# Understanding CLIP Robustness

Yuri Galindo, Fabio A. Faria*
Institute of Science and Technology, Universidade Federal de São Paulo, São José dos Campos, Brazil
yurioliveiragalindo@gmail.com, ffaria@unifesp.br

## Abstract

*Neural networks show a lack of robustness under adverse conditions, such as dealing with new datasets/distributions and adversarial perturbations. Some works in the literature, based on experiments with Resnet models trained on Imagenet, elect possible culprits such as the vulnerability to high frequency disturbances and dependence on non-robust features. Contrastive Language-Image Pre-training (CLIP) has been proposed as a new learning procedure which has improved robustness to new distributions but low robustness to adversarial examples. Therefore, CLIP presents an ideal opportunity for measuring how robust features and frequency sensitivity are associated with robustness to data shift. In this sense, we measure the vulnerability of CLIP model to high frequency perturbations, and perform image generation and inpainting tasks for assessment of robust features. In the performed experiments, the CLIP model is shown to be more robust to higher frequency perturbations and less robust to lower frequency perturbations, indicating a higher dependence on features with lower frequency. Finally, the images generated by CLIP were of low quality, indicating a lack of robust features.*

## 1. Introduction

Deep Neural Networks achieve excellent classification results (over 90% accuracy in the Imagenet [15] dataset), even surpassing humans. However, these learning models do not perform well on new datasets and distributions, showing a considerable drop in performance [13].

Some works existing in literature have shown that these learning models are not robust in another adverse condition: the addition of specific perturbations (adversarial examples) [5]. Although these adversarial examples are mostly imperceptible by humans, they lead the models to make errors with high confidence score. Explaining how these models can surpass humans in some conditions and be so vulnerable in others is still an open research question.

An explanation for the existence of adversarial examples is that models learn non-robust features, i.e., information that is predictive but affected by small variations [7]. Models that are robust to adversarial attacks, on the other hand, rely on robust features, which are more aligned with human perception. Due to this alignment, robust models can be used for image synthesis tasks such as inpainting [12].

Yin *et al*. [14] have analyzed the robustness of Resnet models to different perturbations in terms of frequencies. The experiments showed that models which are robust to adversarial examples rely more on low-frequency information, and are more vulnerable to low-frequency perturbations such as contrast changes. Natural distribution shift such as the change of datasets was not studied.

These works presented interesting hypotheses, however the experiments were carried out using models based on Resnet [6], trained on the Imagenet [11] dataset. Therefore, it is not possible to know whether their findings hold for the general case of deep learning models trained on different ways.

Recently, the CLIP learning procedure (Contrastive Language-Image Pre-training) [10] has been proposed to achieve better robustness to natural distribution shifts. A CLIP model (ViT-B/32) is able to match the Resnet50 model performance on Imagenet without ever seeing any training examples, in a zero-shot fashion. This model has a non-convolutional archicteture, relying instead on Transformers [9], and is trained on a dataset of 400 million image and text pairs, 400 times more examples than Imagenet [11]. It is also trained with a contrastive training objective, employing a text model in order to find the matching text and image pairs.

Due to its differences from previous models and its superior robustness to distribution shift, this CLIP model is the most suitable for verifying how the existing hypotheses for model robustness hold up in this case. Geirhos *et al*. [4] showed that CLIP is closer to human behavior in various aspects, including having less texture bias. Therefore, in this paper, we want to understand what it has changed in terms

*The authors are grateful to the São Paulo Research Foundation (FAPESP – grants #2017/25908-6 and #2018/23908-1) and to NVIDIA for donating a Titan V GPU used in the experiments.

of frequency vulnerability and feature robustness.

## 2. Experimental Methodology

In this section, we performed the experiments using CLIP ViT-B/32 as a zero-shot classifier. It received the images as usual and text prompts in the format "a photo of a *class*", for each possible class of the dataset. For the Imagenet100 dataset, in which each class corresponds to various names, we selected only the first given name. This procedure matches Resnet50 accuracy.

