Abstract

Generative Adverserial Networks (GANs) have achieved remarkable results in the field of image generation. However, at the moment, there is no consensus on how to rate the quality of these GANs.
This paper compares seven different metrics. First, a small literature survey was conducted. Then these seven metrics were used to rate nine different GANs trained on the same dataset.The aim was to show which metrics agreed with each other on the quality of these GANs,
This combination shows that, while some of the more simple metrics did not score well, both in earlier literature and in the experiment, the metrics that scored well in earlier literature also showed agreement with each other in the experiment.

Comparing Quantitative Metrics for Generative Adversarial Neural Networks

Bart Slangewal

Supervisor: D.Tax

Bachelor Thesis TU Delft

29-06-2019

Contents

Abstract i

1 Introduction 1

2 Background 2
- 2.1 What is a GAN. . . . 2
- 2.2 How to measure if a GAN is good 2
- 2.3 Research question 2

3 Literature Survey 4
- 3.0.1 Maximum Mean Discrepancy 5
- 3.0.2 Wasserstein distance 5
- 3.0.3 Inception Score 6
- 3.0.4 Frechet Inception Distance 6
- 3.0.5 Classifier 2 Sample Testing 6

4 Experiment 7
- 4.0.1 Selecting a dataset 7
- 4.0.2 Selecting metrics 7
- 4.0.3 Selecting GANs. . . . 8
- 4.0.4 Running the experiment 8
- 4.0.5 Normalized scores plotted for each GAN 8
- 4.0.6 Scores plotted for each metric 8

5 Analysis 11
- 5.1 Similarity in ratings during training progression 11
- 5.2 Do metrics agree on GAN quality?. . . . 11
- 5.3 Volatility of results 11

6 Conclusion 12
- 6.1 Conclusion 12
  - 6.1.1 Pixel space distance metrics 12
  - 6.1.2 Convolutional space distance metrics 12
  - 6.1.3 Inception score metrics 12
- 6.2 Future Work. . . . 12

Bibliography 13

1 Introduction

Since their conception in 2014, a large number of Generative Adversarial Networks (GANs) [2] has been proposed and developed. GANs have achieved great results in realistic image generation, among other fields. Recently, stunning images have been produced. The theory and application of GANs has received much attention. However, the evaluation of these models has not been studied nearly as extensively.

GANs are often evaluated by visual inspection. This is a time consuming process, that inherently suffers from being subjective. There has been research into quantitative metrics, which can be automated. There has even been some research into the relative merits of such metrics. [1][9][4]. However, the question which metric is the most suitable for evaluating GANs in the field of realistic image generation remains open. It is important that consensus is reached. If there is no agreed upon method of objectively measuring progress, it is hard to say which techniques are effective. This impedes the entire field of study.

This paper hopes to contribute to answering this question by applying a variety of proposed metrics to a variety of different GANs. This way, it is possible to tell which metrics agree with each other, and which rate GANs differently. If all, or many, of the likely metrics proposed in earlier work agree with each other, this is a good sign that they are rating GANs objectively. The results of this experiment are combined with a very brief survey of earlier work, in order to recommend some quantitative metrics for rating GANs.

2
Background

2.1. What is a GAN

A GAN, or Generative Adversarial Network, is a framework for generating new samples that resemble an original data set, but are not drawn from it. It was first proposed by Goodfellow et al [2]. GANS can be used for a multitude of topics, such as generating speech and natural language. In this paper the focus will be on generating images. In order to generate an image, a GAN uses two different neural networks, a generator and a discriminator. The discriminator is trained on a large set of images, and judges the likelihood that an image is part of the set. The generator is given random noise, and generates an image from this. The trained discriminator judges how likely it is that this generated image is part of our original data set. The discriminator gives the generator a score depending on how likely this seems. This score is then used to train the generator. At the start of training, the scores will be very low, since the images would be almost random. However, as the network learns, the image resemble the original dataset more and more.

In a sense, the generator is a counterfeiter, trying to print fake money. The discriminator is the police, who try to find the fake money. When the counterfeiters get better and better, the fakes would be indistinguishable from the real thing, thus generating very realistic images. [2].

