NON-ROBUST FEATURES THROUGH THE LENS OF UNIVERSAL PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work ties adversarial perturbations to so-called non-robust features. These are features which are susceptible to small perturbations and believed to be incomprehensible to humans, but still useful for (generalizable) prediction. We study universal adversarial perturbations and demonstrate that the above picture is more nuanced. Specifically, even though universal perturbations—similarly to standard adversarial perturbations—do leverage non-robust features, these features tend to be fundamentally different from the "standard" ones and, in particular, non-trivially human-aligned. Namely, universal perturbations are more semantic, and have human-aligned locality and spatial invariance properties. However, we also show that these semantic non-robust features have much less predictive signal than general non-robust features, indicating that the bulk of non-robust features requires additional techniques to probe their structure. Our findings thus take a step towards understanding the nature of non-robust features.

1 INTRODUCTION

Modern deep neural networks have been extremely successful across many tasks, but they exhibit a major vulnerability to adversarial examples (Szegedy et al., 2014). The models' brittleness to small, imperceptible perturbations highlights one particular way in which models' decision making deviate from that of humans. Recent work gives evidence that this difference may arise from the presence of certain useful *non-robust* features in our datasets (Ilyas et al., 2019). These are brittle features that are sensitive to small, imperceptible perturbations, yet provide enough predictive signal to generalize well on the underlying classification task. Our models' reliance on these features is the key driver of their vulnerability to adversarial examples.

While prior work has shown that non-robust features exist in natural datasets, there is little understanding of what these features might be. After all, the entire premise of adversarial examples is that their visualizations are visually incomprehensible to humans, so it is unclear what structure the associated non-robust features possess. A number of works suggested that these features may exploit properties of the dataset that are misaligned with human perception (Ilyas et al., 2019), such as high-frequency information (Yin et al., 2019). However, we lack a more direct study of the properties of non-robust features. In fact, we do not understand well how models make decisions in general, so it is unclear how much we can understand about these features that are a priori imperceptible.

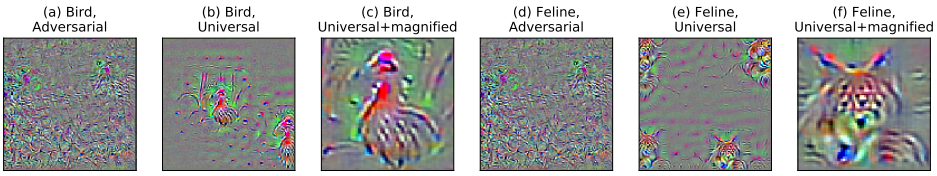


Figure 1: ℓ_2 adversarial perturbations at $\epsilon = 6.0$ for different target classes on ImageNet-M10: (a,d) standard adversarial perturbations on a single image, (b,e) universal adversarial perturbations, and (c,f) zooming in on the most semantic patch.

In this work, we show that by adding additional constraints to adversarial perturbations, we can probe the associated non-robust features and gain new insights into their nature. In particular, we study

non-robust features via a careful analysis of universal adversarial perturbations (Moosavi-Dezfooli et al., 2017a), i.e. adversarial perturbations that generalize across many inputs. We demonstrate that universal perturbations possess properties quite different from standard adversarial perturbations (beyond their generalization across images), and this allows us to uncover new aspects of non-robust features. Concretely, our findings are:

Universal perturbations have natural, more human-aligned properties. We show that universality alone can be used as a prior for generating more interpretable adversarial perturbations. Specifically, universal perturbations are *highly semantic* compared to "standard" adversarial perturbations (e.g., Figure 1). In particular, (1) the most semantic parts of the perturbation contain the most signal; and (2) universal perturbations are approximately *spatially invariant*, in that they are still effective after translations. Overall, universal perturbations tend to be more human-aligned than the standard ones.

Non-robust features can be semantic. We leverage the properties of universal perturbations to show that some of the non-robust features can be semantic. Specifically, we compare the sensitivity of natural and (adversarially) robust models to rescalings of these perturbations to demonstrate that universal perturbations likely rely on non-robust features (rather than robust ones). We then show that most of the corresponding non-robust signal comes, in turn, from the local patches with the most semantic content.

Universal perturbations contain less non-robust signal. We find that the non-robust features leveraged by universal perturbations are weaker than those leveraged by standard adversarial perturbations, despite being more semantic. We measure both (1) generalizability to the original test set and (2) transferability of perturbations across independent models, following the methodology of Ilyas et al. (2019). Under these metrics, universal perturbations consistently obtain non-trivial but substantially worse performance than standard adversarial perturbations.

2 PRELIMINARIES

We consider a standard classification task on input-label pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ sampled from a data distribution $\mathcal{D}$ in which the goal is to learn a classifier $C : \mathcal{X} \to \mathcal{Y}$ that predicts y on input x. Following Ilyas et al. (2019), we introduce the following terminology.

Non-robust vs. robust features. A *useful* feature for classification is a function that is (positively) correlated with the correct label in expectation. A feature is *robustly useful* if, even under adversarial perturbations (within a specified set of valid perturbations Δ) the feature is still useful. Finally, a *useful, non-robust* feature is a feature that is useful but not robustly useful. These features are useful for classification in the standard setting, but can hurt accuracy in the adversarial setting (since their correlation with the label can be reversed). For conciseness, throughout this paper we will refer to such features simply as non-robust features.

Universal perturbations. A *universal adversarial perturbation* (or just *universal perturbation* for short), as introduced in Moosavi-Dezfooli et al. (2017a), is a perturbation δ that causes the classifier C to predict the wrong label on a large fraction of inputs from $\mathcal{D}$. We focus on *targeted* universal perturbations that fool C into predicting a specific (usually incorrect) target label t. Thus, a (targeted) universal perturbation $\delta \in \Delta$ satisfies $P_{(x,y)\sim\mathcal{D}}[C(x+\delta) = t] = \rho$, where ρ is the *attack success rate* (ASR) of the universal perturbation. The only technical difference between universal perturbations and (standard) adversarial perturbations is the use of a single perturbation vector δ to be applied to all inputs. Thus, one can think of universality as a constraint for the perturbation δ to be input-independent. Universal perturbations can leverage non-robust features in data, as we show here; we refer to these simply as *universal non-robust features*.

ℓ_p perturbations. We study the case where Δ is the set of ℓ_p-bounded perturbations, i.e. $\Delta = \{\delta \in \mathbb{R}^d \mid ||\delta||_p \leq \epsilon\}$ for $p = 2, \infty$. This is the most widely studied setting for research on adversarial examples and has proven to be an effective benchmark (Carlini et al., 2019). Additionally, ℓ_p-robustness appears to be aligned to a certain degree with the human visual system (Tsipras et al., 2019).

2.1 COMPUTING UNIVERSAL PERTURBATIONS

We compute universal perturbations by using projected gradient descent (PGD) on the following optimization problem:

$$\min_{\delta \in \Delta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[\mathcal{L}(C, x+\delta, t) \Big] \tag{1}$$

where $\mathcal{L}$ is the standard cross-entropy loss function for classification and t is the target label[1]. As observed in Moosavi-Dezfooli et al. (2017a), universal perturbations trained on only a fraction of the dataset still generalize to the rest of the dataset, so in practice it suffices to approximate the expectation in equation 1 using a relatively small batch of data drawn from $\mathcal{D}$. We call the batch we optimize over the *base set*, and use K to refer to its size. We describe the hyperparameters used for the optimization and their selection in Appendix B.

2.2 EXPERIMENTAL SETTINGS

Dataset. We conduct our experiments on a subset of ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky et al., 2009) datasets. The Mixed10 ImageNet (Engstrom et al., 2019) dataset, which we refer to as ImageNet-M10, is formed by sub-selecting and grouping together semantically similar classes from ImageNet[2], providing a more computationally efficient alternative to the full dataset, while retaining some of ImageNet's complexity.[3] In the main text, we focus our results on ImageNet-M10, and in the Appendix we provide additional results across different datasets and settings.

Models. We use the standard ResNet-18 architecture (He et al., 2016) for most of our experiments.

3 PROPERTIES OF UNIVERSAL PERTURBATIONS

We find that universal perturbations, in contrast to adversarial perturbations, are much more aligned with human perception. Beyond visualizations, we demonstrate this in the following ways: (1) local patches that exhibit class-specific characteristics also contain more predictive signal, and (2) these perturbations are approximately spatially invariant.

