MLP Coursework 4: Evaluating Relativisitic Generative Adversarial Networks in Artwork and Shoe Generation

s1832582, s1876158
G075

Abstract

Recent developments in Generative Adversarial Networks has led to a renaissance of sorts in AI art. We investigate some common GAN architectures, trained on paintings and images of shoes, to see how useful they might be in the creative process. The problem of evaluating the output of Generative Adversarial Networks (GANs) trained on images is well documented (Barratt & Sharma, 2018). In an effort to overcome issues with the Fréchet Inception Distance (FID) (Heusel et al., 2017) being ImageNet specific we propose a modification to the standard wherein the architecture that provides the embedding trained on the images of interest. Unfortunately our proposed method still suffered from many of the same issues as FID, and did not provide a good metric with which to compare outputs. We find that the GANs with relativistic objective functions (Jolicoeur-Martineau, 2018) outperform those without in the task of producing realistic shoes and art.

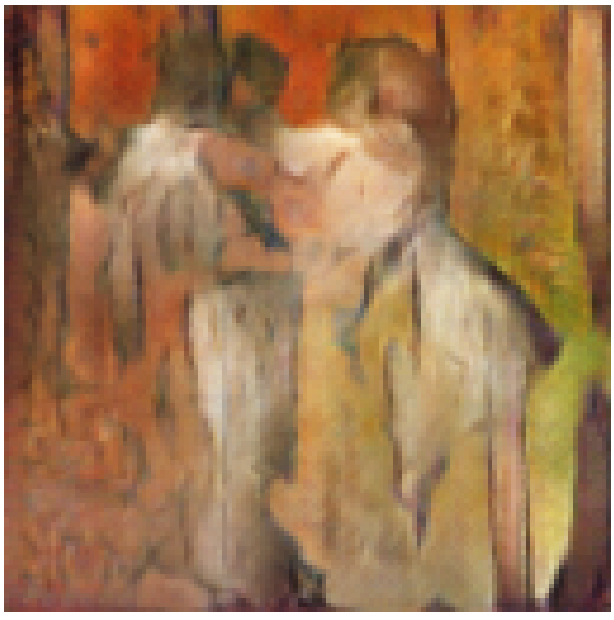

Figure 1. Les Danseurs - Paul CezGANne (Generated by GAN)

1. Introduction and Motivation

In October of 2018 the auction house Christies listed the "Portrait of Edmond Belamy" by the collective Obvious for auction (Christies, 2018). The work had been produced by the collective using Robbie Barrats's art-DCGAN code. The collective trained art-DCGAN on 18th century portraits and the output looked like a reasonable, albeit clearly "GAN-ist" interpretation of the same (Barrat, 2017). To accompany this was a wave of speculation about the role of technology and more particularly AI in art (Vincent, 2018). Some claim that one of the main utilities of modern art is to launder and store wealth (Saunders, 2013) so perhaps the biggest sign that AI art had "arrived" is that the piece sold for $432,500, vastly higher than the $7000 - $10,000 that was expected (Wetzler, 2018).

After the invention of GANs in 2014 by (Goodfellow et al., 2014) interest in applying a GAN to the creation of art quickly followed. Work by Mario Klingemann (Poscic, 2018) (whose own work is up for auction in March 2019 at Sotheby's) and Robbie Barrat shows that there is some merit to using GANs (and generative models more generally) as part of the process of producing art.

This has coincided with great interest from researchers in GANs, and several flavours of GAN architecture have been developed. Recently there has been interest in Relativistic GANs (Jolicoeur-Martineau, 2018), where the loss function is designed such that training the generator both increases the probability the discriminator assigns to fake data, but also decreases the probability it assigns to real data.

Our original objective was to see if we could reproduce artwork using GANS that could be seen as real. Our intention was to use a few different models and compare the results to see which could fool a person into believing it was art made by a real human the best. After running some initial GANs we found GANs to be more unstable than originally predicted. We had trouble getting realistic art images using the architecture we selected.

The other issue we ran into is that GANs notoriously have poor evaluation methods (Barratt & Sharma, 2018). Our method to eyeball generated images from the GAN to see which could pass as art proved to be a large task that would not yield adequate results. The reason being as two MSc students we have a narrow view of what is one person could consider good art. The automated scores such as Inception Score (IS) (Salimans et al., 2016) and FID have their own

issues that we will elaborate more on in the evaluation subsection.

Originally, inspired by (Denton et al., 2015), we intended to use human volunteers to provide evaluations of outputs of the models. Unfortunately we were not able to do this.

