An Extended Study of Human-like Behavior under Adversarial Training

Paul Gavrikov[1*] Janis Keuper[1,2*] Margret Keuper[3,4]
[1]IMLA, Offenburg University, [2]Fraunhofer ITWM, [3]University of Siegen
[4]Max Planck Institute for Informatics, Saarland Informatics Campus
{paul.gavrikov,janis.keuper}@hs-offenburg.de, keuper@mpi-inf.mpg.de

Abstract

Neural networks have a number of shortcomings. Amongst the severest ones is the sensitivity to distribution shifts which allows models to be easily fooled into wrong predictions by small perturbations to inputs that are often imperceivable to humans and do not have to carry semantic meaning. Adversarial training poses a partial solution to address this issue by training models on worst-case perturbations. Yet, recent work has also pointed out that the reasoning in neural networks is different from humans. Humans identify objects by shape, while neural nets mainly employ texture cues. Exemplarily, a model trained on photographs will likely fail to generalize to datasets containing sketches. Interestingly, it was also shown that adversarial training seems to favorably increase the shift toward shape bias. In this work, we revisit this observation and provide an extensive analysis of this effect on various architectures, the common ℓ_2- and ℓ_∞-training, and Transformer-based models. Further, we provide a possible explanation for this phenomenon from a frequency perspective.

1. Introduction

ImageNet [1] trained convolutional neural networks (CNNs) have been shown to predominantly classify images by the observed texture, whereas, humans rather tend to consider global object shapes as the predominant cues [2]. In this context, *Geirhos et al*. provided an initial analysis of robust models and provided initial evidence, that the initial texture bias in CNNs is shifted towards shape-based decisions under adversarial training (AT) [3]. However, the authors have limited their analysis to a ResNet-50 trained on ImageNet using AT against an ℓ_2-bound adversary. To allow for a more conclusive evaluation, we expand their analysis to the more common ℓ_∞-setting for AT and analyze additional CNNs like ResNet-18 and Wide-ResNet-50-2, as well as Transformers (XCiT-S/M/L12), which are known to behave differently from CNNs regarding their inductive bias. In our study, we evaluate models trained on clean data as well as under AT with different norms and budgets with respect to their generalization ability to out-of-domain (OOD) data [3, 4], with special emphasis on the shape-texture *cue-conflict*, that has been used as a measure of the misalignment between human and neural network based image classification. In this context, we provide an extensive evaluation and discussion of the different behavior of CNN and Transformer models.

Further, we analyze the generalization ability of adversarially-trained networks from a frequency perspective. Specifically, we investigate the frequency spectra of different OOD image categories and provide possible explanations for the following two questions: (i) Why does adversarial training lead to an accuracy decay on some OOD datasets? and (ii) Why is the *cue-conflict* between shape and texture affected by AT?

We summarize our key findings as follows:

- Training against ℓ_∞-bound adversaries generally results in similar trends regarding human-like behavior with respect to the shape-texture bias as ℓ_2-bound adversarial training. However, ℓ_∞-robust models perform better on high-frequency, and worse on low-frequency data.
- Observations made by prior work on ℓ_2-bound ResNet-50 scale to other CNNs and Transformers relative to parameter sizes.
- Although Transformers also experience a drop in OOD performance after adversarial training, they perform better in OOD generalization and are more human-like than robust CNNs, and even outperform humans on many benchmarks.
- From the analysis of the images frequency spectra, we provide a possible explanation of why adversarial training can lead to a decay of model accuracy on OOD data.

*Funded by the Ministry for Science, Research and Arts, Baden-Wuerttemberg, Grant 32-7545.20/45/1 (Q-AMeLiA).

- We also provide a possible explanation of why adversarial training reduces texture bias and increases shape bias.

2. Related Work

This work focuses on the intersection between adversarial robustness and "human-like" behavior which we briefly sketch in this section.

Adversarial robustness. Neural networks have a tendency to overfit the training data distribution, which makes them fail to generalize beyond it. As a result, their predictions are often highly sensitive to small changes in input [5,6], even if those changes are imperceptible and meaningless to humans. This phenomenon can be formally described as an adversarial attack, where the goal is to find an additive perturbation to the input sample that maximizes the loss function [7–10]. To constraint attacks, perturbations are only sought within a specified radius ϵ (budget) of the original input. The radius is typically bounded by the ℓ_2 or ℓ_∞-norm.