3.1 PERCEPTUAL ALIGNMENT

Visual comparison with standard adversarial perturbations. Standard ℓ_p adversarial perturbations are often thought to be incomprehensible from a human perspective (Szegedy et al., 2014; Ilyas et al., 2019). Even when magnified for visualization, these perturbations are not identifiable with their target class. In contrast, universal perturbations exhibit vastly different properties—the perturbations, when amplified for visibility, contain patches that are easily recognizable as the target class (see Figure 2 for example universal perturbations, and Figure 1 for a comparison to standard perturbations). While observations along this line are certainly not new, we investigate this much further with additional analyses below.

Predictive signals are localized to semantic patches. While universal perturbations contain parts that look semantic to humans, they also contain less interpretable regions; a priori, it is unclear which parts the model actually relies on. To investigate this, we search over 64 different 80×80 random patches across the image, and evaluate their ASR. For ℓ_2 perturbations, the patches are visibly variable in norms, so a possible concern is that the model is only reacting to the patches with the highest norm; to correct for this, we linearly scale up all patches to have the same norm as the largest norm patch. Even after normalization, the patches with the highest rates were much more characteristic of the corresponding class (Figure 3). For ℓ_∞ perturbations, where patches all have similar norms (so no further normalization was done), we also observe that their ASR is proportional

[1] While many different algorithms have been developed for computing universal perturbations with varying attack success rates, we do not consider them as achieving the highest rate is not necessary for our investigations.

[2] For example, Mixed10 consists of the ten coarse-grained class labels—`dog`, `bird`, `insect`, `monkey`, `car`, `feline`, `truck`, `fruit`, `fungus`, `boat`—with 7,200-7,800 training and 300 test images per class.

[3] ImageNet-M10 is similar to restricted ImageNet, which has been used for adversarial robustness research.

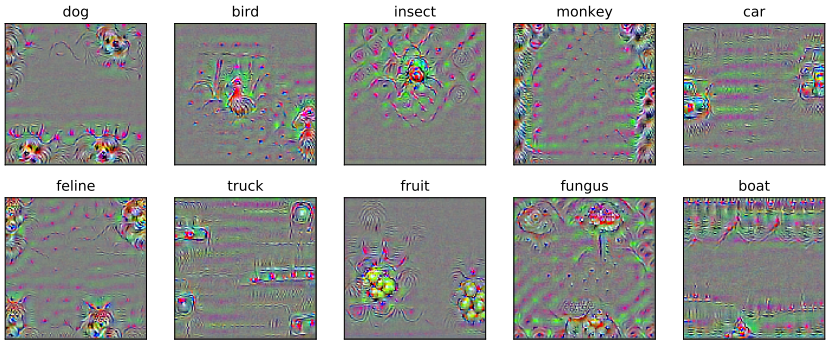


Figure 2: ℓ_2 universal perturbations at $\epsilon = 6.0$ for different target classes. We observe patterns that closely resemble dog and cat (feline) faces, birds, outlines of cars and trucks, an insect, and mushrooms (fungus).

to their visual saliency. This demonstrates that the model is indeed reacting mostly to the most salient parts of the perturbation.

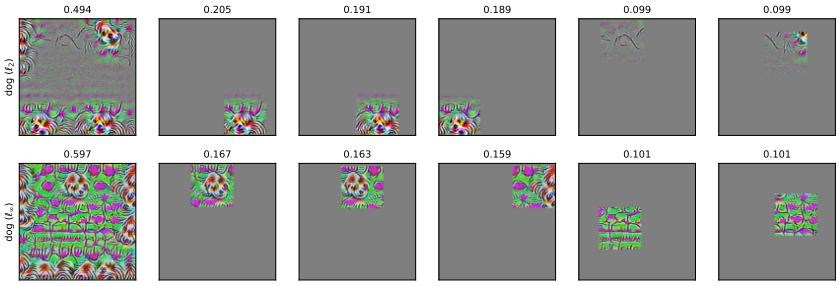


Figure 3: First column shows the original perturbation for the target class (top: ℓ_2; bottom: ℓ_∞). Next five columns show three patches with the highest ASR, and two with the lowest (among 64 patches tested). Top of each perturbation indicates its ASR on the test set.

3.2 Spatial invariance

Not only are adversarial perturbations undecipherable, but they are also extremely brittle to translations, as we demonstrate. In contrast, we find that universal perturbations are substantially translationally invariant. As spatial invariance is one of the key properties of the human visual system (Hubel & Wiesel, 1968), this indicates that adversarial perturbations can be made more human-aligned by imposing universality.

We quantify spatial invariance by measuring the ASR of translated perturbations. Specifically, we compute targeted standard adversarial perturbations on a sample of 256 images from the ImageNet-M10 test set. Then, we evaluate the ASR of different translated copies of these perturbations over the test set (copies of each perturbation are evaluated only on its corresponding image). For comparison, we take a precomputed set of universal perturbations of the same norm, and evaluate them across all images. Figure 15 shows that universal perturbations still achieve non-trivial ASR after translations of varying magnitudes. In contrast, standard adversarial perturbations achieve a chance-level 10% ASR when shifted by more than eight pixels (two grid cells in Figure 15). This illustrates that the model is reacting to universal perturbations in a different manner from that to adversarial perturbations.

4 Universal Perturbations Leverage Non-Robust Features

The predominant belief about non-robust features is that they are incomprehensible to humans. We saw, however, that universal perturbations are significantly more interpretable and possess other

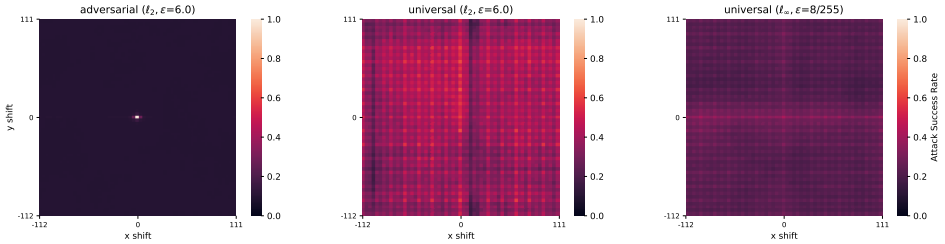


Figure 4: Evaluation of translated adversarial and universal ℓ_2, ℓ_∞ perturbations for the target class `bird` on ImageNet-M10. We evaluate a subsampled grid with strides of four pixels. The value at coordinate (i, j) represents the average ASR when the perturbations are shifted by i pixels right and j pixels up, with wraparound to preserve information; the center pixel at $(0, 0)$ represents the ASR of the original unshifted perturbations. For universal perturbations, the ASR at each location is averaged over ten perturbations.

human-aligned properties. Since universal perturbations are bounded in norm, they must be leveraging non-robust features. Then, we might conclude that there exist in fact non-robust features that have semantic properties.

However, this is not necessarily the case, as there is a possibility of *robust feature leakage* (Goh, 2019): that is, that an adversarial perturbation primarily relies on gradients of robust features. While a small perturbation cannot fully manipulate a robust feature (i.e., flip its correlation), it can still affect the feature to a degree. Thus, it is possible that the semantic parts of universal perturbations were attenuating signal from the robust features, and that allowed the non-semantic aspects of the perturbations to manipulate non-robust features enough to change the overall classification.

