Image Synthesis With a Single Robustness Classifier

Shining Yang
Minjung Lee
Abhijith Tammanagari

Duke
PRATT SCHOOL of ENGINEERING

Abstract

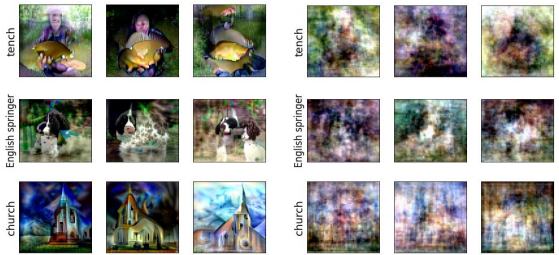
In this project, we performed several image synthesis tasks (image generation, image inpainting, super resolution, image-to-image translation, and sketch-to-image translation) using a pre-trained adversarially robust classifier released by Santurkar et al (2019). Due to the limitation in computation resource, we used ImageNette dataset, which is a subset of 10 easily classified classes from Imagenet. In line with the paper, we found that the features learned by a basic classifier are sufficient for all these tasks, provided this classifier is adversarially robust. The idea for these augmentations, is taking the original adversarial model and using L2 FGSM attacks in various means to achieve the synthesis tasks.

Image Generation

Our image generation is based on maximizing the class score of the desired class using the pre-trained robust model. To generate a sample of class y, we sample a seed and minimize the loss L of label y:

$$x = \underset{\|x'-x_0\|_2 \le \varepsilon}{\arg\min}\ \mathcal{L}(x', y), \qquad x_0 \sim \mathcal{G}_y,$$

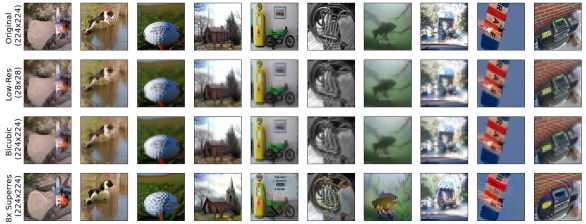
The images generated this way look plausible and diverse.
In comparison, we also we also did the same things on a standard non-robust model (in this case, a standard resnet50 model). And we found that the non-robust model performed poorly on image generation and didn't show clear visual patterns. The results are presented below.



a) Image generation with robust classifier

b) Image generation with non-robust classifier

Super-Resolution



In the case of super resolution, we utilized the standard L2 FGSM attack on the images with the standard loss function. The images that were the input to the L2 FGSM trained images are the downsampled and then upsampled original images. The outcomes can be seen above. This method seems to perform pretty well across all the classes besides parachute where it distorts the image drastically. One interesting take away is with the french horn which is initially a black and white image which gains color as well as increased resolution. On the other hand with the fish underwater, we see that the fish becomes very well defined but the surrounding water becomes more distorted. Lastly, in the case of the chainsaw we notice that the chainsaw is indeed super resolute but we see that the fallen tree in the background takes on the shape of a chainsaw blade.

Image-to-Image Translation

Instead of loading in their pre-trained A2B model, we stick to the ImageNet model and want to see how the model generalize on different task. We used standard loss function with the only preprocessing being changing the labels passed into L2 FGSM to the translation target. Our results show that this robust model performs well on image-to-image translation once the original and targeted classes are provided. It again demonstrates that the adversarial robust model learns meaningful features and is generalizable.

Image Inpainting

The goal of inpainting is to recover the missing pixels using robust models in a manner that is perceptually plausible with respect to the rest of the image. Given a robust classifier pre-trained on uncorrupted data, and a corrupted image x with label y, we solve

$$x_I = \underset{x'}{\arg\min}\ \mathcal{L}(x', y) + \lambda||(x - x') \odot (1 - m)||_2$$

where L is the cross-entropy loss, m is the mask, and λ is a constant (which is 10). We used L2 FGSM attack on ImageNette dataset, and for x', we picked random seeds as 1) pure random noise and 2) sample from class-conditional seed distribution. The following results were obtained using pure random noise, and the noise of the results was severe.

(a) Image inpainting using robust models with random sample from pure random noise

Therefore, we used class-conditional seed distribution instead, and finally got the following results.

(b) Image inpainting using robust models with random sample from class-conditional seed distribution

Although these results do not perfectly match the original images, but it's not too difficult to recognize the images, and most of them look plausible.

References

Santurkar et al., "Image Synthesis with a Single (Robust) Classifier," (https://arxiv.org/pdf/1906.09453.pdf)
MadriLab, "Robustness Applications", (https://github.com/MadryLab/robustness_applications?)
Fastai, "ImageNette", (https://github.com/fastai/imagenette)