This motivated what we believe to be a novel approach. We trained a Variational Autoencoder (VAE) as a baseline with which to compare the GAN architectures we were interested in, as well being an integral part in a new scoring metric we wished to try the Fréchet Variational Distance (FVD). Our idea was to modify the FID to use VAE which will be explained in more detail in section 3.

We are sill interested in seeing which GAN architectures are most conducive to producing art. We will be comparing DCGAN, Relativistic DCGAN and our baseline VAE, both subjectively, and using quantitative scores like IS, FID and our FVD. Our objective is to see if there are noticeable differences, either in terms of images produced or performance between these methods. Lastly, we hoped to generate some novel and interesting images in the process.

After the interim report we also chose to include another dataset, slightly different from art. We chose to see if we could produce new styles of shoes using a large dataset of shoe images. Producing a new style of shoes can be seen as a form of art and thus is somewhat related, but gave us a more realistic and extra dataset to test our methods and architecture.

2. Dataset and task

2.1. Dataset - Art

One of the datasets we looked at was the Painter by Numbers dataset found on Kaggle's openly available datasets (Nichol). This dataset was originally originally intended to be used to try to identify which author painted a given piece of work, or even possibly determine if it was a forgery or authentic. However it provides an excellent resource for training generative models on art, containing a large number of images in a wide array of styles and genres.

The dataset consists of 79434 paintings. Each painting in the directory has a corresponding entry in the CSV file with a filename, title, genre, style, artist, and date which can be used as a look-up. We created a script which allows us to filter paintings by genre, style, or both. This is so that our GAN has more consistent data to train on.

To this end we partitioned off the part of the dataset corresponding to "impressionist" paintings and used these for our training set (8220 images). We chose to partition the training set this way as we were interested in learning style as opposed to content of images.

There was no need to preprocess the data any further than partitioning by category.

2.2. Dataset - Zappos Shoes

We also looked at the UT-Zap50k (Yu & Grauman, 2014; 2017) dataset. This consists of 50025 thousand images of shoes of various types and styles, all sourced from the Zappos online clothing retailer. All the images are centred on a white background and have the same orientation. The images can be divided into four catagories: Shoes (30169 instances), Boots (12832 instances), Sandals (5741 instances), and Slippers (1283 instances).

Similar to the art dataset there was no need to preprocess the images before using them in the GAN.

2.3. Task

Our task is to see which GAN architecture is the best at producing artwork belonging to a certain genre or style. We will be using IS and FID to judge the quality of output from our trained models.

We would like for the models to hopefully generate images that strike a balance between being too novel (and looking noisy) and resembling the inputs too much (and not being creative). To test whether the models are generating outputs that an artist would find useful we intend to use these measures: A nearest neighbours search of the data-set, to check if output overly resembles training data; interpolation over the z-space (introduced in the next section) to see if there is mode collapse, IS, FID, and our own modification to FID to measure the diversity of output images.

3. Methodology

The generative adversarial network (GAN) (Goodfellow et al., 2014) is a neural network architecture inspired by one question: "Can we train neural networks to beat each other?". They consist of two opposed networks: a generator (G) and a discriminator (D). The generator produces "counterfeit" examples of data and defines the "z-space" of the model: an easily sampled from probability distribution, the samples from which are then mapped to generated output images. The discriminator tries to determine whether the input it has just seen is real or fake. This constitutes a minimax game that can be expressed in two loss functions, one for the discriminator:

$$J^D = -\frac{1}{2}\mathop{\mathbb{E}}_{x\sim p_{data}} logD(x;\theta_D) - \frac{1}{2}\mathop{\mathbb{E}}_{z\sim p_z} log(1-D(G(z;\theta_G))) \quad (1)$$

And one for the generator:

$$J^G = -\frac{1}{2}\mathop{\mathbb{E}}_{z\sim p_z} log(D(G(z;\theta_G))) \quad (2)$$

Where p_{data} are the real samples that the discriminator was trained on and p_z refers to input noise, usually set to be Gaussian with mean zero and variance 1. θ are the trained parameters for the discriminator and the generator.

This game has a Nash equilibrium at a local minima,

which is what the network is seeking. These loss functions are called non-saturating since (as opposed to having $log(1 - D(G(z, \theta_G))$ in the generator loss function) it provides stronger gradients during the early stage of training when the discriminator is rejecting samples with high confidence. Note that the loss function for the discriminator is equivalent to the Jensen-Shannon Divergence (JSD) if the discriminator is optimal (Goodfellow et al., 2014; Jolicoeur-Martineau, 2018), this is relevant for our discussion of relativistic loss functions for GANs later.

