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

Vinay Uday Prabhu, John Whaley
UnifyID Inc
San Francisco, CA
{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 deepnets' trained on triplet-loss cost functions.*

1. Introduction

In this short paper, we look at the problem of generating adversarial examples targeting an off-the-shelf black box image classifier. The goal of this dissemination is not to proclaim a novel black-box attack but to instigate a conversation of some issues that we feel are more fundamental, such as what constitutes an image category or a label, and what constitutes a real world image versus an artistic rendering of it and the role that texture plays, especially in the context of commercial off-the-shelf (OTS) APIs. Specifically, we use the `Watson Visual-Recognition- V3 API, version 2016-05-20` API for all the results shown here.

Let's begin by focusing on Fig4. What we see is the image of a cat getting style-transferred into a 'pattern-style-image' using the *arbitrary image stylization* [2] project which comes pre-packaged in with the *Magenta* project [1] for different interpolation weights monotonically increasing from 0 to 1 (from the left to the right). As seen, with the raw image (interpolation weight ($w = 0$)) or style-transferred images with low interpolation weights (up until interpolation weight $w = 0.1$) as inputs, the commercial OTS classification API has, as expected *correctly* classified the image as a `cat` with high confidence scores (0.97 to 0.99). When we increase the interpolation weight slightly to $w = 0.15$, we see a dramatic change in the inference landscape. The top guessed classes dramatically change from `feline, cat and carnivore` to `cellophane, moth and invertebrate`. We now note that from the perspective of the image with $w = 0.1$, the $w = 0.15$ is an adversarially perturbed variant. While the two images share a structural similarity (SSIM) index [4] of 0.969 (translating to $\infty - norm$ distance of 0.125), we see that the labels are indeed different.The local texture based features that the classifier might have learned, has perhaps coaxed it into making erroneous classification, while the image still clearly looks like that of cat. This brings us to the natural question as to whether the artistically style transferred (non-naturally occurring) image (with $w = 0.1$) *deserve* to be classified as a `cat` in the first place. This is akin to another related question of what is the normative expected class when the input is a real world figurine rather than an animate being, which

[1] https://github.com/tensorflow/magenta/tree/master/magenta/models/arbitrary_image_stylization

brings us to the Fig 1. Here, we see the input image[2] literally being that of an artistic cat figurine that results in a high confidence classification of being categorized a `cat` with high confidence score (0.89). The rest of the short paper is organized as follows. In Section 2, we describe the experimentation procedure. In Section 3, we share the results and conclude the paper in Section 4.

2. Procedure

It is indeed legitimate to ask if the example discussed in section-1 was idiosyncratically chosen. In order to assuage those concerns, we did the following experiment. The main querying point behind the experiment was as follows: *Is it indeed the case that images that are style transferred with low interpolation weights do result in mis-classifications?* For this, we extracted 200 randomly chosen cat images from the `Kaggle Dogs and Cats` dataset[3]. We resized all of them to size 299 x 299 and style transferred each one of them using the same style image extracted from the DTD dataset[1] using the style transfer algorithm detailed in [2]. Fig 2 showcases this with a specific example. In order to ensure that the images still looked 'cat-like' the interpolation weight was set to a *low* value of 0.125. One can sift through all the raw images and the style transferred images as a gif animation shared via the following link: `goo.gl/ejYxPw`. Now, both the raw images and the style transferred images were classified using the `Watson Visual Recognition- V3 API, version 2016-05-20` API. The `Accept-Language header string` that sets the language of the output class names was set to `en`. The `owners query array` was set to the default option (`IBM`). The `classifier_ids` was set to `default` that required no training and would *Return classes from thousands of general tags.*. The `threshold query` parameter that represents the minimum score a class must have to be returned was set to 0.5.
The results are covered in the forthcoming section.