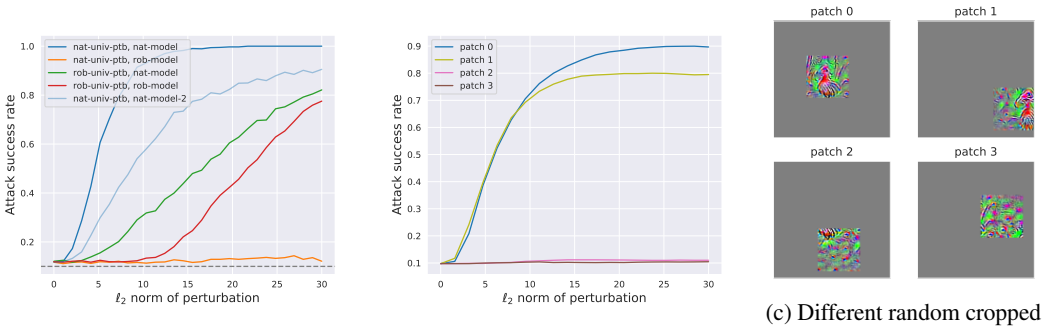
Scaling analysis on natural vs. robust model We provide evidence that robust leakage is unlikely to arise by applying a scaling analysis. Our goal is to evaluate whether ℓ_2 robust models react to these universal perturbations. We do so by "amplifying" the universal perturbations. That is, given a universal perturbation δ, we evaluate the ASR of $t\delta$, where t is a scaling parameter. In Figure 5, we measure the sensitivity of natural and robust models to scaled perturbations, as we scale t. When we perform this scaling on a natural (non-robust) model, the strength of the attack grows as we increase t. This holds even if networks used to train the universal perturbation and the one used to evaluate its ASR differ. Intuitively, the non-robust features in any non-robust model pick up on the perturbation, and as t increases, the signal they pick up increases as well. If there was robust leakage, we would expect a similar qualitative behavior for the robust model. Instead, we find that the robust model does *not* respond at all to any scaling of the perturbation.

In fact, a more fine-grained scaling analysis shows that universal non-robust features predominantly occur in the semantic parts of the image. In Section 3 we had observed that models are most sensitive to the most semantic parts of universal perturbations. We repeat the above scaling evaluation on natural vs. robust models, except we scale only patches of the perturbation, of varying levels of semantic content (Figure 5b). As before, we find that none of the patches, even after scaling, trigger the robust network. On the non-robust models, the relative effectiveness of local patches is consistent with semantic content: patches with more semantic content scale more effectively than patches without it.

This enables us to conclude that some of the non-robust features are leveraging semantic information. This suggests, in turn, that non-robust features may not be entirely unintelligible to humans, as was previously conjectured. Our visualizations(Figure 5c) demonstrate that universal non-robust features may be performing something intuitive, such as detecting the presence of a "diluted" bird. Thus, one source of non-robustness in networks may be reliance on non-robust features that are faint but still human-aligned.

5 Universal Non-Robust Features Convey Less Signal

So far, we demonstrated that universal perturbations tend to leverage highly semantic non-robust features. We now demonstrate that these universal non-robust features have less generalizable "signal"



(a) Scaling universal perturbations

(b) Scaling cropped patches

(c) Different random cropped patches

Figure 5: Evaluating the sensitivity of universal perturbations at different scales: (a) we generate a universal perturbation (`nat-univ-ptb`) of norm $\epsilon = 6.0$ on the natural model (`nat-model`) for the target *bird*, and measure the sensitivity of natural and robust model (`rob-model`) at various rescalings. We also generate a universal perturbation (`rob-univ-ptb`) of norm $\epsilon = 30.0$ for the robust model and use it to show that the robust model *does* react to scalings of some signal. The same `nat-univ-ptb` is also evaluated on an independent natural model (`nat-model-2`) for control; (b) different cropped local patches of the same `nat-univ-ptb` are also evaluated; (c) visualizations of the cropped patches.

than standard non-robust features. As in Ilyas et al. (2019), we measure the usefulness of non-robust features based on: (1) how well a model generalizes to the original test set when it is trained on a dataset where the *only* useful features are those introduced by universal perturbations; and (2) the transferability of universal perturbations across models. Generalization directly measures the usefulness of universal non-robust features for learning, while transferability measures whether different models learn the same features.

5.1 Generalization from universal non-robust features

We construct two datasets, $\widehat{\mathcal{D}}_{univ}$ and $\widehat{\mathcal{D}}_{adv}$, where the only useful features are universal non-robust features and non-robust features, respectively.

Constructing $\widehat{\mathcal{D}}_{adv}$ and $\widehat{\mathcal{D}}_{univ}$. We follow the procedure outlined in Ilyas et al. (2019) to construct $\widehat{\mathcal{D}}_{adv}$ (which corresponds to $\widehat{\mathcal{D}}_{rand}$ from their paper): 1) For each input-label pair (x, y), select a target class t uniformly at random; 2) Compute an ℓ_2 adversarial perturbation δ on x targeted towards class t, and replace the original input-label pair (x, y) with $(x + \delta, t)$. By construction, x and t are decorrelated and the only features correlated with t are those leveraged by δ. Because δ is constrained to be within an ℓ_2 ball, it can only yield non-robust features. Next, we adapt the above procedure to construct $\widehat{\mathcal{D}}_{univ}$. To this end, in step 2, we use a universal perturbation in place of an adversarial perturbation.[4]

Comparing the signal in $\widehat{\mathcal{D}}_{adv}$ and $\widehat{\mathcal{D}}_{univ}$. We train new ResNet-18 models on the $\widehat{\mathcal{D}}_{adv}$ and $\widehat{\mathcal{D}}_{univ}$ datasets and evaluate them on the original test set [5] The best generalization accuracies from training on $\widehat{\mathcal{D}}_{univ}$ and $\widehat{\mathcal{D}}_{adv}$ were 23.2% and 74.5%, respectively. This indicates that universal non-robust features have much less generalizable signal than standard non-robust features.

5.2 Transferability of universal perturbations

Ilyas et al. (2019) suggest that the pervasive transferability of adversarial perturbations can be attributed to different models relying on similar non-robust features. Consequently, perturbations that leverage more such shared non-robust features should transfer better. Transferability of universal perturbations was first observed in Moosavi-Dezfooli et al. (2017a) and many follow-up works. Here

[4]There are a few complications in this new setup, however. See Appendix A.1 for more details.

[5]We report the best results found over a grid of training hyperparameters.

Table 1: Interpolating universality: generalization from $\widehat{\mathcal{D}}_{univ}$ on smaller base sets.

Size of base set (K)	Test Accuracy (%)
1	74.5
2	57.1
4	61.3
8	57.4
16	34.3
32	21.8
256	19.1

Table 2: Generalization from class-universal and sub-class universal non-robust features. We find minimal evidence to suggest that non-robust features are aligned with (sub)classes.

Source Class	Test Accuracy (%)
Random	23.2
Single Class	23.9
Single Sub-class	27.1

we are primarily interested in transferability as a proxy for detecting shared non-robust features, and compare the transferability of adversarial and universal perturbations.

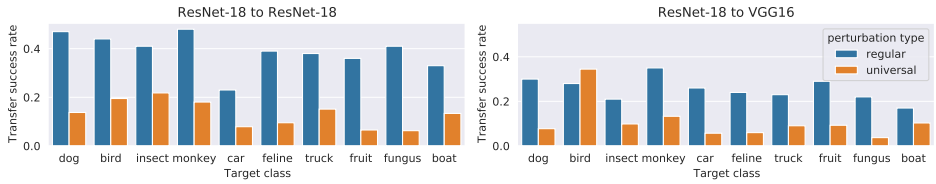


Figure 6: Comparing transferability of adversarial and universal ℓ_2 perturbations at $\epsilon = 6.0$. Perturbations are generated on the source model (ResNet-18), then transferred to a different target model: (left) another ResNet-18 model trained from a different random initialization, and (right) a VGG16 model.

To measure transferability, we (1) perturb examples using either an adversarial perturbation or a universal perturbation on the *source* model, and (2) measure the probability that the perturbed input is classified as the target class on a new model (a ResNet-18 trained independently). Similarly to when we study generalization, we only consider perturbations that succeed at fooling the source model. Furthermore, we only evaluate transfer on images that have labels different from the target label. Figure 6 shows that universal perturbations transfer much less, suggesting that models share fewer universal non-robust features. This is consistent with results from Section 5.1.

5.3 Interpolating Universality

The above experiments have demonstrated that while universal non-robust features are highly semantic, they also have less generalizable signal. To explore the potential underlying tradeoff, we investigate to what extent one can interpolate between the properties of universal and standard non-robust features. To this end, we explore two different methods of "interpolating universality."

Interpolating via changing the base set size. One natural way to interpolate universality is to vary K, the size of the base set (cf. Section 2.1). $K = 1$ corresponds to adversarial perturbations, and large enough K corresponds to fully universal perturbations. Table 1 shows that generalization behavior generally decreases with the size of the base set. [6] Notably, generalization begins to suffer even for relatively small values of K. In contrast, the semantics of universal perturbations improves as the base set becomes larger, in line with the hypothesized tradeoff (cf. Appendix A.3).

