	0.005266	0.004498	0.003102	0.002032
MIFGSM [27]	0.002450	0.002152	0.001663	0.001202
DIFGSM [12]	0.003377	0.002991	0.002307	0.001715
TIFGSM [28]	0.007059	0.006373	0.004973	0.003610
Square [14]	0.015541	0.013974	0.010610	0.007937
Jitter [15]	0.003101	0.002799	0.002170	0.001558

Figure 5: Full table of 14 adversarial attack methods average epsilon required to fool a standard ResNet-50 per human viewing time difficulty subset.