3. Results

In Fig 3, we see the counts of the most probable classes that the API returned. As seen, the top 4 classes that encompassed more than 50% of the test images were `crazy quilt`, `camouflage`, `mosaic` and `patchwork`. In Fig 5, we see the scores as well as the histogram of scores related to the 200 classification trials. As seen, we have an overwhelmingly large number of cases where the mis-classifications were made with high confidence scores associated. In Fig 6, we see the 5 images that the API classified correctly and in Fig 7, we see randomly chosen 10 examples of style transferred images that were classified incorrectly.

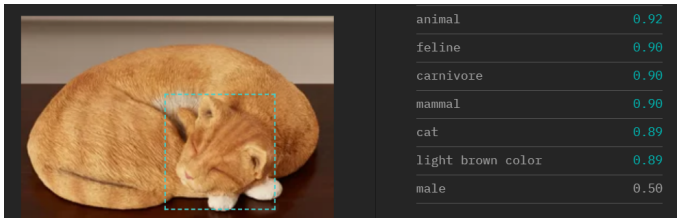

Figure 1. Cat figurine: Art and not a cat?

4. Conclusion and Future Work

Due to limitations of API usage for free-tier users, we could not extend the experiment for larger datasets, which is our immediate goal. Besides this, another question that we would like to explore is the choice of the style image. We selected an image for the texture dataset on account of 2 reasons. The first being that a pre-trained style transfer model was readily available. The second reason was based on a hunch that texture, would be in fact be the right aspect of the image to *perturb* to induce a mis-classification. As stated in the abstract, our intention is not to proclaim a new black-box attack or to berate the commercial API.
Besides showcasing the potential of looking at style transfer as an adversarial example generating technique, we also wanted to draw attention to the inherent fuzziness that surrounds the definition of what constitutes an image class/category or 'tags' in the case of such APIs and what entails an image mis-classification. The API that we used describes[4] the technology as: `Watson Visual Recognition's category-specific models enable you to analyze images for scenes, objects, faces, colors, foods, and other content.` With regards to the specific API documentation[5], it was stated that with upon usage with `Pre-trained models` (in lieu of a custom trained classifier), the API `Returns classes from thousands of general tags.` On the concluding note, we would like to remark that we also ascertained the efficacy of these style-transferred based black-box attacks using the universal adversarial images for different Deep-nets from [3] as the style image, the results of which we plan to disseminate in the full version of this work.

[2]The image was sourced from `https://www.wayfair.com/keyword.php?keyword=outdoor+cat+sculptures`. We find this specific shopping portal to be an especially good source of such figurine art examples

[3]https://www.kaggle.com/c/dogs-vs-cats

[4]`https://www.ibm.com/watson/services/visual-recognition/index.html`

[5]`https://www.ibm.com/watson/developercloud/visual-recognition/api/v3/curl.html?curl-classify`