2.2. How to measure if a GAN is good

Stunning results can be achieved by well trained GANs. However, it is hard to quantify if one GAN is better than another, let alone how much better. There is no consensus on what metric should be used [1]. The question whether one image is better than another is not an easy one. Saliemans et al [7] have tried offering multiple images to people using MTurk, and asking which one they thought was real. Asking people which image they prefer is not a bad way to judge relative quality, but it is costly and takes a long time. Also, it will never be completely objective. This way of judging a GAN is also know as a qualitative metric [1]. A different approach is to mathematically or algorithmic-ally judge a GAN. This is also known as a quantitative metric. The advantages are that an algorithm is faster, does not need pay, and is more objective. However, there is no consensus on which quantitative metric to use [1]. Many quantitative measures have been proposed, and some research has been done comparing them. This paper seeks to investigate if the more popular metrics agree with each other on the relative quality of different GANs, and are thus a usefull way of comparing future research to the current state of the field.

2.3. Research question

This paper seeks to answer the following question:
Does running different proposed and commonly used metrics on the results of many different GANS show an agreement on the quality of said GANS?
To help answering the question, the following questions also will be answered:

1. Which metrics are suitable for running on any GAN, regardless of architecture?

2. Which types of metric disagree with each other? How can this be explained?

3. Is it possible to propose a small set of metrics that can more objectively rate GANs then is now possible?

3

Literature Survey

Xu et al[9], Jiwoong Im et al [4] and Borji [1] have each rated GAN metrics on different disirable qualities.

- Xu et al considered the following metrics: Inception Score, Mode Score, Mean Maximum Discrepancy in pixel and convolutional space, Wasserstein Distance in pixel and convolutional space, Frechet Inception Distance, Classifier Two Sample Test
- Jiwoong im et al considered the following metrics: Inception Score, Mean Maximum Discrepancy, Wasserstein Distance, Frechet Inception Distance,
- Borji considered 29 metrics, all the above are included in his survey.

Note, Xu et al calls Classifier Two Sample Test the 1-Nearest Neighbor Classifier.
The Mode Score was proven by Xu et al to be equivalent with the Inception score [10], and will thus not be considered any further.

From these studies, four criteria were chosen that are important for good GAN performance. Each of these criteria is judged by at least two studies. The outcome of judging the candidate metrics on the criteria is seen in table 3.1. To make it easier to compare the results, all tests have been converted to the model Borji uses. Metric suitability is rated on a high, moderate and low suitability, based on the conclusions of the author of the test.

- Agreement with human judgment
  The intuitive way to judge is one picture is better than another would be to ask humans to rate the pictures. After all, the pictures are usually generated for human consumption. Thus, it is important that a metric agrees with human perception. Jiwoong Im et al have run experiments where they let a metrics and a person determine which looked more realistic, a real image or a generated sample. The metric was then rated based on how often it agreed with the person.
- Discriminability
  Disciminability is defined as the ability for a metric to distinguish images generated by the GAN from real images. [9] They created a set of 2000 real images S_r and a set of the same size with a mix of real and generated samples, S_g. When increasing the amount of generated samples in S_g, it is expected that the distance between S_r and S_g, as calculated by the metrics, increases. Because the Inception Score is not based on distance between distributions, they used the relative IS between S_r and S_g. A metric passes the test if the distance (or relative score) increases when the fraction of fake images increases. Borji also rates the discriminiability is his survey [1]. He bases his conclusion on literature review.
- Robustness to transformation
  In realistic images, small translations and rotations often do not meaningfully change the image. A face that has been rotated slightly or flipped is still a face. A good metric should not punish a GAN for such

		WD Pix	WD Conv	MMD Pix	MMD Conv	IS	FID	C2ST
Human judgment	Jiwoong	2	-	2	-	2	-	-
Human judgment	Borji	-	-	-	-	2	2	-
Discriminability	Xu et al	0	2	0	2	0	2	2
Discriminability	Borji	2	-	2	-	2	2	2
Transformation	Xu et al	0	2	0	2	2	2	2
Transforamtion	Borji	-	-	-	-	1	1	-
Diversity	Xu	1	2	2	2	1	2	-
Diversity	Borji	1	-	0	-	1	1	0
Average score		1	1.6	1	2	1,22	1,71	1,5

Table 3.1: Comparison of metrics based on work by other authors. Metric suitability is rated on a high, moderate and low suitability, based on the conclusions of the author of the test. These scores are rated on a 0, 1, 2 scale. A - means the metrics was not rated, and is not counted when calculating the average score.

small changes. [1]. Xu et al tested robustness to transformation by running the metrics on two datasets. Dataset S, with 2000 real images in it, and S' in which a portion of the images is shifted up to 4 pixels, or rotated up 15 degrees. The scores are then compared. Since these are real images, a good metric should rate them very highly. The transformation should not matter, since they are still real images.