Adversarial attacks can be found in both white-box [8–10] and black-box [11–14] settings, with gradient-based attacks being particularly effective. Models that are not trained with adversarial defenses are typically only robust to low budgets attacks, if at all. Adversarial training (AT) [8] is a solution to this problem, as it trains the model on worst-case perturbations found during training, effectively making out-of-domain attacks become in-domain samples. However, this approach can result in overfitting to attacks used during training. Early stopping [15] and the addition of external (synthetic) data [16–18] have been proposed as effective solutions to address this problem.

However, adversarial robustness does not necessarily correlate with improved generalization and can even hurt it [19]. Supposedly again due to overfitting of training data, *e.g.* models can still be susceptible to adverse weather conditions, image artifacts due to image compression, changes in lighting, etc. [20, 21].

Measuring "human-like" behavior. *Geirhos et al.* propose to measure "human-like" reasoning via out-of-distribution (OOD) generalization to datasets and consistency in predictions with humans [3].

Regarding OOD, they propose to benchmark against a set of 12 ImageNet modification datasets [22] at various intensities/conditions. At first glance, this may sound familiar to ImageNet-C [20], but benchmarks a different set of modifications: *(the absence of) colour*, *contrast (changes)*, *eidolon I/II/III*, *false-colour*, *high/low-pass (frequency filtering)*, *phase-scrambling*, *power-equalisation*, *rotation*, *uniform-noise*. Additionally, they propose to benchmark

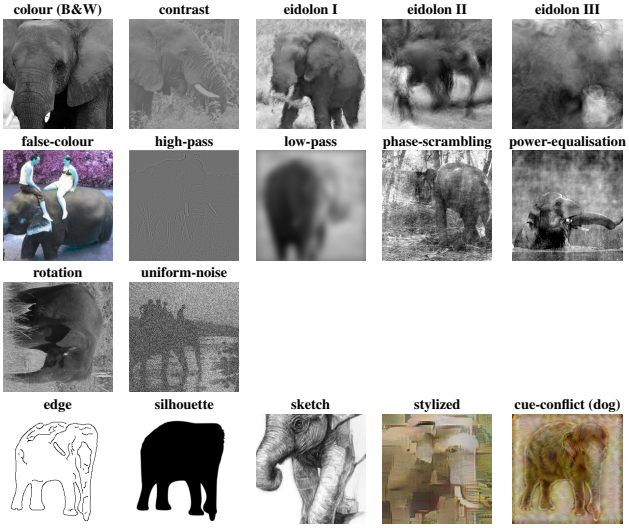

Figure 1. OOD examples from [2–4, 22] for the ImageNet class "elephant".

against a set of five OOD datasets aiming to identify the shape-texture-bias [2]: *stylized*, *edge*, *silhouette*, *texture/shape cue-conflict*, and *sketch* (the latter provided by [4]). All datasets contain samples that belong to 16 ImageNet classes and are therefore classifiable by ImageNet models. For all datasets, the authors include a baseline obtained in lab settings over 4-10 human annotators. The *cue-conflict* dataset is of particular interest, as neural networks are not only prone to overfit but - at least in the vision domain - they also tend to compute predictions based on details such as the texture of images rather than shapes, which does not align with human vision [2]. For example, an image of an elephant with an overlaid lion texture will most likely result in a prediction as "lion", while most humans would predict "elephant" as the true label when given the choice between both. It is worth noting that the authors also mention that ImageNet can be largely accurately classified solely based on texture. As such, ImageNet performance is insufficient as an indicator of "human-like" decision making, and *Geirhos et al.* propose to additionally report the *cue-conflict* score to quantify this phenomenon. Examples of all datasets are shown in Fig. 1.

As an additional metric to accuracy, [23] propose to evaluate the agreement in predictions. In particular, they analyze false predictions (*error consistency*) as well as the intersection rate of predictions where both humans and models have made a correct prediction (*observed consistency*).

The authors maintain a leaderboard of the most "human-like" models, which is currently dominated by Transformers such as *ViT* [24, 25] and *CLIP* [26], or large convolutional neural networks [27, 28] - all being pre-trained on massive datasets.

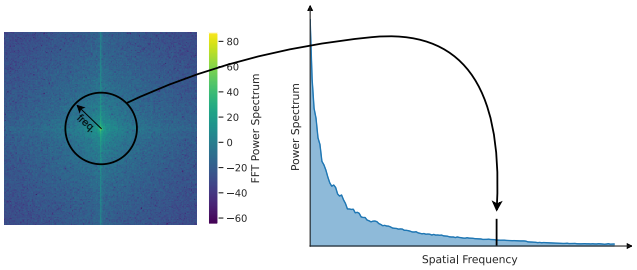

Figure 2. Visualization of how we obtain the spectrum plots. Each frequency measurement in the spectrum plot corresponds to the integral over the FFT power spectrum (frequency increases from the center to outer edges) up to that particular frequency.

