Data Augmentation with Style Transfer for Training Object Detection Models on Stylized Art

Oliver Brady[1]

Abstract

Modern object detection models, often trained on hundreds of thousands of photos and bounding box labels, are incredibly powerful tools for locating objects in an image. When the input image, like a piece of stylized art, is far from the distribution of photos the model was trained on, however, these models fare far worse. One of the may challenges of applying machine learning models to art is that the datasets are far smaller. To increase the accuracy of object detection models on abstract art, I take the underlying photograph dataset that a popular model, YOLOv4, was trained on, and use a style transfer model, CycleGAN, to create a stylized labeled dataset for training. I then train YOLOv4 on the new synthetic art dataset. The results from the a brief training on the augmented dataset show for common labels, like person, the new model likely outperforms the standard YOLOv4 on a dataset of Cezanne paintings. This is an encouraging sign of the validity of this technique as an approach for improving object detection in art.

1. Introduction

Object detection models are incredibly useful to art scholars for labeling and categorizing works for databases, and analyzing macro trends across many works and throughout time. With a good object detection model, an art scholar could track the presence and frequency of a certain object, say a skull, across a century or more of art works. Current object detection models that have been trained on photographs do not transfer very well to art, especially stylized art. Object detection models trained wholly on artwork do not preset a viable solution because art datasets are so much smaller than photograph datasets. In addition, labeled art datasets are incredibly hard to come by.

The goal of this project is to use the weights of a pre-trained style transfer model, called CycleGAN, to convert a labeled photo database into Cezanne-styled images. Then by fine tuning the YOLOv4 object detection algorithm on this new stylized dataset, YOLOv4 will learn to better detect objects in real Cezanne works.

2. Related work

2.1. Object Detection

Object detection is the process of finding and identifying objects in an image. Currently, the top models are deep convolutional neural networks (CNNs) that are trained on millions of images (Girshick, 2015). Many of these models transform object classification engines into object detection by having a sliding box move across the image looking for objects. The model in this project, YOLOv4, takes a slightly different approach in that it only looks at each image once, hence the name You Only Look Once (YOLO) (Bochkovskiy et al., 2020).

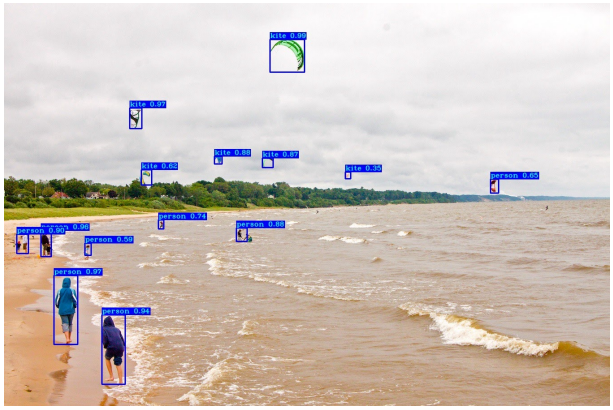

Figure 1. A example output of YOLOv4

2.2. Generative Adversarial Networks

Generative Adversarial Networks (GANs), invented in 2014, are a class of machine learning models where two neural networks, a generator and a discriminator, compete (Goodfellow et al., 2014). In the simplest case, the generator takes a noise vector as an input and attempts to generate an image.

[1]Stanford University

The discriminator looks at the underlying dataset and attempts to discern if the image is real or fake. The generator cannot see the underlying dataset, but gradually learns to fool the discriminator. When the model is done training, the result is generator weights that can generate an image similar to the underlying distribution of the image dataset.

2.3. Image Translation and Style Transfer

Image translation is a class of computer vision tasks to translate an input image into a corresponding output image. Traditionally, this task has been tackled by building each part of the translation pipeline by hand and manually extracting features (Efros & Freeman, 2001). I recent years, however, a seminal paper, pix2pix, showed that with a paired dataset a GAN could take an image instead of a noise vector as an input, and give the image from dataset one the style from dataset two and produce that as an output (Isola et al., 2017). A paired dataset means two datasets with different styles but each image is paired with one in the other dataset.

A further breakthrough, CycleGAN, pioneered the use of unpaired datasets for image translation (Zhu et al., 2017). CycleGAN has two sets of generators and discriminators, one for each direction of translation, and cycle consistency over the two sets of GANs is enforced. Therefore, there is no need for paired images.