Interpolating via restricting the source class. One may expect that images from the same class share similar features, and that the same is true for non-robust features as well. This motivates us to interpolate universality along a different axis: the level of semantic hierarchy. Specifically, we consider *class*-universal and *subclass*-universal perturbations, where we restrict the base set to examples from a single class (e.g., dog) or subclass (e.g., Terrier). If certain non-robust features exist

[6] See Appendix A.2 for a follow-up experiment and more discussion.

only in images from the same class, we would expect *class*-universal perturbations to be more likely to find them. To examine this, we repeat the experiments from Section 5.1 using these perturbations (cf. Table 2); the generalization on the original test set hardly improves. This suggests that if non-robust features are specific to small subpopulations of images, then these subpopulations must be more fine-grained than division into different (sub)classes.

Discussion We observed that we can interpolate the signal present in universal non-robust features by varying the base set's size, and to a smaller extent, its class diversity. This can be intuitively captured by the following toy model: Each example comes with a set of non-robust features, and different examples share a varying subset of features depending on their closeness. [7] By computing perturbations of over a set of images, we effectively filter the set of perturbable non-robust features to their intersection; it would then follow naturally that this intersection, and hence signal, decreases with more images and as the images become farther apart.

Existing theoretical models (Tsipras et al., 2019; Allen-Zhu & Li, 2020) fail to distinguish between regular and universal perturbations, and hence fail to capture the phenomena observed here. We believe that more empirical observations are necessary to identify a precise theoretical model.

6 Related Work

Prior works observed that patterns that appear unintelligible to humans can induce misclassification while also being useful for generalization Jetley et al. (2018); Ilyas et al. (2019). Ilyas et al. (2019) refer to these as non-robust features, and posit that they may explain phenomena like the widespread existence and transferability of adversarial examples Szegedy et al. (2014); Papernot et al. (2016). Conversely, several works discuss the usefulness and perceptual alignment of robust features Tsipras et al. (2019); Santurkar et al. (2019). Our work goes beyond the simple view represented in these works that non-robust features are unintelligible while robust features are meaningful by finding semantic, albeit weakly transferable, non-robust features. We probe non-robust features via universal perturbations, which were first studied in Moosavi-Dezfooli et al. (2017a). Many follow-up works explore its various aspects (Moosavi-Dezfooli et al., 2017b; Poursaeed et al., 2018; Wu & Fu, 2019; Hayes & Danezis, 2019; Gupta et al., 2019; Shafahi et al., 2020), and some have also observed their semantic properties.

Beyond adversarial robustness, recent works have studied other ways in which models learn to use signals that appear that deviate from those used by humans, including texture, high-frequency content, and background (Geirhos et al., 2018; Yin et al., 2019; Xiao et al., 2020). Finally, many works propose saliency map methods to construct visualizations of features learned by networks Simonyan et al. (2013); Ribeiro et al. (2016); Sundararajan et al. (2017); Carter et al. (2019). Despite the visual appeal of many of them, they involve post-processing that can lead to misleading results Adebayo et al. (2018). In contrast, our visualizations universal perturbation involve no post-processing.

7 Conclusion

We revisited universal perturbations and used them to analyze non-robust features in a more fine-grained manner. We found that universal perturbations are highly semantic and spatially invariant (in contrast to adversarial perturbations), and gave first evidence of existence of perceptually aligned, interpretable non-robust features. But we also showed that they account for only a small part of all non-robust features. Understanding the structure and properties of the remaining bulk of non-robust features is a natural direction for future work. We explored just one method for restricting adversarial perturbations, so it is likely that more can be uncovered by using different restrictions (for instance, restricting signals to a certain subspace). More broadly, it would be valuable to develop a theoretical framework for reasoning about and studying features in a more fine-grained manner—one that goes beyond their correlation and robustness.

[7] While the exact definition of closeness is hard to define, we can imagine, for example, two images of dogs in the same pose with similar background share more non-robust features than with an image of a car.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020.

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

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

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. *arXiv preprint arXiv:1811.12231*, 2018.

Gabriel Goh. A discussion of 'adversarial examples are not bugs, they are features': Robust feature leakage. *Distill*, 2019. doi: 10.23915/distill.00019.2. https://distill.pub/2019/advex-bugs-discussion/response-2.

Tejus Gupta, Abhishek Sinha, Nupur Kumari, Mayank Singh, and Balaji Krishnamurthy. A method for computing class-wise universal adversarial perturbations. In *ArXiv preprint arXiv:1912.00466*, 2019.

Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *ArXiv preprint arXiv:1708.05207*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. In *The Journal of physiology, 195(1), 215-243.*, 1968.

Andrew Ilyas, Shibani Santurkar, Logan Engstrom, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Saumya Jetley, Nicholas Lord, and Philip Torr. With friends like these, who needs adversaries? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554*, 2017b.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. In *ArXiv preprint arXiv:1605.07277*, 2016.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S. Davis, and Tom Goldstein. Universal adversarial training. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Junde Wu and Rao Fu. Universal, transferable and targeted adversarial attacks. In *ArXiv preprint arXiv:1908.11332*, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Neural Information Processing Systems (NeurIPS)*, 2019.

APPENDIX

In Appendix A, we provide deferred details and analyses from Section 5.

In Appendix B, we provide details of the experimental setup and all hyperparameters used in the experiments.

In Appendices C to F, we give a selection of additional results for experiments in the main paper across different settings.

In Appendix G, we discuss the diversity of universal perturbations.

A DEFERRED DETAILS FROM SECTION 5

A.1 DETAILS OF $\widehat{\mathcal{D}}_{univ}$ CONSTRUCTION

Here we provide additional details for constructing the $\widehat{\mathcal{D}}_{univ}$ dataset in Section 5.1. As it is rather expensive to compute a separate universal perturbation for each example, we compute and use the same universal perturbation per batch of examples. One possible concern is that re-using same perturbations results a reduction in "diversity" of features leveraged. Our experiments indicated that the precise number of universal perturbations does not impact generalization.

Another complication here is that for a given ϵ, universal perturbations fool models only on a fraction ρ of all examples (while adversarial perturbations almost always fool models). Hence, we only include examples on which the universal perturbation succeeds (i.e., a pre-trained model classifies the perturbed input as t). This filtering step reduces the size of the dataset, so we decrease the number of examples in $\widehat{\mathcal{D}}_{adv}$ accordingly in our evaluation. We also make sure, as before, that the distribution of new and original labels is uniform and independent so that the original label is uncorrelated with the new label.

A.2 ADDITIONAL ANALYSIS ON K-INTERPOLATION

One plausible hypothesis for degradation of signal quality between $K = 1$ and higher K is that optimizing the perturbation over more images splits the effective perturbation "budget," and is akin to optimizing over each image separately with a reduced budget, say ϵ/K. To test this hypothesis, we constructed a $\widehat{\mathcal{D}}_{adv}$ with a reduced budget of $\epsilon = 3.0$. Accuracy of 76.6% in Table 5 shows that the generalization is on par with $\epsilon = 6.0$ (and is much higher than $K = 2, \epsilon = 6.0$), so it seems unlikely that this simple budget sharing can explain the above interpolation phenomena. What is more likely happening is that different non-robust features in different images are interacting.

A.3 INTERPOLATING UNIVERSALITY

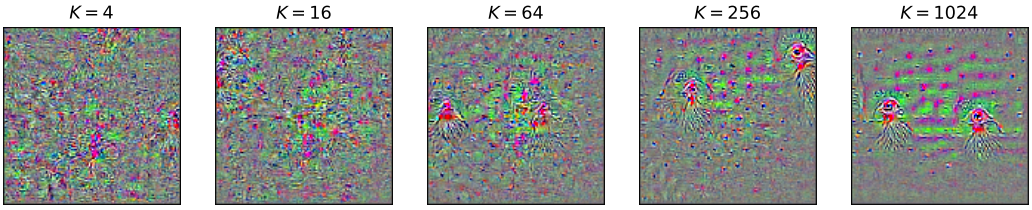


Figure 7: Visualization of universal perturbations for target *bird*, computed over base sets of different size K. We observe that the visual quality improves with the size of the base set.