3. Method

To study the likeliness to human-like behavior of adversarially-trained models in greater detail, we use publicly available checkpoints and perform an analysis according to the setting proposed in [3]. We analyze pre-trained *ResNet-18, ResNet-50, WideResNet-50-2* models trained against ℓ_∞-bound adversaries with $\epsilon \in \{0.5/255, 1/255, 2/255, 4/255, 8/255\}$, and against ℓ_2-bound adversaries with $\epsilon \in \{0.01, 0.03, 0.05, 0.1, 0.25, 0.5, 1, 3, 5\}$, and clean baselines (all provided by [29]). Further, we analyze *XCiT-S/M/L12* Transformer models trained against ℓ_∞-bound adversaries with $\epsilon = 4/255$ provided by [30] and a clean *XCiT-S*[1] baseline provided by [31]. Lastly, to better understand the differences between CNNs and Transformers, we also analyze a clean *ConvMixer-768-32* [32] checkpoint, again obtained from [31]. All models were trained on ImageNet [1] without any additional pre-training.

For all models, we measure the accuracy of all datasets by reporting the mean overall conditions in the dataset where average human performance was above 20% accuracy. Lastly, we determine the *observed* and *error consistency* against human annotators again as a mean over all datasets and conditions. As there are multiple annotators per dataset, we calculate consistencies against each annotator and report the mean.

We first provide a more extensive evaluation of models that have been trained using ℓ_2-AT. Then, we provide insights on how models trained with ℓ_2-AT behave compared to models that are trained using ℓ_∞-AT. Comparing these two training types is not straightforward, due to the different types of perturbations they cause. As the ℓ_2-norm penalizes the euclidean distance, perturbations can locally be more severe than under ℓ_∞. Yet, if the perturbation magnitude increases the area of perturbations has to decrease under ℓ_2-norm, while attacks under the ℓ_∞-norm can add perturbations to the entire image without constraints except for the magnitude. Thus, there are multiple options for choosing comparable budgets between the two norms. We choose a straightforward way and select budgets for both norms that approximately result in the same clean accuracy shown in Tab. 1. As we have more checkpoints for ℓ_2-AT training, we only use a subset of those in the following analysis.

[1]Clean pre-trained *XCiT-M/L12* with the same configuration were not available.

Table 1. Our chosen ϵ budgets for comparisons between ℓ_2- and ℓ_∞-bound training.

ℓ_2	0.1	1	3	5
ℓ_∞	0.5/255	1/255	4/255	8/255

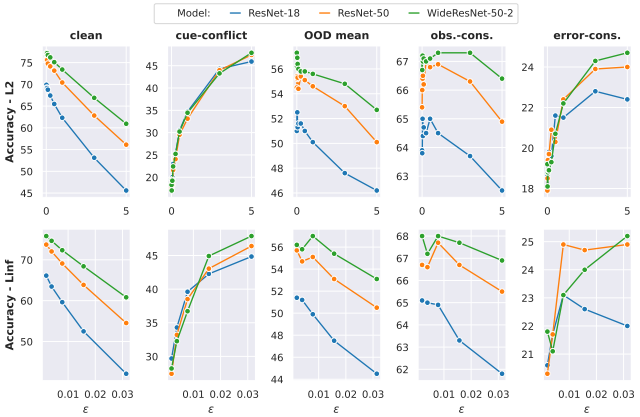

Figure 3. Performance of ℓ_2 vs. ℓ_∞-AT-trained *ResNet-18, ResNet-50, WideResNet-50-2* on clean data, texture/shape bias cue-conflict datasets, the average mean of all OOD datasets, and observed/error consistency compared to humans under increased training attack budget ϵ.

Next, we compare the behavior of CNN and Transformer models under these training settings. Based on these experiments, we then discuss whether AT is an effective tool to induce a more human-like behavior in trained models. Finally, we impose a frequency perspective on OOD performance and shape bias under AT. To back this analysis, we plot the frequency distribution for each OOD dataset, and clean ImageNet validation samples belonging to the same classes. Then we compare each OOD distribution to the clean distribution to understand where shifts in the frequency distribution are located. We obtain the frequency distribution plots as introduced in [33]: we compute the log-scaled FFT power spectrum and compute the radial integral under increasing frequency resulting in a frequency power distribution (Fig. 2). For comparability, we scale the resulting distributions by their integral.

