Exploring Image Synthesis Using a Robust Classifier

Francesco Mastrocinque
Owen Gibson

Duke
PRATT SCHOOL of ENGINEERING

Introduction

By utilizing a robust classifier, we show that a projected gradient descent (PGD) adversarial attack is able to manipulate input images driven towards a target class with unique salient and perceptible class features. We used this model to accomplish several image synthesis tasks, such as image generation, inpainting, super-resolution, and image-to-image translation. For image generation, we explored using several seed distributions as an input and compared the performance to using a standard non-robust classifier.

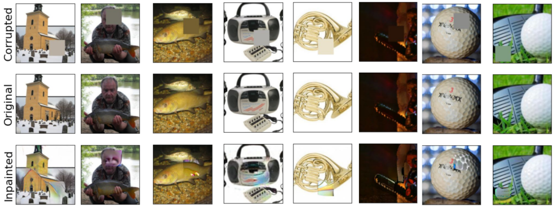
Methodology

Dataset: The Imagenette dataset was used in this study and consisted of 10 classes: tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, and parachute. The dataset contains a total of 13,394 images that were divided into training and validation sets using a 70/30 split. The images were resized to 224 x 224 pixels.

Models: Model architectures used as classifiers in this work are based off of ResNet. For the non-robust classifier, a pre-trained architecture was used with default ImageNet weights. The robust classifier was provided by the Madry Lab as a ResNet50 model that was trained using a robust optimization objective on the full ImageNet dataset, where the worst case loss is minimized over a specific perturbation set:

$$\mathbb{E}_{(x,y)\sim D}\left[\max_{\delta\in\Delta}\mathcal{L}(x+\delta,\, y)\right]$$

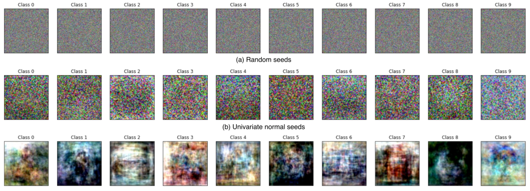
Inpainting



Main takeaway: Input images contained a 60 x 60 pixel patch to hide some features of the image. For most classes, inpainted images looked natural while other inpainted images did not match the surrounding context. Inpainted images were synthesized using the following constraint:

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

Image Generation

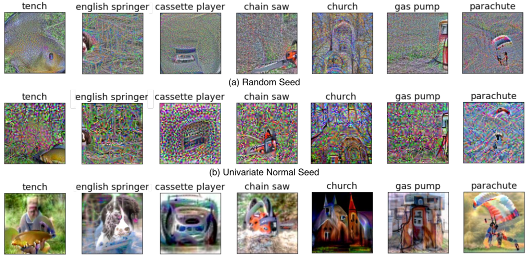


(a) Random seeds

(b) Univariate normal seeds

(c) Multivariate normal seeds

Using Various Seed Distributions and a Robust Classifier:

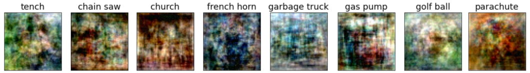


(a) Random Seed

(b) Univariate Normal Seed

(c) Multivariate Normal Seed

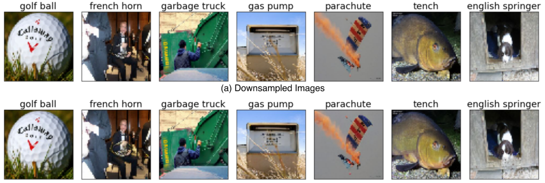
Using a Multivariate Normal Seed Distribution and a Non-Robust Classifier:



Main takeaway: We explored the quality of image generation using random, univariate normal, and multivariate normal seed distributions using a robust classifier. Our results show that multivariate normal distribution seeds perform significantly better compared to the other seed image techniques when used on a robust classifier. On the other hand, using a multivariate normal distribution seed with a non-robust classifier yielded results that underperformed the robust classifier. Generated images were synthesized using the following constraint:

$$x = \arg\min_{\|x'-x_0\|_2\le\varepsilon} \mathcal{L}(x', y), \qquad x_0 \sim G_y$$

Super-Resolution



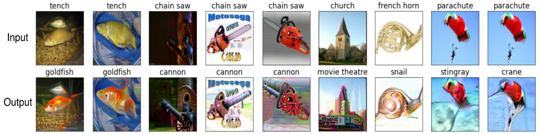
(a) Downsampled Images

(b) Super-resolution Images

Main takeaway: Performing image super-resolution on downsampled images achieved higher resolution images that were natural looking without any significant distortion in most instances. Class specific features were enhanced while surrounding background features remained relatively unchanged. Super-resolution images were synthesized using the following constraint:

$$x = \arg\min_{\|x'-\uparrow(x_L)\|_2<\varepsilon} \mathcal{L}(x', y)$$

Image-to-Image Translation



Main takeaway: We show that image-to-image translation can be achieved using a general (naive) robust classifier trained on an entire dataset, as opposed to training a new model for each image translation pair. Our technique allows for a much higher throughput post-training, as any image from a given class can be translated to any target domain. Translation images were synthesized using the following constraint:

$$x = \arg\min_{\|x'-x_0\|_2\le\varepsilon} \mathcal{L}(x', y), \qquad x_0 \sim G_y$$

Conclusions

Our findings demonstrate that the adversarially robust classifier greatly outperforms the benchmark classifier for image generation tasks. Furthermore, our work shows that robust classifiers can be used as generative models that can accomplish many tasks involving image synthesis with results that contain compelling salient features.

References

Santurkar et al., "Image Synthesis with a Single (Robust) Classifier," (https://arxiv.org/pdf/1906.09453.pdf)