Recall from Table 1 that generalization decreases with the size of the base set. In contrast, the semantic equality of universal perturbations improves as the base set becomes larger (Figure 7). This suggests a possible tradeoff between semantic quality and generalizable signal, at least along this particular axis. However, a more careful examination suggests that the picture might be more nuanced: better semantics only solidify at higher values of K ($K \gtrsim 64$), whereas generalization

suffers even at relatively small values of K ($K \sim 16$). Understanding whether these results are both manifestations of the same underlying phenomena require further work.

B EXPERIMENTAL SETUP

All experiments in the paper involve a combination of: training a model, computing adversarial perturbations, and computing universal perturbations. Below, we provide details for all the hyperparameters and their selection.

Robustness threshold We use the same robustness threshold throughout: $\ell_2, \epsilon = 6.0$ and $\ell_\infty, \epsilon = 8/255$ for ImageNet datasets, and $\ell_2, \epsilon = 1.0$ for CIFAR-10. (There is no one "right" ϵ, so choose standard values and just maintain consistency.)

B.1 TRAINING MODELS

Training base models All models are trained with SGD with momentum and standard data augmentation. For corresponding robust models, the same optimization parameters are used with PGD training (Madry et al., 2017) using 3 steps and attack step size $\frac{2}{3}\epsilon$ (for both ℓ_2 and ℓ_∞). Weight decay of 5e-4 was used in all cases. Their accuracies are shown in Table 3.

Table 3: Accuracies of standard and robust models used in our datasets. For standard models, the robust accuracy is close to 0% for both ℓ_2 and ℓ_∞ perturbations.

Dataset	Model	Standard Test Accuracy (%)	Robust Test Accuracy (%)
ImageNet-M10	Standard	95.7	< 1.0
	ℓ_2	86.7	59.8
	ℓ_∞	87.6	59.2
CIFAR-10	Standard	94.8	< 1.0
	ℓ_2	80.0	50.7

Training models on constructed datasets The models on the constructed datasets, including $\widehat{\mathcal{D}}_{univ}$ and $\widehat{\mathcal{D}}_{adv}$ (Section 5.1), were trained over the following grid of hyperparameters: three learning rates (0.01, 0.05, 0.1), two batch sizes (128, 256), including/ not including a single learning rate drop by a factor of 10. All models were trained with 400 epochs, standard data augmentation, and weight decay $5 \cdot 10^{-4}$.[8] In general, there were several settings of hyperparameters with similar quality; we only report the ones corresponding to the best runs in Table 4.

As an exception, for the K-interpolated datasets (Section 5.3), we fixed a single set of hyperparameters to train models.

B.2 COMPUTING ADVERSARIAL PERTURBATIONS

Adversarial perturbations For a given ϵ, attack step size of $\frac{\epsilon}{3}$ with 10 steps was used (no random restarts.)

Universal perturbations The hyperparameters used in the PGD algorithm for universal perturbations were: learning rate schedule[9], number of epochs, and batch size.[10] In order to see the impact of the choice of these parameters, we ran a grid search over the parameters; a sample of these selections is visible in Figure 8. Overall, we observed that the exact choice of these parameters did not have a large impact on the final training accuracy (and the quality of the resulting perturbations, both in terms of visuals and generalization to other images) as long as (1) the initial learning rate is not too low or too high, and (2) a sufficient number of epochs is used. So in all of our experiments, we fixed a learning rate of 2.0 and a batch size of 128 or 256, and trained a sufficient number of epochs (max

[8] Without data augmentation, the model overfit on these constructed datasets.

[9] Sometimes this is known as step size in the adversarial examples literature.

[10] Batch size only matters if it is smaller than the size of the base set, K.

Table 4: Default and best hyperparameters for training models on different datasets.

Source Dataset	Dataset	Epochs	LR	Batch Size	LR Drop
ImageNet-M10	$\mathcal{D}$ (original)	200	0.1	256	50, 100, 150
	$\widehat{\mathcal{D}}_{adv}$ (ℓ_2)	400	0.01	128	250
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_2, random class)	400	0.01	128	250
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_2, K-interp)	400	0.01	256	250
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_2, same class)	400	0.05	64	None
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_2, same subclass)	400	0.05	128	250
	$\widehat{\mathcal{D}}_{adv}$ (ℓ_∞)	400	0.001	256	None
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_∞)	400	0.1	128	None
CIFAR-10	$\mathcal{D}$ (original)	150	0.1	128	50, 100
	$\widehat{\mathcal{D}}_{adv}$ (ℓ_2)	400	0.05	256	250
	$\widehat{\mathcal{D}}_{univ}$ (ℓ_2)	400	0.05	256	None

of 100).[11] The use of a learning rate decay did not seem to have much impact, so we used the same learning rate throughout.

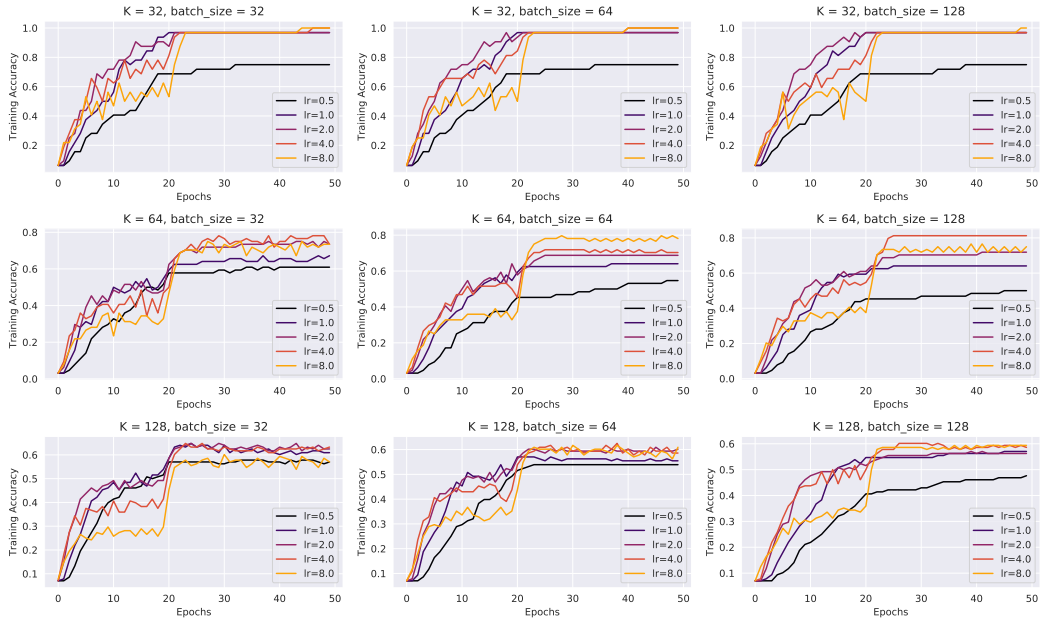


Figure 8: The learning dynamics of computing universal perturbation (epoch vs. training accuracy on the base set) are shown across a selection of learning rates, batch sizes, and three intermediate values of K.

[11]In this case K is much larger than batch size, fewer epochs suffice as the effective number of updates increases due to bounded batch size. In particular, for a fully universal base set (e.g. the entire training set), as few as five epochs seed to suffice.

C VISUALIZATIONS

We show visualizations of universal perturbations across different datasets (CIFAR-10, ImageNet-M, and ImageNet), architectures (VGG for Mixed10), and norm constraint (ℓ_∞ for Mixed10).