4. Results

Width and depth of ℓ_2 CNNs. First, we want to evaluate the effect of ℓ_2-training on CNN architectures: *Wide-*

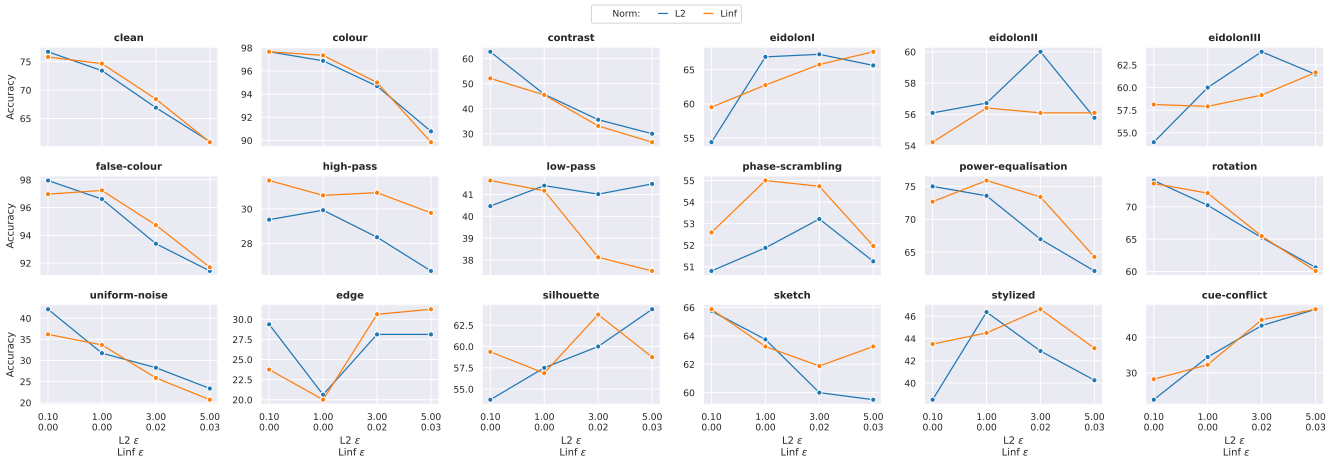

Figure 4. Comparison of performance on OOD datasets between robust *WideResNet-50-2* trained against ℓ_2- (upper ϵ values) and ℓ_∞-bound (lower ϵ values) adversaries under increasing training budget ϵ. ϵ are selected in a way that clean accuracy approximately matches between the norms.

ResNet-50-2, *ResNet-50* and *ResNet-18* (Fig. 3, top row), which allows investigating the effect of ℓ_2-AT depending on model depth and width.

We observe that increasing both, width and depth improves clean performance, and performance on all OOD datasets, as well as the observed and error consistency. As such, *Wide-ResNet-50-2* performed best on clean performance, OOD mean, and observed/error consistency. It is also worth noting, that switching from *ResNet-50* to *WideResNet-50-2* has a smaller impact on performance than switching from *ResNet-18* to *ResNet-50*. Also, we observe that in some cases *ResNet-18* shows opposite trends with respect to training budget than 50-layer deep *ResNets*, *e.g.* for *power-equalisation*. Still, *ResNet-18* performs best on the *edge* dataset for large training budgets (not shown due to space limitations). Overall, this suggests, that increasing parameterization of ℓ_2-bound adversarially-trained models correlates with an increase in human-like behavior.

ℓ_2- vs. ℓ_∞-bound Adversarial Training. Next, we compare how ℓ_2-AT relates to ℓ_∞-AT with respect to human-like reasoning. Exemplarily, we analyze the trend under comparable budgets of a *WideResNet-50-2* (Fig. 4). We observe that on some datasets there is barely any perceivable difference as the budget increases (*colour, contrast, false-colour, uniform-noise, rotation, cue-conflict*), but there are cases where one norm or the other clearly performs better. ℓ_∞-robust models seem to be more robust against *high-pass*, *phase-scrambling*, and *power-equalisation*. On the other hand, ℓ_2-robust models appear to perform better on *low-pass*, and *eidelonII/III*. Lastly, there are also some inconclusive settings where one or the other performs better depending on the budget (*silhouette, eidolon I, stylized*). Besides *cue-conflict*, none of the OOD categories clearly benefit from AT for WideResNet-50-2. These observations only partly transfer to other CNN architectures in Tab. 3. In general, Fig. 3 (bottom) shows a similar trend for ℓ_2 and ℓ_∞-AT, and all results support the finding that the *cue-conflict* score increases consistently under both types of AT, *i.e.* the behavior becomes more human-like towards shape bias in both cases. Therefore, we conclude that the more commonly used ℓ_∞-AT is equally effective in inducing human-like behavior in CNNs, with respect to *cue-conflict*, and *consistency*.