### 2.1. Frequency sensitivity

This experiment measures the accuracy of the model for perturbations of different frequencies. We build the "Fourier heat map" described in [14].

The Fourier basis vector $(i, j)$ is the image $U_{i,j}$ with norm 1 such that the discrete Fourier transform of the image, $DFT(U_{i,j})$, has the element $(i, j)$ different from 0 and all other elements equal to 0 [1]. For each image $X$ of the validation set and each Fourier basis vector $(i, j)$, we obtain a perturbed version $\tilde{X}_{i,j}$ of the image for that specific frequency vector. This is described on the Equation 1, in which $r$ is randomly chosen as $-1$ or 1 and $v$ is the norm of the perturbation. As in [14], we used $v = 4.0$ for CIFAR10 and $v = 15.7$ for Imagenet.

$$\tilde{X}_{i,j} = X + rvU_{i,j} \tag{1}$$

We then measure the accuracy obtained by the model averaged over the images perturbed with this Fourier basis vector $(i, j)$, which corresponds to the point $(i, j)$ of our graph, centered on the origin. By obtaining the accuracy for the perturbation of all Fourier basis vectors, we can observe how the model behaves for perturbations of the whole frequency spectrum.

We performed this experiment on the test set of the CIFAR10 dataset [8], consisting of $10,000$ images with dimensions of $32 \times 32$ divided in 10 categories, and on the Imagenet100 validation set, consisting of a subset of the Imagenet dataset composed of 100 randomly selected categories [1]. The images have dimensions of $224 \times 224$, with 50 examples per category. For the Imagenet experiments, we build a heatmap of size $63 \times 63$, discarding the higher frequencies. This is done to replicate the heatmaps produced in the original paper [14] and better observe the frequencies used by the models. Experiments were built upon an existing implementation [2].

### 2.2. Robust Features

This experiment is qualitative, and aims to verify visually the quality of the features learned by the model. We follow the experimental methodology for inpainting and image generation described in [12]. Experiments were built upon the official implementation [3]

For image generation, first we obtain the mean and standard deviation of the pixels of the images of the class of interest $y$. Then, we sample pixels assuming a normal distribution with the observed mean and deviation. Now, we apply projected gradient descent (PGD) in order to maximize the probability assigned by the model to the modified image $x'$ while keeping the norm of the perturbation as less than $\epsilon = 40$.

The Equation 2 shows the minimization objective, in which $\mathcal{L}$ is the Cross Entropy loss and $x_0$ is the image sampled by the class distribution $\mathcal{G}_y$. $C$ refers to the classification model, which takes an image as input and outputs class probabilities.

$$x = \underset{s.t. \|x' - x_0\|_2 \leq \varepsilon}{\arg\min_{x'}} \mathcal{L}\left(C(x'), y\right), \quad x_0 \sim \mathcal{G}_y \tag{2}$$

For the inpainting experiment, we randomly assign a region of the image of size $60 \times 60$, and substitute it by the mean of the pixels of the region, over each channel. Now, we apply PGD in order to maximize the probability assigned by the model while minimizing the pixels altered outside the region and keeping the norm of the perturbation as less than $\epsilon = 21.6$.

Equation 3 shows the minimization objective, in which $\lambda$ is a constant, $m$ is the mask matrix which is equal to 1 on the affected region and 0 elsewhere. $\odot$ corresponds to element-wise multiplication. Other symbols have the same meaning as in Equation 2. In our experiments, $\lambda = 10$.

$$x_I = \underset{s.t. \|x' - x_0\|_2 \leq \varepsilon}{\arg\min_{x'}} \mathcal{L}\left(C(x'), y\right) + \lambda \left\|(x - x') \odot (1 - m)\right\|_2 \tag{3}$$

## 3. Results and Discussion

This section presents the experiments performed in this paper and discusses the achieved results.