Normalization for visualization. For all visualizations, perturbations were rescaled as follows (for each input channel): (1) truncated to three standard deviations, (2) rescaled to $[-1, 1]$, and (3) shifted to mean 0.5 (to fit inside $[0, 1]$)

C.1 DIFFERENT DATASETS

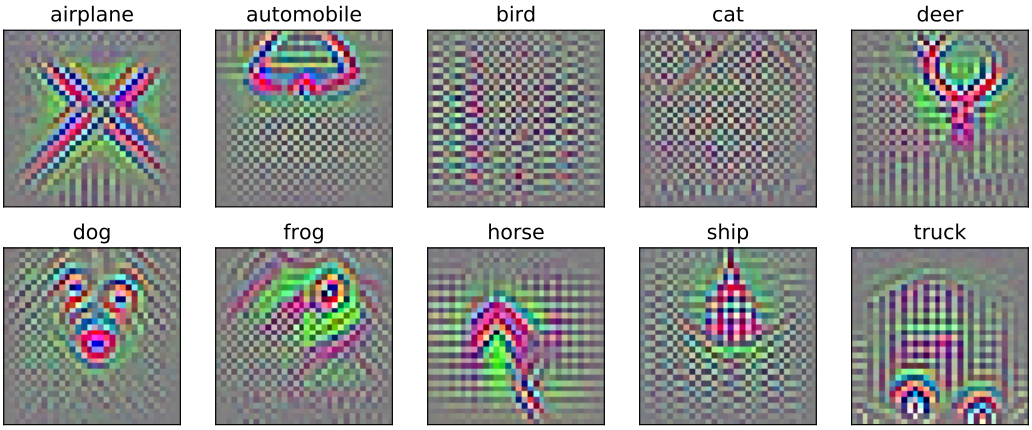


Figure 9: Visualization of ℓ_2 universal perturbations at $\epsilon = 1.0$ for a **CIFAR-10** model.

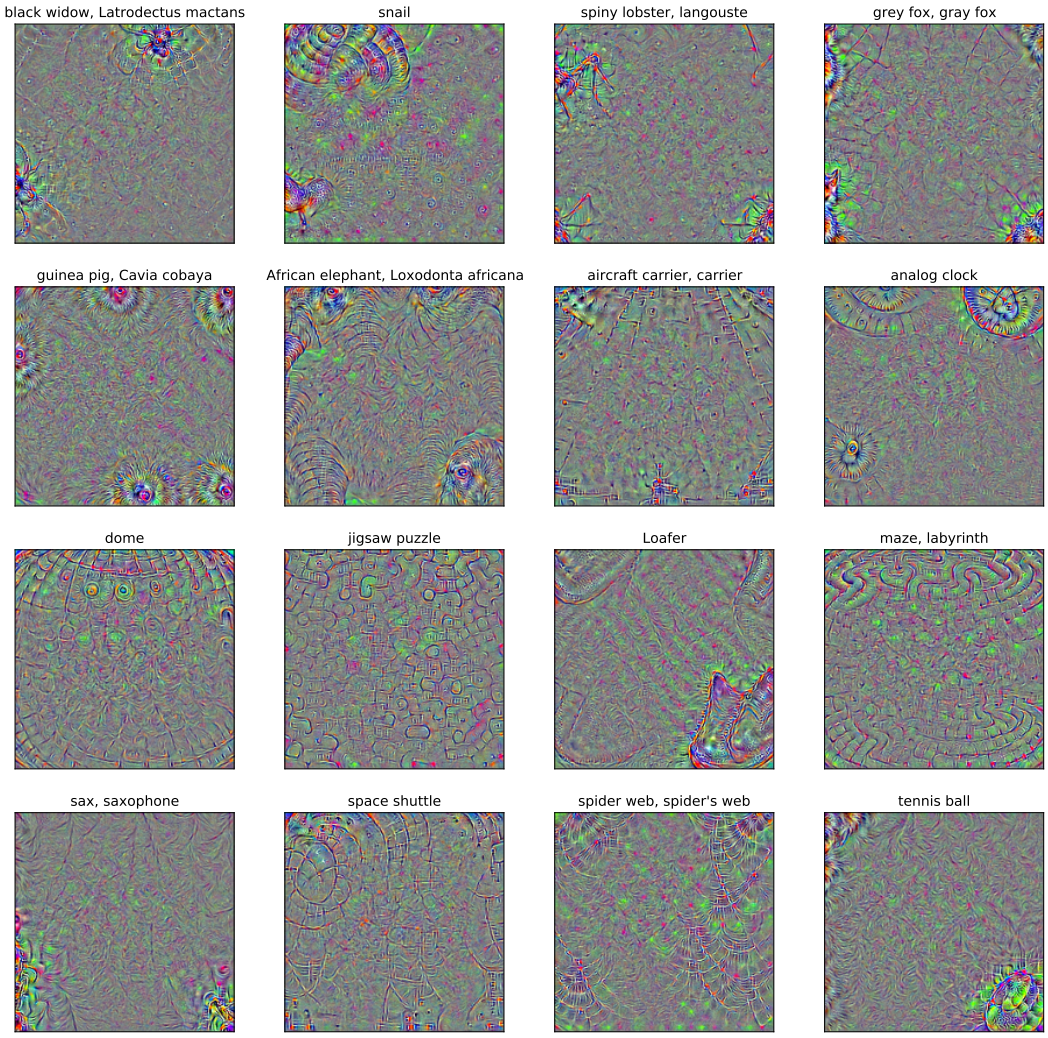


Figure 10: Visualization of ℓ_2 universal perturbations at $\epsilon = 6.0$ for a full **ImageNet** model. We hand-picked the classes with the most salient perturbations. Though not all perturbations are immediately identifiable as the corresponding classes, many of them are visually reasonable.

C.2 OTHER ARCHITECTURES

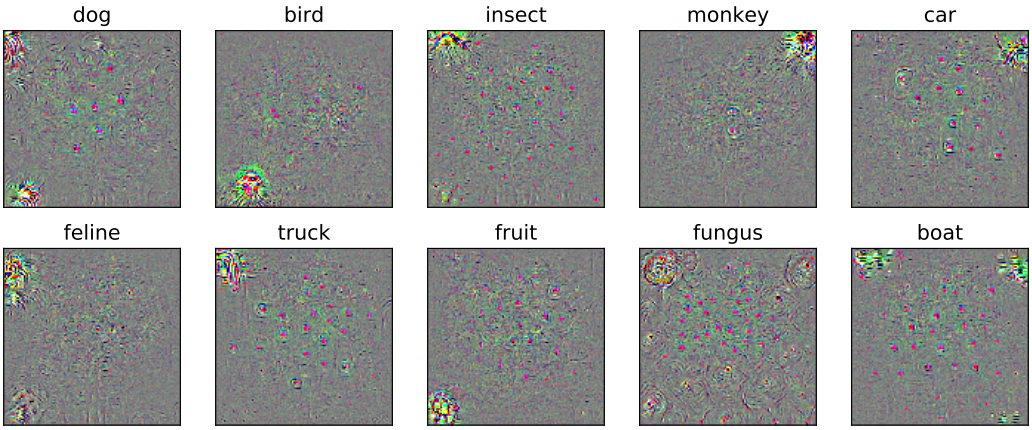


Figure 11: Visualization of ℓ_2 universal ($K = 512$) perturbations at $\epsilon = 6.0$ for ImageNet-M10, generated this time on a **VGG11** network Simonyan & Zisserman (2015).

C.3 ℓ_∞ PERTURBATIONS

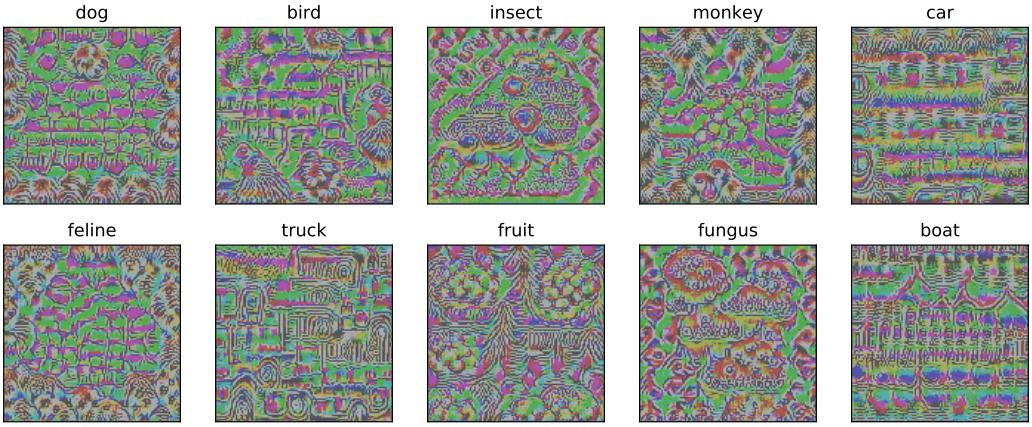


Figure 12: Visualization of ℓ_∞ universal perturbations at $\epsilon = 8/255$ for different target classes. We observe recognizable patterns for each class similar to those of ℓ_2 universal perturbations, with a more noticeable background texture.