Table 2. Comparison between parameters of analyzed models.

Model	Inductive Bias	Parameters
ResNet-18	CNN	10.4 M
ResNet-50	CNN	25.6 M
WideResNet-50-2	CNN	68.9 M
ConvMixer-768-32	Hybrid	21.2 M
XCiT-S12	Transformer	26.3 M
XCiT-M12	Transformer	46.4 M
XCiT-L12	Transformer	103.8 M

CNNs vs. Transformers. Finally, we expand our analysis to Transformer architectures (*XCiT*) for which we only report clean and ℓ_∞-training performance. On clean training, even the smallest Transformer (*XCiT-S12*) which has a comparable number of parameters to *ResNet-50* (Tab. 2), performs significantly better than the largest CNN. Contrary to CNNs it is also able to surpass human performance on *eidolon II/III*, *high-pass* (with an impressive improvement of approx. 35% above CNNs), *phase-scrambling, power-equalisation, uniform-noise*, and *stylized*. Under AT, we largely see the same shift as with CNNs with one exception: while AT improves *stylized* performance of CNNs, it

Table 3. Results in [%] of generalization performance and consistency with human predictions/errors. For robust models we only report $\ell_2, \epsilon = 3$ (ℓ_2) and $\ell_\infty, \epsilon = 4/255$ (ℓ_∞) for brevity. Models without adversarial training are highlighted in gray. **Bold** values indicate the best performance amongst all models.

Model	Clean	Out-of-distribution Performance																	Mean	Consistency	
		colour	contrast	eidolon I	eidolon II	eidolon III	false colour	high pass	low pass	phase scr.	power equal.	rotat.	uniform noise	edge	silh.	sketch	styliz.	cue conflict		correct	error
R18	69.79	95.47	71.88	47.50	51.88	49.38	93.39	32.66	37.73	48.21	61.25	68.36	34.22	18.12	41.88	59.00	36.00	19.61	51.00	63.90	18.60
R18 (ℓ_2)	53.12	86.25	27.50	60.25	49.53	51.46	85.27	24.14	35.39	47.50	51.96	55.23	20.78	27.50	61.25	51.12	39.50	44.30	47.60	63.70	22.80
R18 (ℓ_∞)	52.49	84.69	23.62	61.12	50.94	51.67	83.57	25.86	35.55	47.05	56.07	55.23	18.83	26.88	56.88	50.88	40.62	42.27	47.50	63.30	22.60
R50	75.80	97.19	83.62	49.12	52.66	51.04	95.62	33.67	38.98	49.11	70.71	73.91	37.97	23.75	48.12	61.25	34.38	17.42	54.50	65.40	17.90
R50 (ℓ_2)	62.83	92.81	32.12	66.12	56.41	62.71	90.71	26.17	40.31	53.84	63.57	63.75	26.09	25.62	60.62	59.38	41.75	43.98	53.00	66.30	23.90
R50 (ℓ_∞)	63.86	91.25	29.25	64.25	54.37	57.50	91.07	30.70	38.52	53.39	68.04	64.06	26.25	25.62	58.75	60.50	43.25	43.05	53.10	66.70	24.70
WRN50-2	76.97	98.28	82.38	51.00	54.69	54.17	97.23	34.92	40.62	50.98	75.18	75.39	42.27	28.75	56.88	64.12	36.50	18.28	57.30	67.20	19.20
WRN50-2 (ℓ_2)	66.90	94.69	35.62	67.25	60.00	63.96	93.39	28.36	41.02	53.21	66.96	65.23	28.28	28.12	60.00	60.00	42.88	43.28	54.80	67.30	24.30
WRN50-2 (ℓ_∞)	68.41	95.00	33.12	65.75	56.09	59.17	94.73	30.94	38.12	54.73	73.39	65.47	25.86	30.63	63.75	61.88	46.62	44.92	55.40	67.70	24.00
ConvMixer-768-32	80.16	**99.22**	98.00	50.62	56.72	56.25	98.04	39.77	43.91	56.43	86.25	80.23	**56.02**	26.88	64.38	70.75	44.50	22.73	63.30	69.50	19.50
XCiT-S12	**81.97**	98.91	**98.88**	55.12	59.38	64.17	**98.75**	**69.84**	**46.72**	**62.14**	**91.07**	**81.41**	55.62	**37.50**	61.88	71.12	**57.75**	25.55	**68.90**	70.90	19.50
XCiT-S12 (ℓ_∞)	72.34	96.88	47.62	66.50	58.91	61.04	96.88	36.95	39.77	56.61	82.14	70.70	40.47	31.87	63.75	70.75	48.75	46.80	60.60	70.00	24.10
XCiT-M12 (ℓ_∞)	74.04	97.34	48.25	66.88	60.16	62.29	96.96	36.80	39.06	57.59	81.43	70.86	41.17	26.25	66.88	71.00	52.62	47.27	60.90	70.40	**24.80**
XCiT-L12 (ℓ_∞)	73.76	98.12	47.38	**69.38**	**60.62**	**64.58**	98.66	41.95	41.72	58.21	84.11	70.62	42.27	35.62	**69.38**	**74.00**	54.12	**48.83**	63.40	**71.10**	22.70
Humans	-	88.67	66.09	60.75	58.28	63.91	88.82	46.43	56.09	55.11	75.89	**84.51**	55.37	**87.12**	**75.31**	**91.62**	47.12	**77.55**	-	-	-