### 3.1. Frequency sensitivity

On both CIFAR10 and Imagenet100 datasets, CLIP ViT-B/32 displayed a different behavior from what was previously observed in the Resnet50 model. We found that CLIP is less robust than the Resnet50 to perturbations in the low-frequency domain, and more robust to perturbations in the highest frequencies. This can mean that CLIP relies more on low-frequency information, when compared to Resnet models [14].

In Figure 1, we can see the error rates of both models for the CIFAR10 dataset. We can see that the Resnet isn't affected by the lowest frequency perturbations, shown in the

[1] kaggle.com/ambityga/imagenet100/metadata
[2] github.com/gatheluck/FourierHeatmap
[3] github.com/MadryLab/robustness_applications

blue circle in the center. However, it is affected by perturbations of the highest frequencies, shown by the red on the extremities. CLIP, on the other hand, shows higher vulnerability on the lowest frequencies and less vulnerability on the highest frequencies.

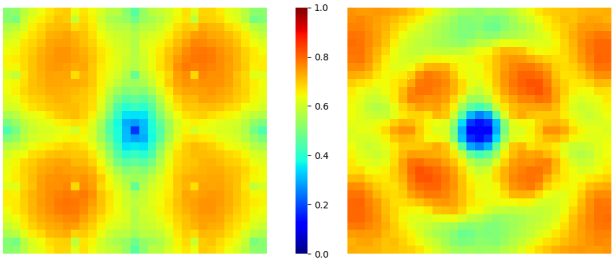


Figure 1. Error rates of the CLIP ViT-B/32 (left) and Resnet50 (right) models for perturbations over the frequency spectrum on the CIFAR10 dataset.

Figure 2 displays the error rates on Imagenet100, in which we can see more clearly the low frequency bias of CLIP model. CLIP model maintains high error rates in the lowest frequencies up to a point, which then decreases for higher frequencies. The Resnet50, on the other hand, shows low error for the lowest frequencies (displayed on a blue circle in the center) and high error on the other frequency regions, including close to the extremities – the highest considered frequencies.

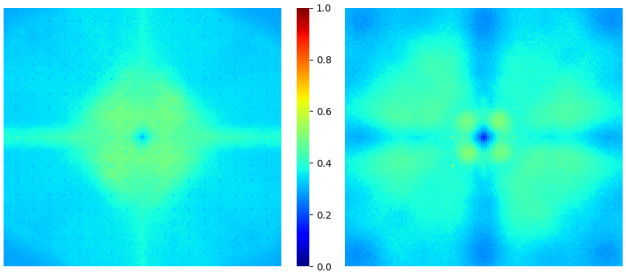


Figure 2. Error rates of the CLIP ViT-B/32 (left) and Resnet50 (right) models for perturbations of various frequencies on the Imagenet100 dataset. The images are cropped with $63 \times 63$ size, centered at the lowest frequency on the Fourier domain.

### 3.2. Robust features

In the inpainting experiment, the robust model described in [12] was capable of finding meaningful reconstructions. The images generated with the robust model were semantically similar to the images before corruption, and were perceptually plausible to humans even in the case of mistakes.

The CLIP model, however, does not reconstruct the corrupted images. As showcased in Fig. 3, CLIP model barely changes the corrupted patches. The alterations are limited to almost imperceptible squiggles. To the model, however, these alterations are highly accurate and even capable of changing a wrong classification to the correct class. The inpainting loss (3), which is based on the probability assigned by the model for each image, decreased from 3.8 to 2e-4.


Figure 3. Inpainting task, performed using the CLIP ViT-B/32 model.

The image generation experiment is illustrated on Fig. 4, and obtained similar results. The images generated by the robust models in [12], while not realistic, showcased perceptually relevant features such as feathers, eyes, fur and noses. The images generated by CLIP model, on the other hand, do not display such features. Any similarity to the original class is due to the pixel sampling process, with the final result looking like incomprehensible noise. To the model, however, these images have high likelihood of pertaining to the correct class. The generation loss (2), based on the probability assigned by the model, decreased from 3.5 on the raw sampled pixels to 3e-5 on the final images.