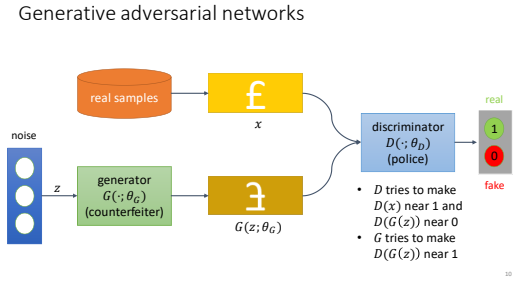

Figure 2. Simple GAN architecture for an example of producing fake currency. The generator creates samples it thinks are similar to real currency samples using noise. The discriminator then tries to determine if the input is from a real or fake sample.

3.1. Variational Autoencoders

A Variational Autoencoder (VAE) consists of two networks, an encoder and a decoder trained end to end. The encoder takes images and encodes them into a vector, constrained to roughly follow a unit Gaussian. This defines a latent z-space, since the decoder takes samples from this unit Gaussian and produces images that should resemble the samples the model was trained on (Doersch, 2016; Kingma & Welling, 2013). We used this to obtain a baseline with which to compare the models, as training a VAE is (relatively) quick and simple. We used a Convolutional-VAE (CNN-VAE) where the encoder and decoder used are convolutional and deconvolutional neural nets respectively. We also use the obtained weights in a modification of the FID that we hope will address some of that distance's issues. We used CNN-VAE code by Chandel (2018).

3.2. Deep Convolutational Generative Adversarial Network

DCGAN (Radford et al., 2015) is a GAN wherein the discriminator uses strided convolutions with global average pooling instead of fully connected layers for downsampling the input images, and the generator uses transpose convolutions for up-sampling the input noise z, along with Batch Norm in both the generator and discriminator. The convolutions and transpose convolutions should make the network learn its own spatial up and downsampling. This approach has been shown to be successful on many image processing tasks (Rawat & Wang, 2017; Krizhevsky et al., 2012). Removing the fully connected layers is intended to improve model stability and using batch norm is intended

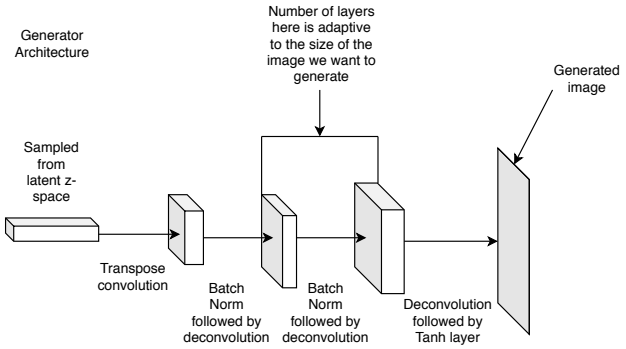

Figure 3. Generator architecture used in DCGAN

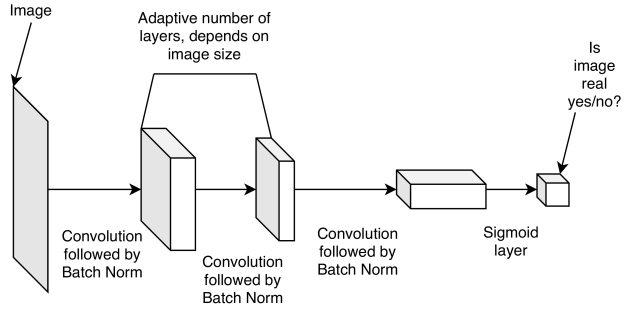

Figure 4. Discriminator architecture used in DCGAN

to counteract problems with poor initialisation and to help with gradients backpropagating in deeper models.

3.3. Relativistic GAN

We make an implicit assumption that GANs do not know that half of the data they are being fed is fake (Jolicoeur-Martineau, 2018). If we revise this assumption we get some potentially beneficial results and this is what the Relativistic GAN (RSGAN) attempts to do. Firstly this lets the discriminator take advantage of a priori knowledge that half of the example in each mini-batch is fake. Secondly allowing the network to increase $D(fake)$ as well as decrease $D(real)$ is more in line with how one would minimise the Jensen-Shannon Divergence normally. Thirdly when the discriminator is trained to optimality it stops paying attention to what it is for samples to be "real", the gradient ignores real data. This results in fake samples not becoming any more realistic and training stalling. However if $D(real)$ always decreases when $D(fake)$ increases, real data is always incorporated into the gradient and thus the network should be able to learn in more difficult settings.