- Generating diverse samples
  A good metric should favor GANs that produce diverse images. If a GAN trains on a set of data for too long, it is possible it will only generate samples of a certain type. For example, if the aim is to generate samples of both cats and dogs, and our GAN only produces very high quality samples of cats, that means it is not generating representative samples. [2] A good metric is sensitive to this.

- Maximum Mean Discrepancy; uses a Kernel function to estimate the two distributions, the real and the fake one. This makes it possible to calculate the distance between them.

- Inception Score; uses a classifier to label the samples, and rate them on quality and diversity.

- Frechet Inception Distance; computes the distance between two distributions, that are extracted from the generated images and the real data.

- Classifier Two Sample Test; uses a classifier to try to figure out which images are generated and which are real. If it fails to do so accurately, the GAN has worked well.

3.0.1. Maximum Mean Discrepancy

To calculate the Maximum Mean Discrepancy (MMD), first a set of real and a set of generated images is gathered. The dissimilarity of the distributions of these sets is then calculated. This is done with the help of a kernel, a function that can be used to appoximate the distribution of the dataset. A low MMD means the two datasets are very similar, and thus that the generated samples are like the real ones.

3.0.2. Wasserstein distance

The Wasserstein Distance (WD) is also know as the earth mover distance. It is a measure of distance between two distributions that can be explained as follows: If the graphs are seen as two mountains, it would be the optimal way to move the earth on one mountain around, so it looks exactly the same as the second mountain. For a GAN, this means constructing all possible transport functions to transform one distribution into the next, and then choosing the optimal one.

Feature Space

Both the MMD and WD metrics are applied in so called pixel space and convolutional space. MMD and WD calculate distance between two distributions. In pixel space, this distribution is simply formed by taking each pixel in an image as a data point. However, Theis et al [8] show that the suitability of such distance calculation is questionable.

3.0.3. Inception Score

First proposed by Saliemans et al [7], Inception score is a widely accepted measure. It uses a third, pre-trained neural network to classify samples. In the standard implementation this classifier is trained on imagenet.

The Inception Score consists of two parts. The first part of inception score is the probability a sample should get a certain label. How certain is the classifier that the digit is, for instance, a seven? This probability is calculated.

The second part of the Inception Score seeks to calculate the following. Do we get all digits equally as often? In MNIST, do we get all ten digits, or does the GAN only generate some digits? If we generate many samples, we expect to see all possible digits. Thus we can score the variety of the samples.When this score is combined, we can calculate how different the distribution of the set of generated images is from the original set of training images.

3.0.4. Frechet Inception Distance

Frechet Inception Distance a modification of the Inception Score. It was first proposed by Heusel et al [3]. It does not use the separately trained classifier simply to label the samples, and score the quality that way. Instead, the values of the neurons of an intermittent layer of the classifier are compared. These values are combined into distribution, one for generate samples, and one for real images. The distance between the distributions is calculated using the Frechet Distance technique. Frechet Distance is often explained as follows: Imagine a man and a dog walking together. They both follow their own path, but the dog is on a leash. If we view the paths of man and dog from above, we get two curves. The Frechet Distance is the minimum length of the leach needed so both can follow their path. Instinctively, the shorter the leash, the more like the path of the man the path of the dog must be, and thus the more similar the curves. The smaller this distance, the more like real images the samples are, and the better the GAN has performed. Heuser et al [3] have shown FID to be consistent with human judgment, making it a promising metric.

3.0.5. Classifier 2 Sample Testing

In Classifier Two Sample Testing [6], the idea is to train a network that will determine whether an image is a generated sample or a real image. To do this, a large amount of generated samples and real images are combined to a new data set, and split into a training and a test portion. During training, the classifier knows the correct answer, and can thus adjust its weights, learning which distribution fits the samples, and which the images. Then, when the testing phase begins, if the original GAN worked well, the distributions of the real images and the samples are very alike. Thus the classifier has nothing to go on and its decision is a guess. Because it is known if an image is real or generated, we can check the answers. The better the classifier worked, the worse the GAN has worked.