D LOCALITY ANALYSIS

We show full results of randomized locality analysis on ImageNet-M10 in Figures 13 and 14. We did not perform the analysis on CIFAR-10 perturbations, as the perturbations are more "global" by nature (e.g. the semantic part seems to occupy a large fraction of the perturbation).

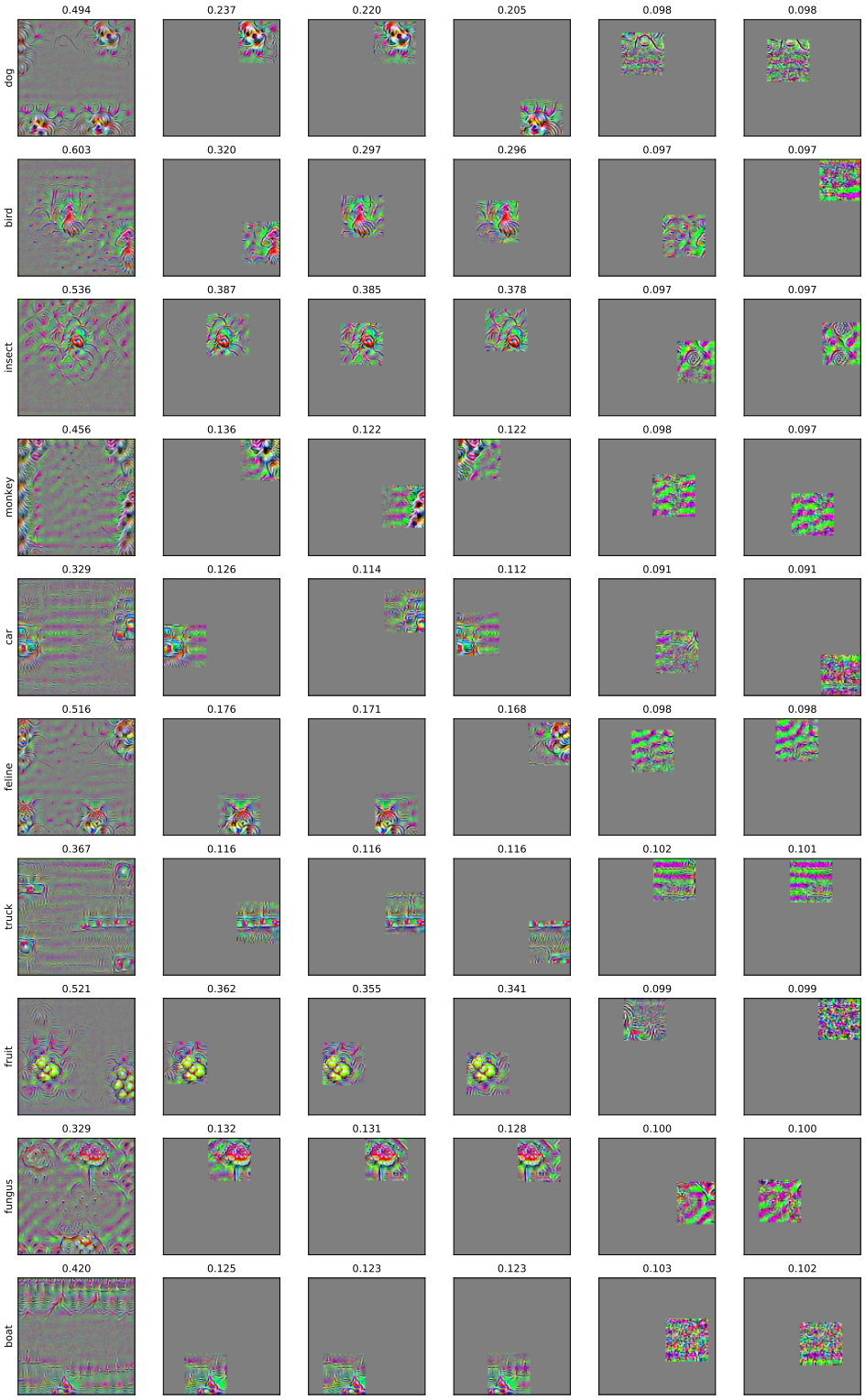


Figure 13: Locality analysis of ℓ_2 universal perturbations on Imagenet-M10. Each row shows the perturbation for the particular class, followed by three patches with highest ASR and two patches with lowest ASR; ASR is written above each patch.

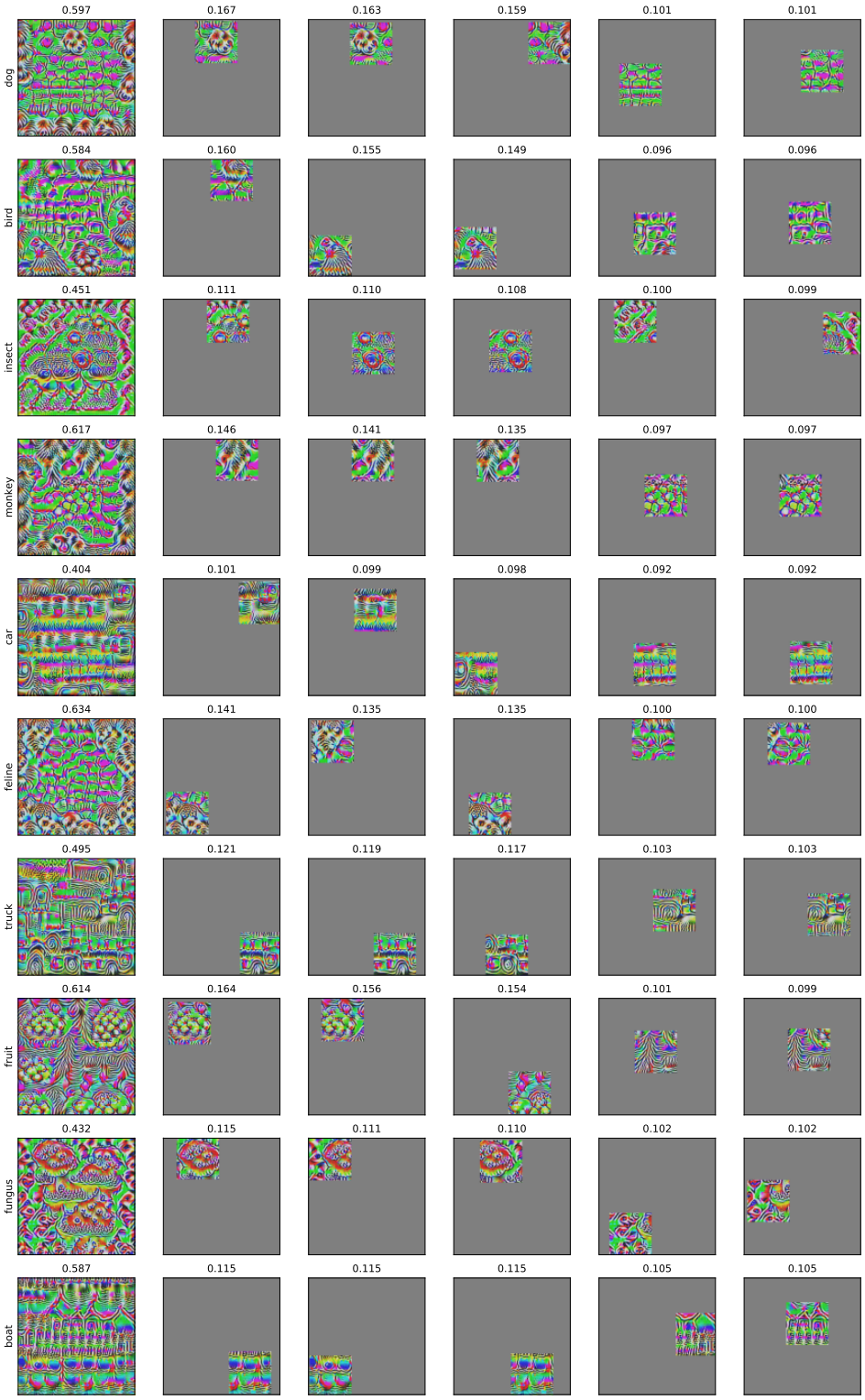


Figure 14: Locality analysis of ℓ_∞ universal perturbations on Imagenet-M10. Each row shows the perturbation for the particular class, followed by three patches with highest ASR and two patches with lowest ASR; ASR is written above each patch

E SPATIAL INVARIANCE

In the main paper, we saw a sample of translation invariance evaluation for perturbations on ImageNet-M10. Here, we include samples from evaluation for ℓ_2 perturbations on CIFAR-10, to show that the pheonemena is general.