We can achieve this by using relativistic loss functions as follows:

Let x, z be real and fake samples respectively, and $C(x)$ to be the discriminator without its activation function applied (the critic function), and f to be an activation function. Then we can define non-saturating relativistic loss functions for both the generator and discriminator:

$$J^D = -\mathbb{E}_{(x,z)\sim(p_{data},p_z}[\mathrm{f}(C(x) - C(z))] \quad (3)$$

$$J^G = -\mathbb{E}_{(x,z)\sim(p_{data},p_z}[\mathrm{f}(C(z) - C(x))] \quad (4)$$

These loss functions have a different interpretation to standard loss functions used in GANs: instead of asking "what is the probability that the sample is real", it is asking "what is the probability the given real data is more real than randomly sampled fake data" (Jolicoeur-Martineau, 2018).

3.4. Evaluation Methodologies

There are two main methods of automatically scoring GANs, IS and FID.

IS was created to measure two things: quality and diversity of generated images. (Salimans et al., 2016) We want to see that the images generated are generated with high entropy.

$$IS(G) = exp(\mathbb{E}_{x\sim\ p_a} D_{KL}(p(y|x)) \parallel p(y))) \quad (5)$$

Here D_{KL} is the Kullbach-Leibler Divergence and is a measure of the divergence between two probability distributions, $p(y|x)$ is the conditional class distribution, and $p(y)$ is the marginal distribution over classes, both defined by InceptionNet trained on ImageNet. IS wants $p(y|x)$ to be predictable (high entropy), so when we are given an image we can tell its ImageNet label easily, and for $p(y)$ to be uniform or high entropy so that it generates a diverse set of images. A higher score is "better".

The FID, proposed by Heusel et al. (2017), embeds generated and real samples into the feature space given by a specific layer of Inception Net. It does this by taking the activations of the 3rd max pooling layer with the (real or fake) sample as input. It treats these embeddings like vector representations of multivariate Gaussians. The mean and covariance for the real and generated samples can then be calculated and the Fréchet Distance of the two distributions is calculated as so:

$$\mathrm{FID}(x, g) = \|\mu_x - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_x + \Sigma_g + 2(\Sigma_x\Sigma_g)^{\frac{1}{2}}) \quad (6)$$

Where (μ_x, Σ_x) and (μ_g, Σ_g) are the mean and covariance of the embedding distributions of the data and model respectively (Lucic et al., 2018).

The justification behind this is that if the FID score is lower, the generated samples should be closer to that of the real samples. Unfortunately there it is not clear that this is always the case. We used code by Seitzer (2019) to obtain the FID scores for the models, and modified this code for FVD. We used code by Barratt (2018) to obtain the inception scores for the models.

3.5. Evaluating Evaluation

There is currently a dearth of choice when it comes to evaluating the outputs of GANs trained on images, and even these measures we have chosen are not without criticism

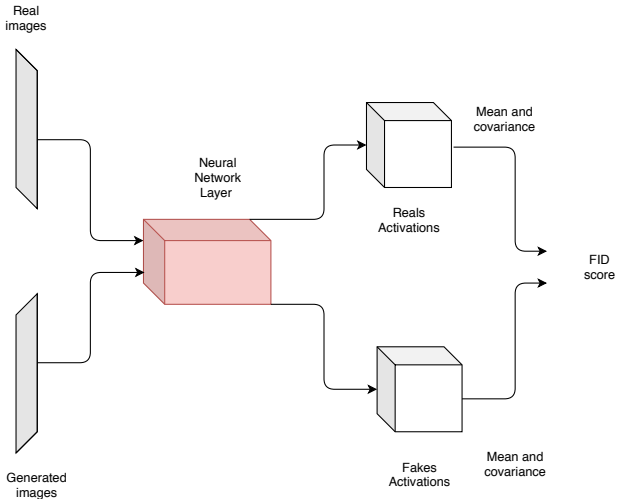

Figure 5. Schematic of how FID score is calculated. The choice of neural network layer used to calculate the real and fake activations is flexible.

(Barratt & Sharma, 2018; Borji, 2018). The most relevant criticism to us is that the Inception v3 network used to calculate IS is trained on ImageNet, so whether it will be able to give a meaningful marginal probability distributions for other datasets is questionable (Rosca et al., 2017). Additionally, the IS rewards crisp, unblurry images whilst, depending on what type of artworks we have trained the GAN on, we may want to generate images with a painterly or impressionistic quality, where a degree of blurriness isn't the end of the world. Another criticism is that IS does not give any indication of overfitting, if our GAN learns to perfectly recreate the data-set images it could still get a good score on IS despite not generating the sort of images we want, since it does not compare generated results to real examples. To discover whether the model is over-fitting we intend to perform a nearest neighbours search on sample outputs from the GAN: if the output is too close to the inputs we know we are over-fitting.