4

Experiment

The experiment was setup as follows:

- Selecting a dataset.
- Selecting metrics.
- Selecting GANs.
- Running the experiment.
- Graphing the results.

In order to implement the experiment, the selections had to meet the following criteria.

- First it is necessary that implementations of the metric and the GANs are available in python. This language was chosen because many usable gans are implemented in python/pytorch. Due to the limited scope of the research, it is necessary to use readily available code. This also has the upside of having material to compare our results to.
- Second it was decided to only rate performance on a single dataset. This means any selected GAN must be trainable on this dataset.
- Third, the metrics must be based on different concepts. Many metrics are proposed that are tweaks or optimizations of a know concept. These optimizations can certainly have merit, but this paper looks to compare a diverse range of metrics. In the future, developments for the most promising of these concepts can be researched.

4.0.1. Selecting a dataset

In order to be able to compare the GANs, they are all trained on the same dataset. Xu et al [9] have done extensive empirical testing on the LSUN dataset of natural images. Since their implementation of the chosen metrics is used, this provides an opportunity to compare results. Thus a different dataset is selected.

CIFAR-10 is a collection of small, 32x32 colour images, divided into 10 different classes. Its images are small, making the training times of GANs short. It is also an easily available and widely used dataset. These characteristics make it perfect for a short research such as this one. The most important drawback is that many GANs are developed specifically to be trained on large images. However, since the point of this paper is not do determine which GAN is the best, but which metrics are usuable to determine this, this is not a large problem. It can also be mitigated by selecting GANs to test that are suitable for generating small images.

4.0.2. Selecting metrics

The metrics discussed in the literature survey have been made available by Xu et al [9]. They are implemented in pytorch, and are able to run on CIFAR-10. As shown before, the metrics are based on a variety of different concepts.

4.0.3. Selecting GANs

A small selection of GANs that have been implemented by Kang [5] have been trained. They were chosen because their implementation is available in pytorch, making them compatible with the rest of the project. Also, they are relatively easy to train in a short time, and suitable for generating CIFAR-10 images. These GANs are rated by the discussed metrics. This is done at 100 points in their training, after each so called epoch. Thus it would be expected that the scores would improve, since the metrics should recognize the difference between a well trained and a poorly trained GAN.

The following GAN types were selected, because their implementation worked well with the implementation of the metrics, and because they present a wide variety of different GANs:

- ACGAN
- BEGAN
- CGAN
- DCGAN
- DRAGAN
- GAN
- EBGAN
- LSGAN
- WGAN

Each GAN references the paper in which they originally presented. Note that the implementation of DCGAN came from Xu et al [9], since it was included with the metrics, and Kang did not present an implementation.

4.0.4. Running the experiment

Note that DCGAN has only been evaluated for 25 epochs, not 100. This is because the results were taken from a pretrained GAN by Xu et al. [9]

4.0.5. Normalized scores plotted for each GAN

The results of the metrics on each of the selected GANs are visualized in figure **??**. The scores were normalized with the following formula: $s_i = (s_i - s_min)/(s_max - s_min)$ This brings each of the scores between 0 and 1, making it easier to compare them. Also, all scores except IS were flipped, so in all graphs, a higher score means a better GAN.

The metrics were divided in three graphs per GAN, for ease of reading. The three categories were chosen so the most similar results are in one graph. This led to the following division; the distance metrics in pixel space, the distance metrics in convolutional space, and the inception net based metrics.

