Extended abstract

Simon Théorêt

Abstract

Data augmentation is a key component to training robust models and to prevent overfitting in computer vision. The Fourier perspective [Yin et al., 2020] gave us a better understanding of the process behind such improvements by showing us the tradeoffs between protecting against corruptions in low frequency and high frequency and the limitations of data augmentation. These insights also raise questions about the bias of gradients towards low frequencies: Does the optimizer and the architecture of a model bias the model to rely on low frequency features. In this work, we investigate the impact of Gaussian data augmentation and adversarial training on a different set of architectures and optimizers. We provide an hypothesis that gradients of adversarially trained models and models trained on Gaussian augmented data are naturally biased towards low frequency features, as they contain more relevant information for classification. To test our hypothesis, we provide an experimental protocol for testing our hypothesis against different architectures and optimizers by computing the accuracy of the trained models on images containing either high frequency features or low frequency features.

1 Introduction

The Fourier perspective introduced by Yin et al. [2020] paved the way for exploring the Fourier space of images' features and how these features are used by models. This approach was used to classify features in two categories: high frequency and low frequency features. The first category includes features such as images' texture, and the second is related to the countours and shapes in images, as stated in Krishnamachari et al. [2023]. Although high frequency features are often invisible to the human eye, the Fourier perspective showed that these features could be successfully used by Convolutional Neural Networks (CNN) in image classification. However, high frequency features are not robust, as shown by Zhang and Zhu [2019] and yet models are often biased toward using these features. On the contrary, low frequency features, such as shape, are often the preferred features of adversarially trained neural networks or networks trained with a Gaussian augmented dataset. The Fourier perspective article Yin et al. [2020] limited their experiments to the ResNet architecture. The lack of empirical research using different architectures and optimizers prompts the question:

Does the architecture and the optimizer influence the bias toward low frequency features in adversarially trained models and models trained on Gaussian augmented datasets?

2 Preliminaries

We use the following notations: $\mathcal{F} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ denotes the discrete Fourier transform (DFT) of an image and we omit the dimensions of the channels, as every channel is treated independently of the other channels. For an image of size $N \times N$, the discrete Fourier transform is defined as

$$\mathcal{F}(k,l) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i,j) e^{-i2\pi\left(\frac{ki}{N} + \frac{lj}{N}\right)},$$

where $f(i,j)$ represents the pixel at position i, j of an image. When we visualize the Fourier spectrum, we always shift the low frequency components to the center of the spectrum. Unless stated otherwise, we only show the magnitude of the Fourier spectrum, not the phase.

To filter components of an image based on their frequency, we use the methodology of Yin et al. [2020]. We set to 0 every point in the Fourier spectrum that is not in the square of width B centered at the highest (lowest) frequency. We then apply the inverse DFT to recover the original image, with the low (high) frequency components filtered out.

Our Gaussian augmentation method follows the methodology of Yin et al. [2020]. We assume that pixels take values in the range $[0, 1]$. Pixel values are always clipped to this range. Gaussian data augmentation with parameters σ is defined as the following operation: i.i.d. Gaussian noise $\mathcal{N}(0, \tilde{\sigma}^2)$, is applied at each iteration and at every pixel. The value of $\tilde{\sigma}^2$ is chosen uniformly at random from $[0, \sigma]$.

3 Problem statement and related works

Our goal is to determine whether optimizers and architecture influence the bias in feature selection of models trained with a Gaussian augmented dataset or adversarially trained. Previous works have tried to formalize the Fourier sensitivity of CNN, as explored by Krishnamachari et al. [2023], and gave experimental insights regarding the tuning of models towards certain frequencies. This has been explored in the works of Krishnamachari et al. [2023], Geirhos et al. [2022], Yin et al. [2020], Mo et al. [2022]. Notably, the works of Park and Kim [2022] demonstrated that multi-headed self-attentions, such as ViT (vision transformers) models, reduce high frequency signals, while CNN amplify them.

4 Proposed experimental protocol

We are planning on using two different architectures for our experiments: the ALL-CNN architecture mostly consisting of stacked convolution layers as described by Springenberg et al. [2015], and the mobileViT architecture, which introduce some transformer modules into a CNN architecture, as described by Mehta and Rastegari [2022]. We opted not to include pure ViT models, due to their high computational demands.

Our experiments consist of training a total of 12 models on the MNIST dataset and evaluating the performance on the validation set with only high frequency features or low frequency features. For both the CNN and ViT architectures, each are trained on the default MNIST dataset, adversarially trained on the MNIST dataset and trained on the MNIST dataset with Gaussian augmentation. This procedure is repeated with two different optimizers. The stochastic gradient descent with momentum and AdamW optimizers will be used in training, as these optimizers were used in the original training of ALL-CNN and MobileViT models.

To accomplish these experiments, we have access to Google Colab Pro account and an NVIDIA GTX 3060 GPU.

References

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022.

K. Krishnamachari, S.-K. Ng, and C.-S. Foo. Fourier sensitivity and regularization of computer vision models, 2023.

S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022.

Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang. When adversarial training meets vision transformers: Recipes from training to architecture, 2022.

N. Park and S. Kim. How do vision transformers work?, 2022.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net, 2015.

D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision, 2020.

T. Zhang and Z. Zhu. Interpreting adversarially trained convolutional neural networks, 2019.