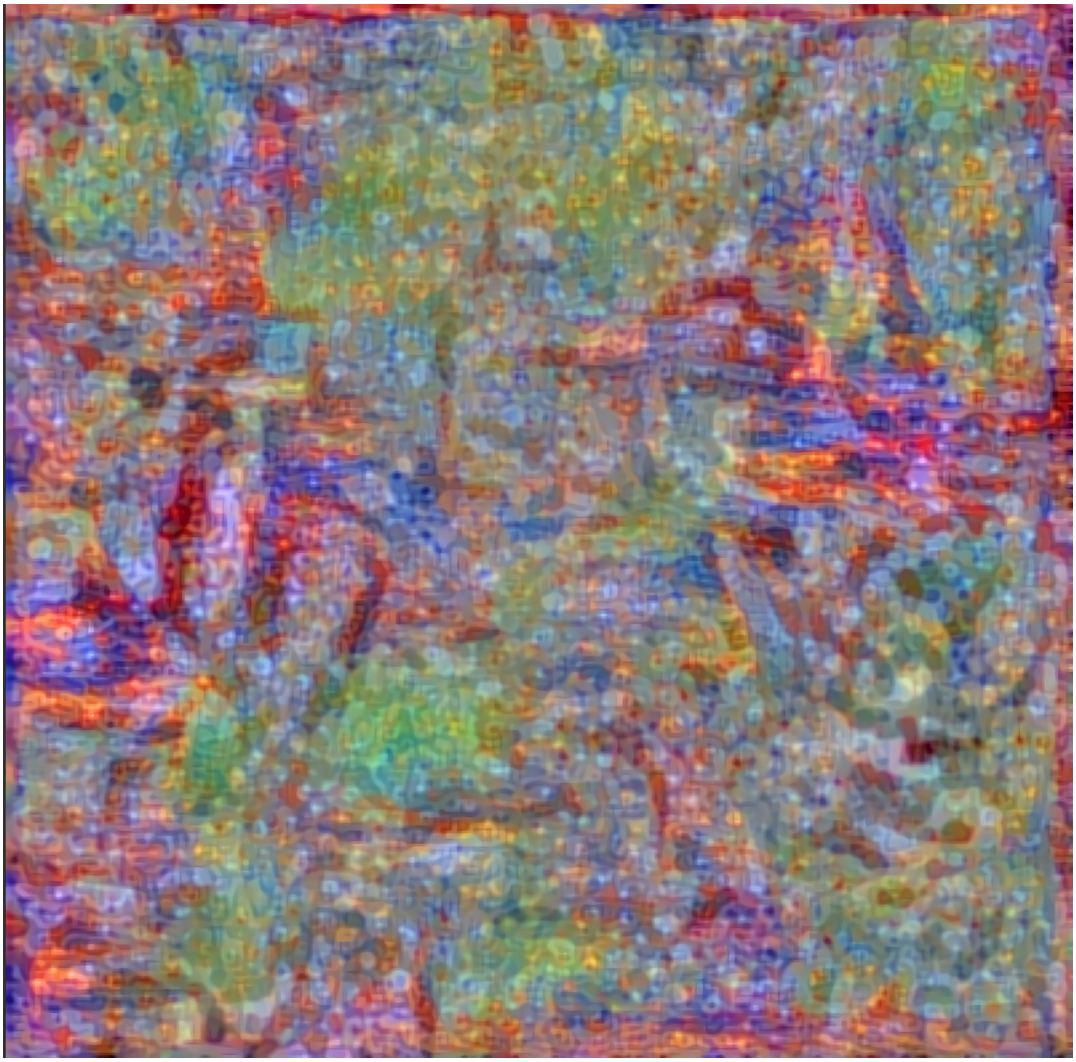
ART AND MACHINE LEARNING
CMU 2019 SPRING
PROJECT 1

ADVERSARIAL ATTACK OF IMAGES

EXAMINE THE BIASES THAT CAUSE MISLABELING IMAGES



Boxiang Lyu / Chuning Yang / Edgar Xi / Tianjun Ma / Zhiyu Bai

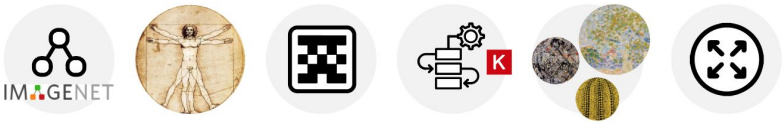
CONCEPT STATEMENT

This project is inspired by the adversarial attacks in machine learning. An adversarial attack consists of subtle modifications an original image that are almost undetectable to the human eye. We want to examine the biases that cause machines to mislabel certain images.

We managed to succeed using several attack algorithms, most notably Fast Signed Gradient Method [1] and the Carlini and Wagner L2 attack [3] to generate these slight modifications.

Style transfer and DeepDream are then used to modify these pictures so that they look more than merely noisy images to represent what these images represent in an artistic setting.