decreased it on Transformers. Still, Transformers achieve higher accuracies than humans in this category. Of all studied models, the adversarially-trained *XCiT-L12* performs best on *eidolon I-III*, *silhouette, sketch, and cue-conflict*. However, it is also worth noting that it contains 50% more parameters compared to the largest CNN we analyze. In general, we can not conclude that more parameters are always better as we *e.g.* see some reduction in error consistency from robust *XCiT-S/M12* to *XCiT-L12*. Further, the clean *ConvMixer* which contains no self-attention but patch-embeddings, shows also an increased *cue-conflict*. Generally, there is a trade-off between CNNs and Transformers in almost all studied datasets. We hypothesize that patch-embeddings may naturally be slightly shifted toward shape bias compared to CNNs.

Is adversarial training a good option to achieve human-like reasoning? While AT does improve *cue-conflict* significantly and shifts the internal decision process toward human-like shape bias behavior, it also decreases OOD performance across many datasets. Most notably, AT causes a significant drop in robustness to changes in *contrast*, *rotation*, and *uniform noise* compared to clean training. Interestingly, it also always reduces *high-pass* performance. In the case of *XCiT*-models this performance is slightly worse than for humans after AT, although the clean model significantly outperformed humans (by approx. 23%). We see the largest OOD drops in *XCiT*, while the *ResNets* show only minor impairments. Based on these findings, AT alone is not sufficient to shift models toward human-like reasoning in all aspects. In the next section, we investigate the frequency spectra of OOD samples and show that they provide an indication of whether AT can, in principle, help to increase performance.

5. A Frequency Perspective on Adversarial Training and Out-Of Distribution Data

In Fig. 5, we plot for all considered OOD image categories their frequency power spectra, radially integrated as described in Fig. 2, and compare the frequency spectra to the spectrum of the clean training images. From this comparison, it is apparent that some OOD categories deviate heavily from the natural image distribution in terms of their spectra. This is obviously true for *high-pass* and *low-pass* images as well as for *uniform-noise* and *edge*, where the differences are particularly strong in the high-frequency regime, but it is also visible for *contrast*, *rotation*, and *power equalization*, or *phase scrambling*, with significant differences in the lowest frequencies. Although adversarial attacks might slightly alter the frequency spectrum of attacked images, they are ϵ bounded and will therefore not significantly change the frequency distribution over all samples. Thus, it would be natural that AT (*i.e.* adding more training samples with a spectral distribution similar to the one of clean images) would harm the transferability of models to such out-of-domain categories. In fact, Tab. 3 shows exactly this trend: both types of AT cause a consistent decay in classification accuracy for the OOD categories *high-pass*, *low-pass*, *uniform-noise*, *contrast*, *rotation* and *power equalization*. When the differences are in the low-frequency regime as for *contrast*, the decay seems to be particularly strong. This observation supports the findings by [19] that AT can harm robustness to other corruption types and provides an initial explanation: Adding more training samples from the original spectral distribution can harm the generalization to other diverging spectral distributions.