4.0.6. Scores plotted for each metric

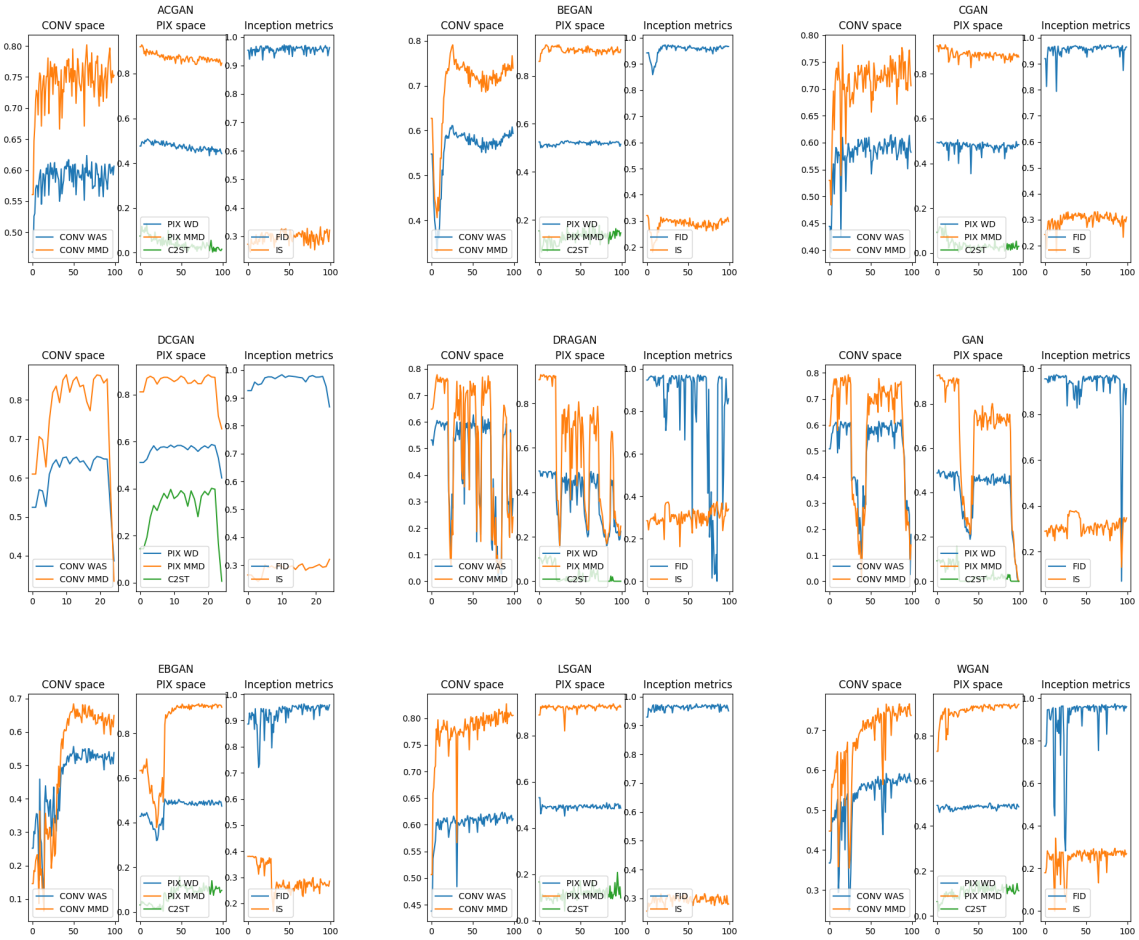

Figure 4.1: Results of the GAN metric experiments. Each figure shows the metric scores for each GAN. PIX and CONV mean pixel space and convolutional space respectively. These results were normalized to be between 1 and 0, and so a higher score means a better GAN.

