Art-attack! On style transfers with textures, label categories and adversarial examples

Vinay Uday Prabhu, John Whaley
{vinay,john}@unify.id

Abstract:

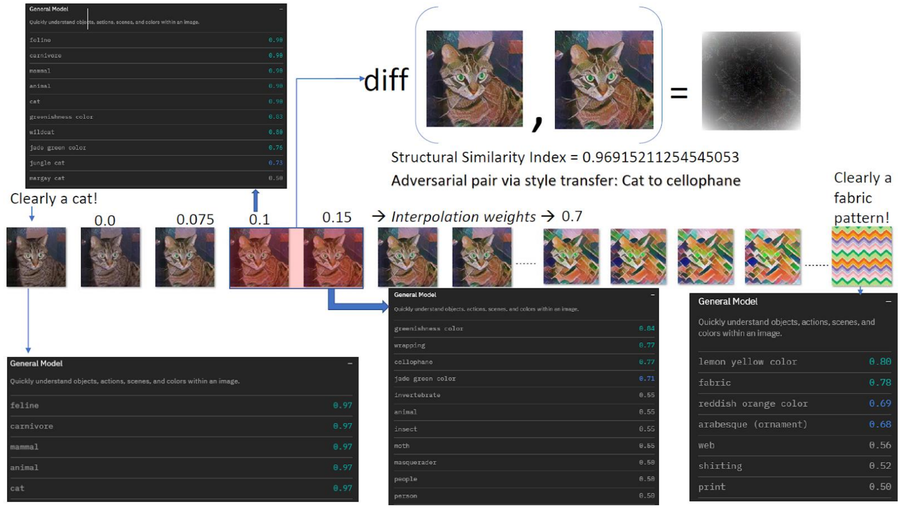
In this short paper, we describe an experiment that entailed using style transferred images to target mis-classification in the context of a specific popular commercial off-the-shelf (OTS) API.

The test images were drawn from the Kaggle ***'Dogs and Cats***' dataset and the style image was drawn from the ***Describable Textures Dataset (DTD).*** The style transferred images achieved adversarial attack success rates of **97.5 % (195 out of 200).**

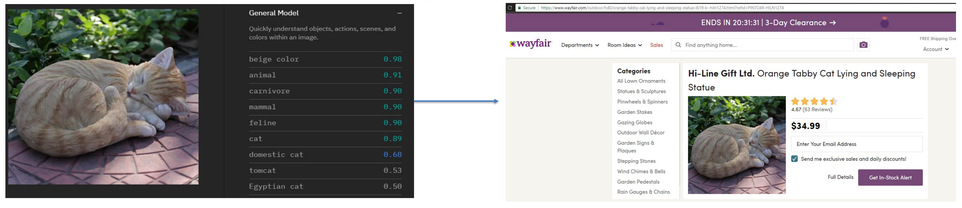
The goal of this paper is **not** to proclaim a new black-box attack recipe or to berate the commercial API we have used, but to merely highlight the following observations.

- The first is regarding the generation of a pair of 'close-by' images using style-transfer that are indistinguishable to the human eye but that elicit very different predictions from a classifier.
- Secondly, on account of the fact that the 'raw image' that is ***adversarially perturbed*** is not necessarily a naturally occurring image and is a style-transferred image itself, we believe this should necessarily instigate a conversation over what constitutes a true image category/class and admit to skepticism if the incorrect response of the classifier would indeed qualify as a mis-classification.
- Lastly, irrespective of what emerges from the above point raised, we would like to highlight the potency of using ***interpolated style transfer as a recipe of generating mutually adversarial pairs*** that can be used for model regularization as well as generating 'challenging' co-class images as inputs into training pipelines for 'embedding deep-nets' trained on triplet-loss cost functions.

Interpolated style transfer to generate adversarial examples?



Are the 'intermediate images' even qualify to be a cat?



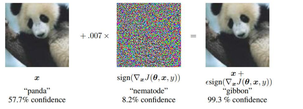


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Let $\boldsymbol{\theta}$ be the parameters of a model, $\boldsymbol{x}$ the input to the model, $\mathbf{y}$ the targets associated with $\boldsymbol{x}$ (for machine learning tasks that have targets) and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ be the cost used to train the neural network. We can linearize the cost function around the current value of $\boldsymbol{\theta}$, obtaining an optimal max-norm constrained perturbation of

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right).$$

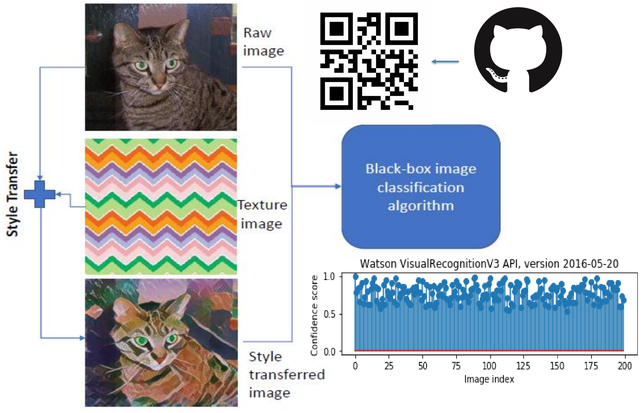
$$x \in X, y \in Y$$
$$f : X \to Y$$
$$\delta : (X \times X) \to \mathbb{R}_+$$
$$\delta(x, \tilde{x}) \leq \varepsilon, f(x) \neq f(\tilde{x})$$

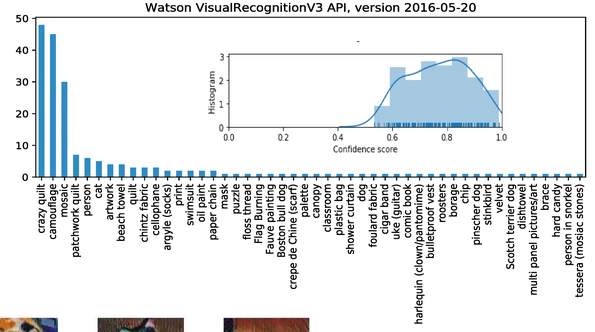
Mutually adversarial pairs for a given classifier and a distance metric

Questions:

1: What is the 'true' label?
2: Do we encounter inescapability of encountering mutually adversarial pairs during interpolated style transfer?

Experiments and results: Github link





Correctly classified images (5)

Incorrectly classified examples