Fig. 5 also shows that some OOD categories have power spectra that are quite similar to the spectra of the original data (and thus of adversarial examples). For these categories, *e.g. eidolon I*, *eidolon II*, *eidolon II*, *false colour* or

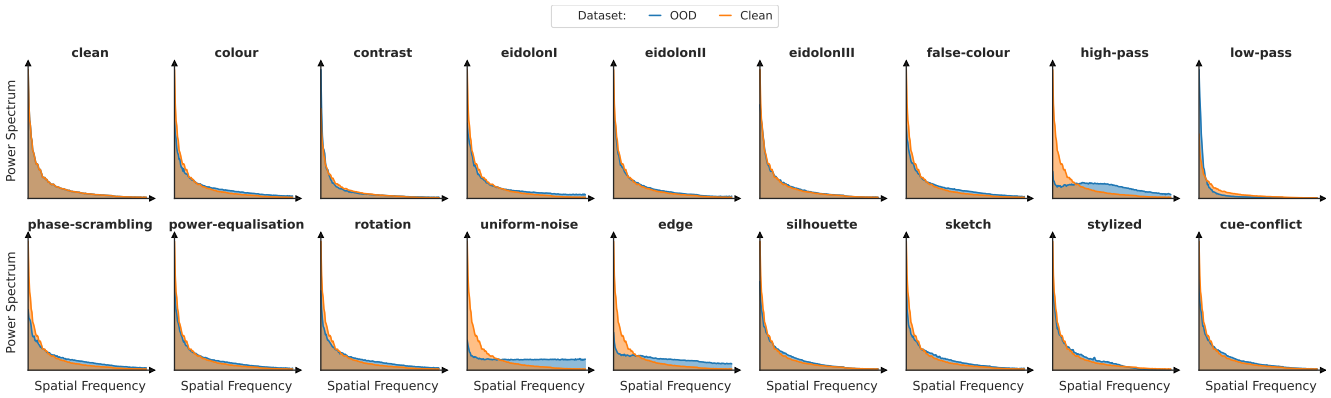

Figure 5. Frequency distribution of the utilized OOD datasets in comparison to comparable ImageNet validation samples (clean). Distributions are normalized by their integral. Frequency increases along the X-axis.

cue-conflict, AT does not lead to a decay in accuracy but can even lead to improvement in some cases. In the following, we will discuss in which cases we might expect this improvement.

From the above observation, we see that the OOD data should share some important properties with the clean data to benefit from AT, *i.e*. the frequency distribution should not differ too much. At the same time, it has been argued that convolutional neural networks tend to decide based on texture information [2], which is local and rather mid to high-frequency. Thus, adversarial examples can attack such models by slightly altering the image in these frequency bands. While this may vary by dataset [34–36], at least some high-frequency is always present as *e.g*. adversarial attacks can be detected in the frequency spectrum [37].

To compensate for these attacks, robust models desensitize to high-frequency and instead shift their decisions towards global cues that involve low-frequency information, which can typically be observed in FFT-spectra of perturbations after AT (*e.g*. [38], Fig. 8). The desensitization of high-frequencies during training also results in more robust models, as shown from various perspectives such as injecting noise patches to inputs [21], blurring feature-maps [39], splitting and regularizing frequency information [19], or low-pass filtering intermediate feature-maps [40] during training. There seem to be sufficient indicators to reasonably assume that shifting the decisions toward low-frequency information by removing the focus from high-frequency is at least a necessary ingredient of robustness. Clearly, AT encourages this shift, which can also be seen in weights of convolution filters of robust models [41, 42].

Likewise, texture bias can also be analyzed from a frequency perspective. Textures contain high-frequency information while shapes can not be represented without low-frequency bands. As non-robust neural networks naturally prefer high-frequency information for predictions they reason based on textures. Under AT, models rely less on local high-frequency information and prioritize the lower-frequent information, that encodes global structures such as shapes. This effect can be well seen in the *cue-conflict* performance where images contain both types of information in the images, and models can choose which information to prioritize. From an information perspective alone, both choices would acceptable.

Ultimately, this perspective does not explain all findings and other mechanics may influence the decision process. For example, *stylized* performance improves under AT for CNNs while the accuracy of Transformers, starting at a higher level, is decreasing. It can just provide an intuition of why the model decisions learn to shift towards a more global, shape bias - given that the overall spectral distribution remains very similar to the original training data distribution in the *cue-conflict* category.

6. Conclusion

We have extended previous experiments that studied the influence of ℓ_2-AT on the reasoning of neural networks in comparison to human reasoning. Our findings indicate, that previous observations scale to ℓ_∞-AT, other CNNs, and even Transformers. In general, we find that robust Transformers appear to be more similar to human reasoning than CNNs as they perform better on OOD datasets and increasingly reason based on shape information. Still, AT results in degradation against some corruptions that do not seem to affect humans or models trained without AT. Finally, we propose an explanation of why AT enforces shape bias from a frequency perspective: AT seems to hurt generalization against OOD datasets where the spectral distribution significantly diverges from the training data. In other cases, AT causes the model to shift its decision from local high-frequency information to global shape information, which better resembles the behavior of humans.