## 4. Discussion

In the first experiment, we observed differences between the Fourier heatmap of a Resnet and CLIP models. We observed that in relation to the Resnet model, the CLIP model has lower robustness to low frequency perturbations and higher robustness to high frequency perturbations. This can mean that this CLIP model depends on lower frequencies than the Resnet model, and as so is more affected by perturbations in these frequencies [14].

This indicates that robustness to natural distribution shifts might be associated to dependency on lower frequency information and higher robustness in high frequency domains. Since the CLIP model is robust to high frequency perturbations and is still vulnerable to adversarial attacks [2, 3], this is also further evidence that adversarial attacks are not exclusively a high frequency phe-

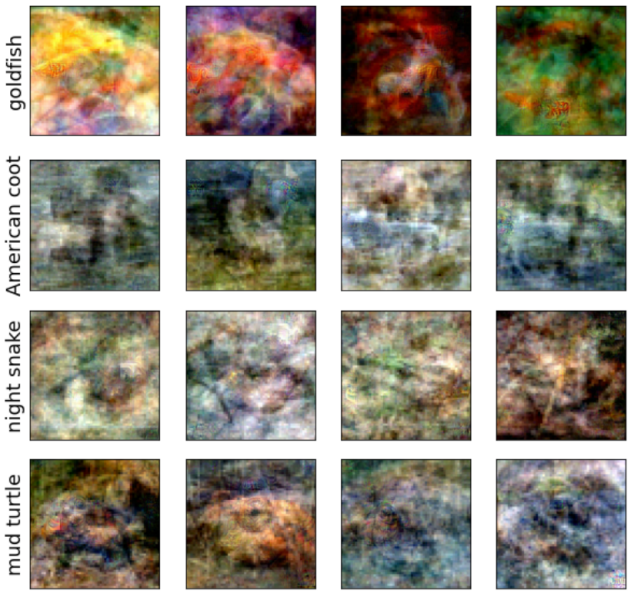


Figure 4. Images generated using the CLIP ViT-B/32 model.

nomenon [14].

In the second experiment, we did not find evidence of robust features. This indicates that robust features are not a requirement for robustness to natural distribution shift, and may be exclusively related to adversarial robustness.

## 5. Conclusion

We replicated experiments for the CLIP ViT-B/32 model, observing how behaviors related to different aspects of robustness look like for this model, which obtains record robustness for dataset shift [10]. We observed that the frequency vulnerability of this CLIP model differs from Resnet models trained in Imagenet, being biased toward lower frequencies. We also observed that it does not possess robust features and human-aligned gradients, leading to the generation of images which fool the model but aren't perceptually plausible for humans.

Future works will aim at expanding these experiments to other models that exhibit robust behavior, such as other versions of CLIP models, in order to understand if low frequency bias and lack of robust features are general trends. Another interesting avenue is to perform an ablation study of CLIP models, observing the frequency vulnerabilities and robustness of the ablated models. Namely, investigating the impacts of the contrastive procedure, training data amount and distribution, and transformer architecture.

## References

[1] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986. 2

[2] Stanislav Fort. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, Jan 2021. 3

[3] Stanislav Fort. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations, March 2021. 3

[4] 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*, 2021. 1

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. 1

[8] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 2

[9] Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. *CoRR*, abs/2008.07772, 2020. 1

[10] 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, 2021. 1, 4

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 1

[12] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Computer vision with a single (robust) classifier. *CoRR*, abs/1906.09453, 2019. 1, 2, 3

[13] Achal Taori, Rohan aernd Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc., 2020. 1

[14] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *CoRR*, abs/1906.08988, 2019. 1, 2, 3, 4

[15] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *CoRR*, abs/2106.04560, 2021. 1