Figure 2. A example output style transfer with CycleGAN

3. Dataset

3.1. Train

The training dataset was the COCO 2014 dataset that the YOLOv4 model was already trained on. I took the first 10,000 images from the 2014 train dataset. That dataset came with labels, so the Cezanne-style weights were applied to each image, and the labels were the scaled to fit the new Cezanne-like COCO dataset.

3.2. Test

This problem is slightly unique in that the train and test set are not from the same underlying distribution. It is not important how the object detection model fares on fake Cezanne images, the only metric that matters is how it performs on real ones. I could not use the real Cezanne dataset for training because we do not have labeled Cezanne images. Thus, I ended up with a training set of fake-Cezanne's and a test set of real ones. I downloaded the Cezanne paintings from the CycleGAN github. The CycleGAN team downloaded their dataset from Wikiart.org and scaled everything to 216x216 pixels. The scaling was done by shrinking each side no matter the dimensions instead of cropping, so the images are slightly distorted (Zhu et al., 2017).

4. Methods

4.1. Coco Image to Cezanne Style

CycleGAN makes the weights for all of their models publicly available on their GitHub. For this step I took each of the 10,000 images that I had selected from the COCO database, and ran them through the CycleGAN photo to Cezanne weights. CycleGAN's weights automatically scaled the images to 216x216 pixels.

Figure 3. One example of Cezanne style transfer on an image from the COCO dataset

4.2. Converting Coco Labels to YOLOv4 Standard

Labels for the COCO dataset are the JSON files that describe the dimensions and locations of the bounding boxes on each image, and category of object that they represent. COCO coordinates are in absolute dimensions, meaning they give pixel location for corners, while YOLOv4 labels are relative, meaning everything is between 0 and 1, given as a percentage of the dimensions of the image. Therefore, once COCO labels were converted to YOLO standard, they could be applied to the scaled CycleGAN output, because the boxes were relative.

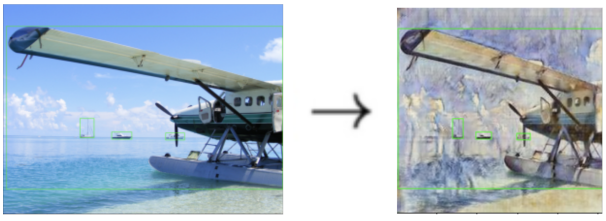

Figure 4. Example of an image from the COCO dataset that has bounding boxes (light green) re-scaled

4.3. Fine Tuning YOLOv4

To begin training, I loaded the YOLOv4 weights that had been trained on thousands of images. I then trained on my new augmented Cezanne styled dataset. Due to timing and computational constraints, I only trained for three hours, and was not able to train YOLOv4 until completion.

5. Results

First I generated prediction with the standard YOLOv4 weights some labels on real Cezanne's. I wanted a baseline before I started training to see if my training had improved the model at all. Some, like Figure 5 were quite impressive, but those were the outliers.

Figure 5. YOLOv4 Prediction on Cezanne before augmented data training. Predicts all people correctly

Much more common were outputs like Figure 6. The Cezanne images are just too different than the photos YOLO is used to, and time and time again it misidentified simple images.

After training, I ran the new model weights on my test set of real Cezanne images. Some notable outputs that provide key insights into the model are displayed in Figures 7 through 11.

6. Discussion

For this section when I refer to "old model or weights" I mean the standard YOLOv4, and "new model or weights" for my model trained on the augmented dataset.

Looking first at Figure 7, we can see that training on the Cezanne weights has improved the ability of the model to detect people in Cezanne paintings. The old prediction just

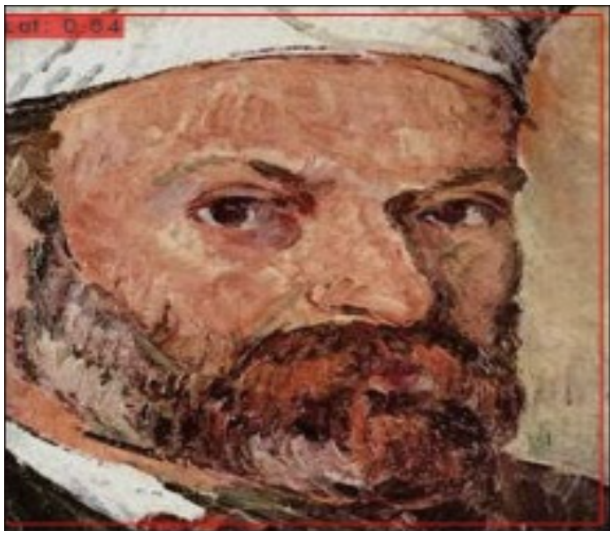

Figure 6. YOLOv4 Prediction on Cezanne before augmented data training. Labels man as cat.