References

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. 1, 3

[2] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019. 1, 2, 6

[3] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel, "Partial success in closing the gap between human and machine vision," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. 1, 2, 3

[4] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing, "Learning robust global representations by penalizing local predictive power," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019. 1, 2

[5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases* (Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, eds.), (Berlin, Heidelberg), pp. 387–402, Springer Berlin Heidelberg, 2013. 2

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (Yoshua Bengio and Yann LeCun, eds.), 2014. 2

[7] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. 2

[8] 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*, 2018. 2

[9] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, 2020. 2

[10] Francesco Croce and Matthias Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," 2020. 2

[11] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song, "Delving into transferable adversarial examples and black-box attacks," in *International Conference on Learning Representations*, 2017. 2

[12] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018. 2

[13] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song, "Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms," in *Computer Vision – ECCV 2018*, pp. 158–174, Cham, Switzerland: Springer, Oct. 2018. 2

[14] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Computer Vision – ECCV 2020* (Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, eds.), (Cham), pp. 484–501, Springer International Publishing, 2020. 2

[15] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Li zhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *International Conference on Machine Learning*, 2020. 2

[16] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann, "Data augmentation can improve robustness," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, eds.), 2021. 2

[17] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann, "Improving robustness using generated data," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, eds.), 2021. 2

[18] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan, "Better diffusion models further improve adversarial training," *arXiv preprint arXiv:2302.04638*, 2023. 2

[19] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox, "Improving robustness against common corruptions with frequency biased models," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10191–10200, IEEE, 2021. 2, 5, 6

[20] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019. 2

[21] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D. Cubuk, "Improving robustness without sacrificing accuracy with patch gaussian augmentation," 2020. 2, 6

[22] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann, "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018. 2

[23] Robert Geirhos, Kristof Meding, and Felix A. Wichmann, "Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds.), vol. 33, pp. 13890–13902, Curran Associates, Inc., 2020. 2

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. 2

[25] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, June 2022. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (Marina Meila and Tong Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021. 2

[27] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan, "Billion-scale semi-supervised learning for image classification," 2019. 2

[28] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby, "Big Transfer (BiT): General Visual Representation Learning," in *Computer Vision – ECCV 2020*, pp. 491–507, Cham, Switzerland: Springer, Oct. 2020. 2

[29] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry, "Do adversarially robust imagenet models transfer better?," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds.), vol. 33, pp. 3533–3545, Curran Associates, Inc., 2020. 3

[30] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal, "A light recipe to train robust vision transformers," in *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023. 3

[31] Ross Wightman, "Pytorch image models." https://github.com/rwightman/pytorch-image-models, 2019. 3

[32] Asher Trockman and J. Zico Kolter, "Patches are all you need?," 2022. 3

[33] Ricard Durall, Margret Keuper, and Janis Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[34] Shishira R. Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava, "A frequency perspective of adversarial robustness," *CoRR*, vol. abs/2111.00861, 2021. 6

[35] Antonio A. Abello, Roberto Hirata, and Zhangyang Wang, "Dissecting the high-frequency bias in convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pp. 863–871, Computer Vision Foundation / IEEE, 2021. 6

[36] Rémi Bernhard, Pierre-Alain Moëllic, Martial Mermillod, Yannick Bourrier, Romain Cohendet, Miguel Solinas, and Marina Reyboz, "Impact of spatial frequency based constraints on adversarial robustness," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, IEEE, 2021. 6

[37] Paula Harder, Franz-Josef Pfreundt, Margret Keuper, and Janis Keuper, "Spectraldefense: Detecting adversarial attacks on cnns in the fourier domain," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021. 6

[38] Julia Grabinski, Janis Keuper, and Margret Keuper, "Aliasing and adversarial robust generalization of CNNs," *Mach. Learn.*, vol. 111, pp. 3925–3951, Nov. 2022. 6

[39] Samarth Sinha, Animesh Garg, and Hugo Larochelle, "Curriculum by smoothing," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds.), vol. 33, pp. 21653–21664, Curran Associates, Inc., 2020. 6

[40] Julia Grabinski, Steffen Jung, Janis Keuper, and Margret Keuper, "Frequencylowcut pooling - plug and play against catastrophic overfitting," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pp. 36–57, Springer, 2022. 6

[41] Paul Gavrikov and Janis Keuper, "CNN Filter DB: An Empirical Investigation of Trained Convolutional Filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19066–19076, June 2022. Selected for Oral presentation. 6

[42] Paul Gavrikov and Janis Keuper, "Adversarial Robustness Through the Lens of Convolutional Filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 139–147, June 2022. 6