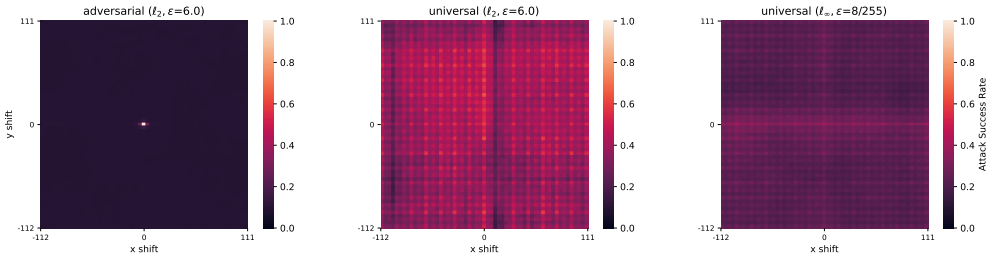


Figure 15: We evaluate attack success rates of translated adversarial and universal perturbations for different target classes over the test set on CIFAR-10. There is no subsampling of the grid here as the dimensions are much smaller. We used a subsampled grid of all the possible offsets with strides of four pixels. As before, the value at coordinate (i, j) represents the average attack success rate when the perturbations are shifted by i pixels along the x-axis and j pixels along the y-axis, with wrap around to preserve information. Evaluation of adversarial perturbation is only shown for one target as it is redundant.

F GENERALIZATION ON NON-ROBUST DATASETS

Here we collect results on generalization from non-robust features across different settings: ℓ_2, ℓ_∞ perturbations in ImageNet-M10 and ℓ_2 perturbations on on CIFAR-10. We did not conduct experiments on ImageNet-R as it is similar to ImageNet-R, only with the added complexity of unbalanced classes, and also not on ImageNet, as the constructions of the dataset is computationally too intensive.

We can see in table 5 that much lower generalization from $\widehat{\mathcal{D}}_{univ}$ compared to $\widehat{\mathcal{D}}_{adv}$ holds consistently.

Table 5: Generalization from constructed datasets.

Source Dataset	Perturbation Set	Constructed Dataset	Test Accuracy (%)
ImageNet-M10	$\ell_2, \varepsilon = 6.0$	$\widehat{\mathcal{D}}_{adv}$	74.5
ImageNet-M10	$\ell_2, \varepsilon = 3.0$	$\widehat{\mathcal{D}}_{adv}$	76.6
ImageNet-M10	$\ell_2, \varepsilon = 6.0$	$\widehat{\mathcal{D}}_{univ}$	23.2
ImageNet-M10	$\ell_\infty, \varepsilon = 8/255$	$\widehat{\mathcal{D}}_{adv}$	78.7
ImageNet-M10	$\ell_\infty, \varepsilon = 8/255$	$\widehat{\mathcal{D}}_{univ}$	26.5
CIFAR-10	$\ell_2, \varepsilon = 1.0$	$\widehat{\mathcal{D}}_{adv}$	64.3
CIFAR-10	$\ell_2, \varepsilon = 1.0$	$\widehat{\mathcal{D}}_{univ}$	23.3

G PERTURBATION DIVERSITY

Throughout our investigations, we observed that the computed universal perturbations seemed to lack in diversity. We observe in Figure 16 (first row) that generating multiple different universal perturbations for a given target leads to visually similar perturbations. [12]

[12] In the original work on universal perturbations, Moosavi-Dezfooli et al. (2017a) made the opposite interpretation that the perturbations generated independently are diverse due to their inner product being small. However, it is unclear whether inner product in the input space is the appropriate measure of their difference. Given the

To increase diversity, we generated universal perturbations with an additional penalty to encourage orthogonality of perturbed representations at different layers of the network.[13] Even with the added regularization, the resulting perturbations appear to be remain similar, but with changes in spatial arrangements of the semantic portions (second row).

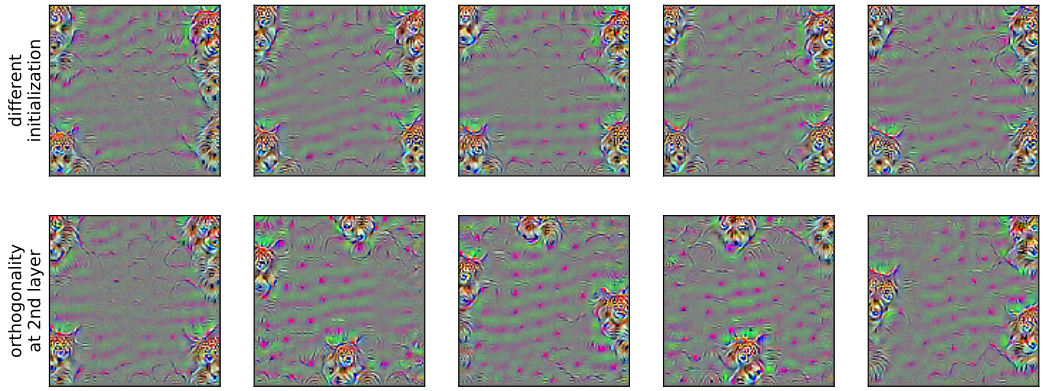


Figure 16: Visualization of ℓ_2 universal perturbations for class feline, computed using different initializations (first row), and an additional penalty to encourage diversity (second row).

One effective way we found to increase the diversity of these perturbations was to compute universal perturbations for specific subclasses of Mixed10. To do so, we trained a fine-grained model using Mixed10's subclass labels, and generated universal perturbations for each of the six subclasses per class. In Figure 17, we show the universal perturbations targeting individual subclasses of the bird class. Notably, we observe that new patterns, which seemingly resemble each distinct species of birds, indeed emerge in these subclass universal perturbations. However, note that the generation of these perturbations required access to more fine-grained labels. An interesting question is whether one can recover such diversity using just the coarse-grained model without the injection of additional metadata.

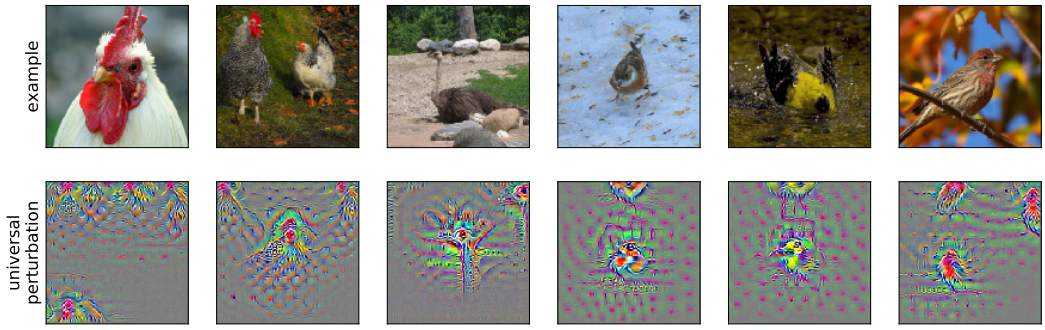


Figure 17: Visualization of ℓ_2 universal perturbations for six individual subclasses inside the *bird* class. These were generated on a *fine-grained* model trained on the subclass labels (60 in total). Observe the distinctive visual characteristics of each class, for instance the outline of a chicken (column 2), the neck of an ostrich (column 3), and the distinctive colors of different species (columns 3-5).

results of our signal analysis in Section 5, it appears more likely that different perturbations are highly similar in their (signal) content.

[13]There is an extensive survey of similar regularization techniques in Olah et al. (2017).