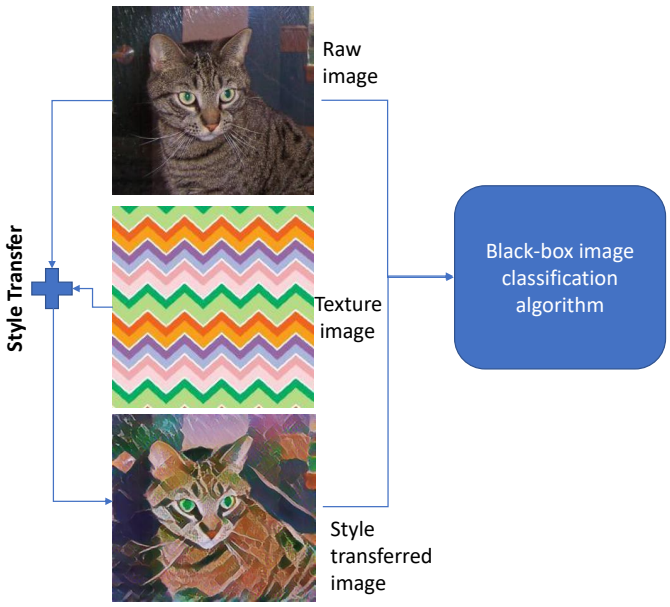

Figure 2. The concept

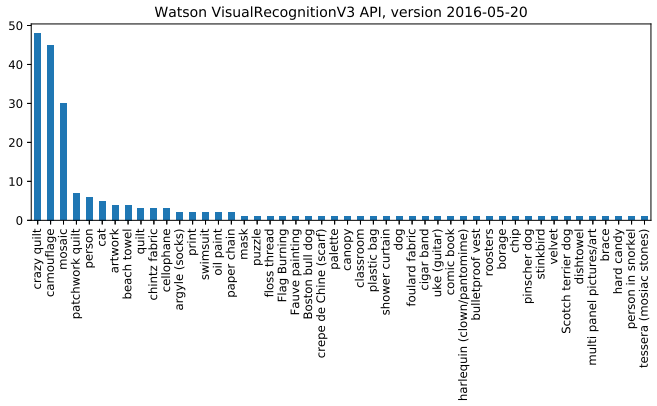

Figure 3. Counts of top guessed classes for the 200 images

References

[1] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3606–3613. IEEE, 2014. 2

[2] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 1, 2

[3] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. 2

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