The FID suffers similar problems to IS, since the statistics it uses to calculate the scores are derived from an embedding in the Inception Net architecture.

This fact leads to the unfortunate situation that IS and FID are both sensitive to the weights of the Inception Net (Barratt & Sharma, 2018), and can give different scores depending on whether the Tensorflow or Pytorch InceptionNet model is used, or which version of InceptionNet is used. This has contributed to difficulties in comparing between published results.

Additionally IS when proposed was only used to evaluate a model that had itself been trained on ImageNet (Salimans et al., 2016). The intuition for applying it to other datasets seems to be due to: 1. the ready availability of pre-trained ImageNet Classifiers, 2. It is a widely used (if flawed) metric, 3. ImageNet is such a large and varied dataset that IS and (and by extension FID) are in some way guaranteed

to measure image diversity.

Unfortunately this may mean that generative models trained on images that do no resemble those in the ImageNet dataset (say, impressionist paintings) may be unfairly scored. Instead we propose a tweak to the FID, wherein the neural net used to acquire statistics with which to compare the generated and real samples, comes from the encoder of a Variational Autoencoder trained on the same images as the Generative model.

This Fréchet Variational Distance (FVD) distance hopes to address the conceptual issue where a scoring method trained on ImageNet data is used to evaluate output from models trained on different datasets.

It may not be the case that for the task of generating art that we want the GANs to be working in the same way as if it were trying to generate realistic pictures of cats for instance. Mario Klingemann has stated that "The most interesting stuff usually happens if the model is outside its defined range of values or pushed into areas it was not trained for. But it needs to be a deliberate decision. It allows me to explore different visual worlds and allows me to discover something that is aesthetically pleasing or disturbing" (Poscic, 2018).

4. Experiments

For a baseline we trained a CNN-VAE model on both datasets. We used an image size of 128x128, 25 epochs, Adam learning rule with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ as advocated in Kingma & Ba (2014), and a latent size of 2048. This gave a relatively small image size to z-size ratio ($\frac{1}{16}$), however this number was chosen to aid the use of the encoder in our FVD distance (the dimension of the 3rd max pooling layer of Inceptionv3 is 2048).

Our second experiment was to run a DCGAN and RSGAN on the two datasets. For both DCGAN and RSGAN we used an image size of 128x128, a batch size of 32, 128 hidden nodes in both the generator and discriminator, 50000 iterations, Adam learning rule with a learning rate of 0.001 , $\beta_1 = 0.5, \beta_2 = 0.999$, and a latent size of 128. This gave us a ratio of 1 for image size to z-space size. We used the default settings provided by Jolicoeur-Martineau's code, as advocated in her paper (Jolicoeur-Martineau, 2018).

We expected RSGAN to have better results than DCGAN because of the relativistic loss functions and previous work in (Jolicoeur-Martineau, 2018) showing that RSGAN outperforms DCGAN.

Training GANs is an expensive endevour: each of our GAN models took between 153600 to 179200 GPU hours to train ($30 - 35$ hours with 4 1060 Ti GTX GPUs with 1280 cores each).

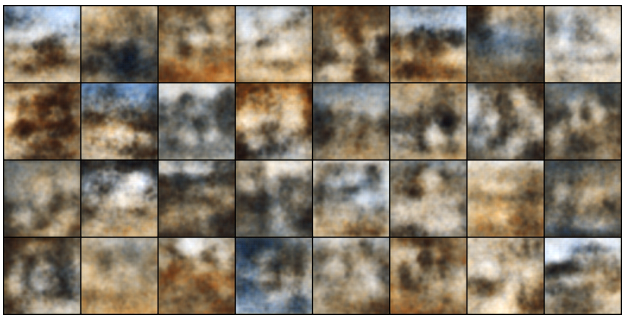

Figure 6. Sample from VAE trained on art

Figure 7. Sample from VAE trained on shoes

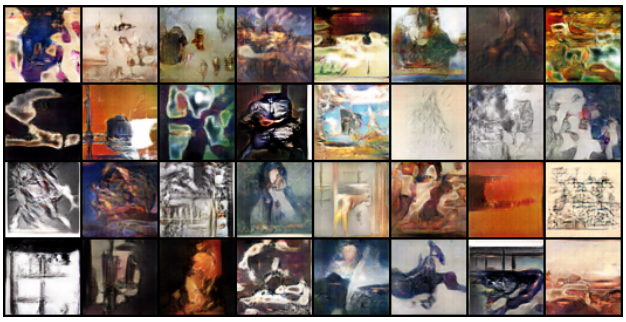

Figure 8. Sample from DCGAN trained on art

Architecture and dataset	IS	FID	FVD
VAE UT-50k	3.18	12.19	N/A
VAE Art	3.50	30.52	N/A
DCGAN UT-50k	4.58	8.21	103.84
DCGAN Art	4.32	5.38	104.65
RSGAN UT-50k	2.56	3.23	106.7
RSGAN Art	3.00	1.77	34.37

Table 1. Results for images generated from the three architectures and two datasets selected.