Figure 7. Old Prediction vs. New Prediction

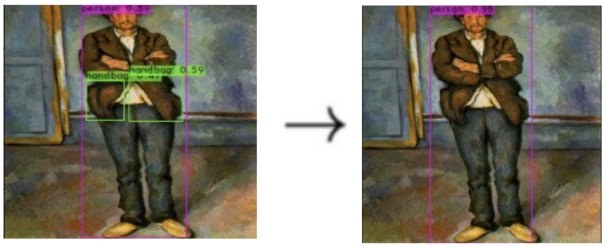

Figure 8. Old Prediction vs. New Prediction

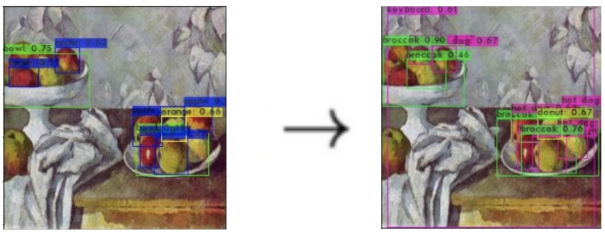

Figure 9. Old Prediction vs. New Prediction

Figure 10. Old Prediction vs. New Prediction

Figure 11. New Prediction

gets the bottles and the apple over the man's shoulder, while the new prediction is able to label the man. This is a very encouraging sign for the usefulness of this technique. Note that the new model does not predict the apple like the old model does. This is because the new model is far worse with labels it has not see very many times, to be expanded upon in a moment.

Figure 8 shows an old prediction on an image of a man, who is labeled correctly, and the predictions for two handbags where there clearly are none. The new model, however, strips the photo of the incorrect handbag labels. Additionally, the old model predicted a person with 0.89 confidence, while the new model is able to increase that confidence to 0.98.

When viewing an object that it has not seen very may times, the new model has very poor accuracy. Figure 10 is a prime example of this issue. The old model gets every fruit correct, while the new model sees where the objects are, but misidentifies them. This is likely due to the fact that the augmented training dataset was small. The model may have only seen one or two apples and they may have had poor Cezanne style transfers. The CycleGAN weights are not that consistent, and if they produce a Cezanne-like object that looks completely unrealistic, and is also quite rare in the dataset, it will trip up the model. These labels are an estimation of ground truth, not ground truth exactly, because there is uncertainty regarding the quality of the style transfer on each object.

7. Conclusions and Future Work

There are many potential improvement to this project, but none are as critical as a quantitative evaluation metric. While the COCO dataset was labeled, all of my contributions to the project were unsupervised, and I never had a labeled set of Cezanne paintings. Thus, the inspection and discussion of my results was qualitative. I could have easily missed some key trends in my outputs; additionally, if my model is a improvement for people recognition, it is unclear by how much.

Another improvement would be a larger augmented dataset and more training time. That could help with the fact that the model flounders when it is not detecting people.

One last improvement would involve a close inspection of the YOLO classes. In Figure 11 the model classifies the table as a keyboard. These art datasets exists in specific time periods and often with unique objects. Cezanne never would have painted a keyboard. Pruning or augmenting the classes could improve detection.

The results of this model show encouraging signs for the potential of this technique, but do not yet provide any con-

clusive results. While it seems that the new model is better at detecting people, and worse at most other objects, the lack of a quantitative evaluation metric makes that difficult to ascertain. Future work will hopefully show that style transfers for training can massively increase the datasets computer scientists use when working in the art world. This approach could be applied to almost all models that currently work on photographs, giving this technique the exciting potential to bring all sorts of powerful tools to art scholars.

References

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection, 2020.

Efros, A. A. and Freeman, W. T. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346, 2001.

Girshick, R. Fast r-cnn, 2015.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.