TECHNIQUES



- Dataset used: ImageNet [2]. The attacks are based on the pre-trained ResNet50 mode, and the style transfers are based on the pre-trained VGG-19 model. Both models come from Keras library.
- “Vitruvian Man” by Leonardo Da Vinci was chosen as our base image for the attacks. This painting embodies what Da Vinci believed to be the ideal human proportions, and we believe that it best represents the essence of a human during the Renaissance.
- We used Foolbox and CleverHans, two libraries focused on adversarial attacks. Fast Signed Gradient Method (FSGM for short) [1] and Carlini and Wagner L2 attack [3] were chosen among the various attacks provided by these methods.
- The code for style transfer and DeepDream are largely taken from Keras examples and the tutorial given in class.
- We noted a close resemblance between the attack and the styles of Georges Seurat, Jackson Pollock, and Yayoi Kusama. Their paintings are used in style transfer and inspire our rendering.
- We used an out-of-box machine learning solution provided by letsenhance.io, which allows us to upscale our images with minimal blurring and loss of details.

PROCESS

Feb. 8th: Brainstormed for the assignment.

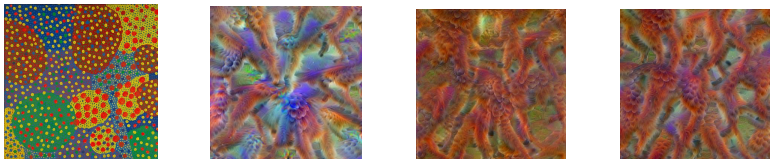
Feb. 15th: Started experimenting with different adversarial attack methods in CleverHans and Foolbox.

Feb. 16th: Started writing the outline for the report. Discovered that several algorithms are often buggy and unable to work. Decided to focus on using FSGM and Carlini and Wagner for the attacks.

Feb. 17th:

1. Finished generating the adversarial attack using FSGM and Carlini and Wagner.
2. Started experimenting with StyleTransfer using paintings by Jackson Pollock and Georges Seurat. Left: Jackson Pollock, “Autumn Rhythm”. Right: Pittsburgh skyline in the style of Jackson Pollock.

3. Experimented with DeepDream following the style of Yayoi Kusama

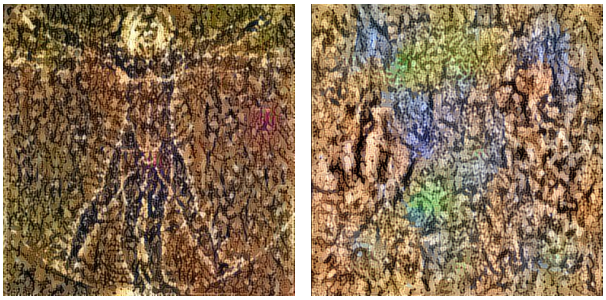


4. Applied style transfer to the attacks using the style of Jackson Pollock and Georges Seurat.

Feb. 18th: Finished experimentations with StyleTransfer. Used letsenhance.io to upscale the images to an acceptable resolution. Finished writing the report.

RESULT

The image in the title page is the attack that causes the vitruvian man to be classified as velvet in the style of Georges Seurat. In addition to that image, we also include the attack and another attack that causes the image to be classified as roundworm in the style of Pollock.



REFLECTIONS

We chose our final selection of images based on the technical difficulties of generating the images, the meaning behind these images, as well as how pleasing they appear. We believe they represent the difference between man and different objects, at least according to machines.

We are satisfied with our idea and techniques for this project because we have successfully visualized the algorithm's bias on images of different categories. We found that the algorithm seems to have a special interest in certain part of the difference image that after doing StyleTransfer, a few parts of the result is in different color compared to the overall image. StyleTransfer adds green in many places of the resulted image while the original style does not contain large packs of green color. This is true for many of our StyleTransfer generated images.

StyleTransfer generated images have too many details at first glance. It is only after applying smoothing algorithms that the image is better suited for human to view. However, the original image might be the true vision of our algorithm.

CONTRIBUTIONS

- **Boxiang Lyu:** implemented and generated the adversarial attacks. Helped choose the target paintings for style transfer and write the report.
- **Edgar Xi:** helped choose base images for style transfer and programmatically generate images.
- **Tianjun Ma:** generated final works for style transfer on the difference between Vitruvian man and velvet. Applied out-of-box upscale algorithms on the outputs and helped on writing the report.
- **Chuning Yang:** wrote part of the final report, raised ideas on style transfer target pictures, reflected our findings on the final result.
- **Zhiyu Bai:** based on the adversarial attack image and the base image, generated the Kusama style imageries through 3 iterations with DeepDream. Upscaled the output. Helped on writing the report.

CODE https://tianjunm.github.io/artml-s19/

REFERENCES

[1] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014 https://arxiv.org/abs/1412.6572

[2] Deng, J. and Dong, et al. ImageNet: A Large-Scale Hierarchical Image Database, CVPR09:2009 http://www.image-net.org/papers/imagenet_cvpr09.bib

[3] Carlini, N and Wagner, D. Towards Evaluating the Robustness of Neural Networks. arXiv preprint arXiv:1608.04644, 2016 https://arxiv.org/abs/1608.04644