Figure 9. Sample from DCGAN trained on shoes

Figure 10. Sample from RSGAN trained on art

Figure 11. Sample from RSGAN trained on shoes

5. Results

By running the three models on the two datasets we obtained a set of quantitative results using the evalutation methods mentioned earlier. The results are shown in table 1 show that for both datasets RSGAN performs better than DCGAN and VAE. Unfortunately it also shows that our new FVD score does not accomplish the task of being a better evaluation method for GANs.

The first sanity check of the models is to look at the real images and see if the quality roughly matches the order of the scores. It is clear that the quality of generated images increases in quality from VAE to DCGAN to RSGAN in order from worst quality to best. The images from all models are not perfect. The images generated in the art domain from VAE (figure 6) are blurry, DCGAN have what look like incomplete images or tears in the images (figure 8), and RSGAN has the closest interpretation of impressionist art, but seems to have some mode collapse (figure 10).

One thing to note is the low IS scores for RSGAN. We think RSGAN generated images have a low IS because the outputs the art model do not resemble classes in ImageNet, while the outputs of the shoe model did not show particular diversity of outputs. Again, this shows the lack of quality in IS and FID as evaluation metrics. Figure 6 shows a collection of art samples generated from the VAE. The VAE images are likely poor and blurry because they were only trained on 25 epochs.

Even though the VAE showed enhanced results according to IS, RSGAN was still able to produce lower values for FID than both DCGAN and VAE on both datasets. This confirms what a human can see by looking at the RSGAN samples (RSGAN shoes: figure 11, RSGAN art: figure 10) that RSGAN produces more realistic looking images. This subjective appraisal by the authors is in line with results from the literature that claim that FID correlates well with human judgement (Lucic et al., 2018). It does appear that RSGAN suffers from some mode collapse by looking at the bottom left of figure 10. Those images seem to have the same colors and same patterns. With a mode collapse we would expect to see poorer results from the quantitative metrics since the images clearly do not show diversity.

Our new evaluation method FVD did not perform the way we expected. We believed that it would be better at clearly discerning between "good" and "bad" generated samples because it is trained on images from the same dataset. The impressionist art dataset shows what we were expecting, RSGAN had a significantly lower score than DCGAN. The opposite case for the shoe dataset was shown where DCGAN scores were better than RSGAN.

We believe that one of the reasons FVD ended up not being a good evaluation metric was due to our VAE being undertrained. The samples from the VAE clearly show undertraining on the datasets. Additionally, the FVD suffers from the issue that it is sensitive to the weights in the VAE, this means that the distances found between the art and the shoe dataset are not really directly comparable to each other. It is also clearly not a particularly robust way of measuring the quality of outputted images, shown by the low score afforded to the RSGAN trained on art which, to our eye at least, can't be said to be 70 "FVD points" better than the output from the RSGAN trained on shoes.

6. Analysis

6.1. Interpolation

Interpolation is a method of showing the diversity of samples the latent space encodes. In our context of GANs this

is using a uniform range of values across the latent z-space to generate different images. Remember that each sampled z is a compressed lower dimensional representation of our output. In an ideal GAN architecture as z changes from one value to another we should see the produced images change proportionally with the value of z.

We can see from figure 15 that the shoe interpolation images generated as Z moves linearly from one value to another the image also changes. All four corners were values sampled from the z-space while everything in between was created by moving linearly from one sampled value to another. For the most part the interpolation in figure 15 shows what we want to see, a smooth change from one unique image to another except as the z-value moves down on the grid there is a sharp change.

Interpolation may show mode collapse if there is sharp change from one image to the next. Mode collapse occurs when the GAN generates a limited diversity of samples. This can happen because the objective functions do not explicitly force the generator to learn how to produce a diverse set of images. The generator may learn to generate similar outputs for a range of z values because that output is scored highly by the discriminator. By looking at figure 15 it can be seen there is some mode collapse. As the shoe changes from a heel to a sandal (moving down the first column) there is a sharp change from row three to four.