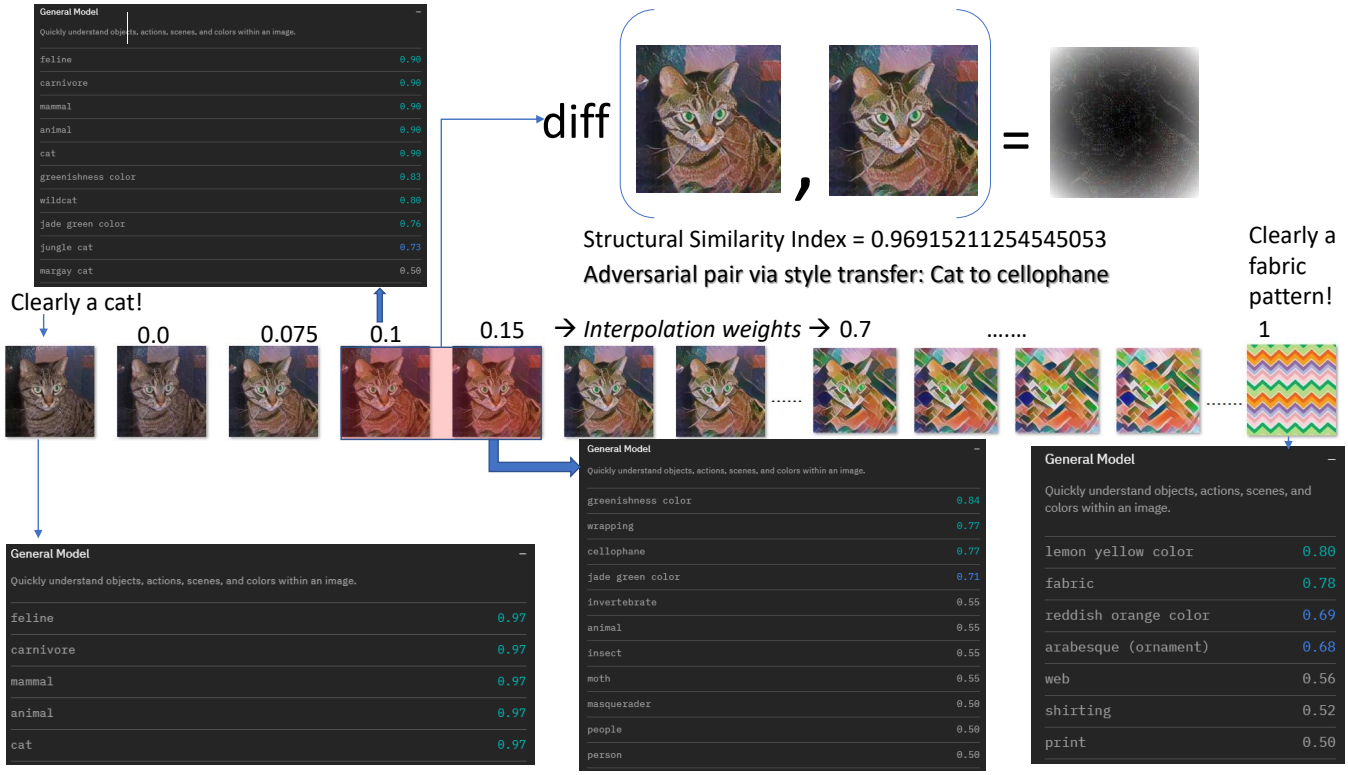

Figure 4. Example of the cat images getting style transferred into a pattern.

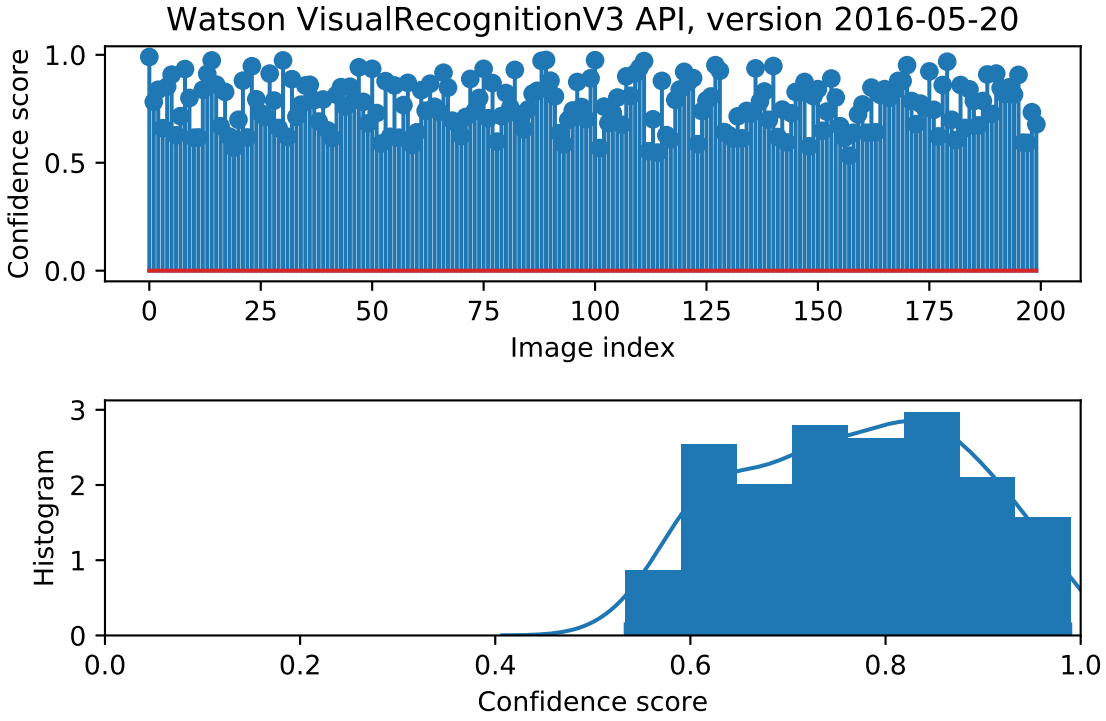

Figure 5. Scores obtained on the 200 test images

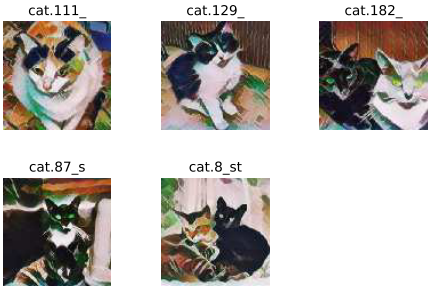

Figure 6. The 5 style transferred images that were predicted correctly

Figure 7. 10 randomly selected example images of incorrectly classified images