Figure 16 show changes from less distinct images. The colors and shapes are similar in the images in each corner. This does show a smoother transition than the shoe interpolation. This may mean our z-space is under-trained, and our GAN has not learned to produce a variety of images. However as images from the appendix and figure 13 show, this is not the case. Instead we thing the lack of diversity shown in the interpolation is down to our interpolation methods and not a deficiency in the model itself.

6.2. Nearest Neighbor

To make sure the images produced by RSGAN looked unique and we were not running our quantitative evaluation methods on mere reproductions of the training data we did a nearest neighbors search. Although a lot of the images generated from the impressionist dataset were blurry images, this is easily explained as many of the images in the dataset look quite blurry due to the nature of the genre. A GAN trying to reproduce this will try to recreate the blurriness without the finesse of a brushstroke and that appears to be happening in our generated images. An example of the blurry image with it's nearest neighbors can be seen in figure 12. Note the clear artifacts left by the transpose convolution procedure.

Finding the nearest neighbors for a random image that is more crisp produces encouragingly good results as well. Look at figure 13 to see that the style and color seems to be consistent between the generated image and the nearest neighbors from the dataset. This also shows that our GAN did not simply copy the images in the dataset, but created new images similar to the dataset.

The generated images from the shoe dataset can also be seen to not be just copying images from the original dataset, but generating similar images. Figure 14 shows a generated shoe and its nearest neighbor. The generated image is of a similar style, brown, boot, and small heel. Yet there is a distinct difference, the generated image appears to have laces whereas the nearest neighbor has a buckle. In conjunction with the interpolation section before this we can assert that the models were creating unique images based on the datasets.

(a) RSGAN generated image

(b) Real Image Nearest Neighbor

Figure 12. A comparison of an image from the impressionist dataset with its nearest neighbor.

(a) RSGAN generated image

(b) Real Image Nearest Neighbor

Figure 13. A comparison of an image from the impressionist dataset with its nearest neighbor.

(a) RSGAN generated image

(b) Real Image Nearest Neighbor

Figure 14. A comparison of an image from the shoe dataset with its nearest neighbor.

7. Related and Further Work

One downside of using deconvolutions for upsampling are the visual artifacts mentioned earlier. Going forward it

Figure 15. Shoe interpolation using 4 unique samples from the z-space. Each corner represents the image sampled from the z-space and each image on the grid between shows a gradual change in Z. The abrupt change from heel to sandal shows mode collapse.

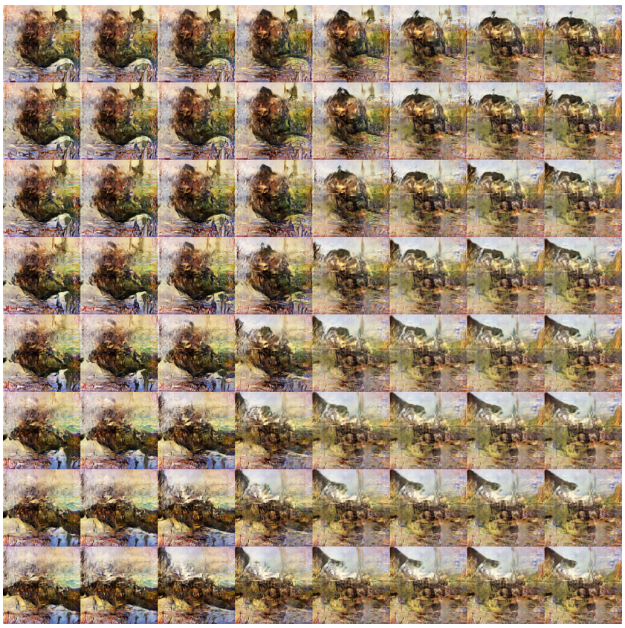

Figure 16. Art interpolation using 4 unique samples from the z-space. Each corner represents the image sampled from the z-space and each image on the grid between shows a gradual change in Z.

may be interesting to experiment with other methods of upsampling, such as a linear upsampling followed by a convolution, or a sub-pixel convolution as described by Shi et al. (2016).

Given more time we would have liked to have applied Wasserstein GANs (WGAN) to our datasets (Arjovsky et al., 2017). These WGANs have generated a lot of research and claims have been made about them dealing well with mode collapse, something our models suffered with to some extent. PixelGAN (Makhzani & Frey, 2017) is another interesting architecture we would like to have experimented with, and compared with the models we investigated in this report.

We would also have liked to do some more experiments on our z-spaces, perhaps seeing if we could ascertain if certain parts of z-vectors encode content or style, or using the z-spaces from VAEs to bootstrap training in GANs.

A problem with our models is that we had no control over what gets generated. There are models that allow the user to condition and generate on certain features such as conditional GANs (Mirza & Osindero, 2014). While this is not something we set out to do, it would help create novel generated paintings. Without it we were still able to find shoes that are unique and paintings that pleasantly fit the impressionist style. Some of our selected shoes and art can be seen in the Appendix as well as figure 1.

8. Conclusion

This project had three main tasks. First could we find an architecture that produced the most realistic images using both subjective and objective metrics. Second, could we improve on the current qualitative measures used to measure GAN success. And third could we produce some images that we felt were enjoyable to view.

We found RSGAN to be the best architecture of those sampled at producing diverse and clear images. RSGAN is clearly superior to DCGAN and VAE when compared using FID, and IS. Even discounting these scores, the output from the RSGAN shows a higher level of realism than both the output from the DCGAN or VAE. The relativistic objective functions really do work!

Scoring generative models continues to pose a challenge. Our attempt at improving the current metrics sadly was not a success. Even though we trained our FVD from the domain we were trying to generate, the metric still has many of the same faults as FID and IS. The values it produces are not easily compared from one dataset to another.

In conclusion, both Christies and Nike have been reticent to approach us with an offer for the output of our GANs.

References

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Barrat, Robbie. art-DCGAN, August 2017. URL https://github.com/robbiebarrat/art-DCGAN.

Barratt, Shane. Inception score for gans in pytorch. https://github.com/sbarratt/inception-score-pytorch, 2018.

Barratt, Shane and Sharma, Rishi. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Borji, Ali. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:1802.03446*, 2018.

Chandel, Shubham. A cnn variational autoencoder (cnn-vae) implemented in pytorch. https://github.com/sksq96/pytorch-vae, 2018.

Christies. Is artificial intelligence set to become art's next medium? | Christie's, Dec 2018. URL https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx.

Denton, Emily L, Chintala, Soumith, Fergus, Rob, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pp. 1486–1494, 2015.

Doersch, Carl. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, and Hochreiter, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Jolicoeur-Martineau, Alexia. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lucic, Mario, Kurach, Karol, Michalski, Marcin, Gelly, Sylvain, and Bousquet, Olivier. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pp. 698–707, 2018.

Makhzani, Alireza and Frey, Brendan J. Pixelgan autoencoders. In *Advances in Neural Information Processing Systems*, pp. 1975–1985, 2017.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Nichol, Kiri. Painter By Numbers. https://www.kaggle.com/c/painter-by-numbers.

Poscic, Antonio. Features | Craft/Work | The Pixels Themselves: An Interview With Mario Klingemann, Aug 2018. URL https://thequietus.com/articles/25188-mario-klingemann-ai-art-interview.

Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rawat, Waseem and Wang, Zenghui. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

Rosca, Mihaela, Lakshminarayanan, Balaji, Warde-Farley, David, and Mohamed, Shakir. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Saunders, Frances Stonor. *The cultural cold war: The CIA and the world of arts and letters*. New Press, The, 2013.

Seitzer, Maximilian. A port of frÃl'chet inception distance (fid score) to pytorch. https://github.com/mseitzer/pytorch-fid, 2019.

Shi, Wenzhe, Caballero, Jose, Huszár, Ferenc, Totz, Johannes, Aitken, Andrew P, Bishop, Rob, Rueckert, Daniel, and Wang, Zehan. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Vincent, James. How three french students used borrowed code to put the first ai portrait in christie's, Oct 2018. URL https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans.

Wetzler, Rachel. How modern art serves the rich. 2018. URL https://newrepublic.com/article/147192/modern-art-serves-rich.

Yu, A. and Grauman, K. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

Yu, A. and Grauman, K. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision (ICCV)*, Oct 2017.

A. Appendix

Figure 17. What are thooooooose?

Figure 18. Blue Suede Shoes

Figure 19. Moon boots

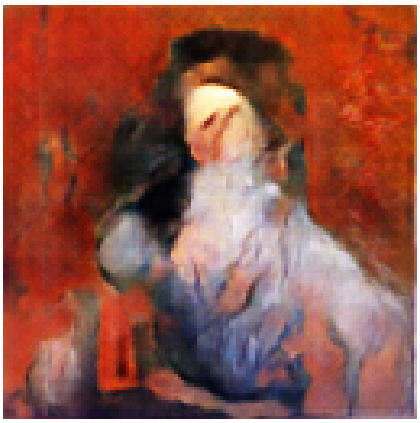

Figure 20. Ghostly figure

Figure 21. A parochial scene

Figure 22. Distant tree