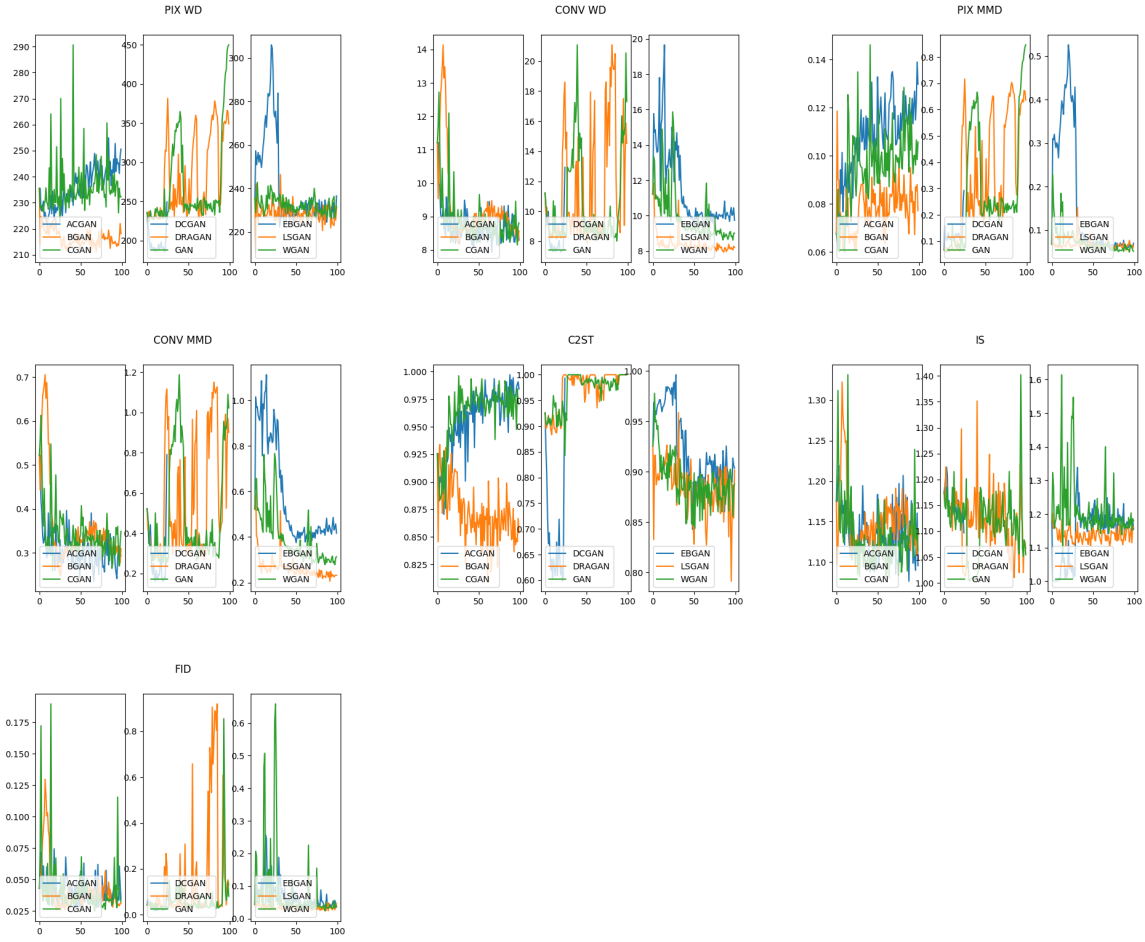

Figure 4.2: Results of the GAN metric experiments. Each figure shows one metric. PIX and CONV mean pixel space and convolutional space respectively. Scores have not been normalized. In all but IS, a lower score means a better GAN

5
Analysis

5.1. Similarity in ratings during training progression

Overall, three categories of metrics stand out. The shapes of these categories match extremely well, as can be seen in Figure 4.1. This van

- The WD and MMD in convolutional space match extremely well.
- The WD and MMD, as well as the C2ST in pixel space are very similar to each other. It is noteable that, out of the three categories, this one is the least similar.
- The FID and IS have fairly similar scores, with the exception that FID occasionally show some large peeks.

It is to be expected that the scores these metrics give a GAN develop in the same way as a GAN improves, as the concepts work in a similar manner.

5.2. Do metrics agree on GAN quality?

- ACGAN, CGAN: These GANs score very similarly in all metrics. From the pixel space metrics, PIX WD and PIX MMD rate ACGAN fairly highly at beginning, then it drops off quickly. C2ST matches this, but drops off far more steeply.
  In the convolutional space, as well as the inception metrics, these GANs score very well.
- BEGAN: Score better than ACGAN and CGAN in pixel space, and very similarly in all others.
- DRAGAN and GAN: These GANs have such volatile results that any useful comparison is impossible
- EBGAN, LSGAN, WGAN: These GANs are rated very similarly by all GANS, steadily climbing to high scores.

5.3. Volatility of results

As can be clearly seen from the graphs, there are steep peaks and valleys in the scores during training. This might be somewhat expected, as the GAN slowly converges. However, 100 epochs is far longer than both Jiwoong Im et al [4] and Xu et al choose [9]. They both train their GANs for only 25 epochs. These results would implicate that this does not give a realiable score, especially since the DCGAN is taken directly from Xu et als implementation, and shows a large drop in GAN quality between epoch 20 and 25.

6

Conclusion

6.1. Conclusion

Running different proposed and commonly used metrics on the results of many different GANS does indeed show an agreement on the quality of said GAN. From the tested popular GAN metrics, the following three categories emerged, not only based on the theoretical principles on which they rely, but also in their rating of nine different GANS.

6.1.1. Pixel space distance metrics

The Classifier Two Sample Test, as well as the pixel space Wasserstein Distance and Maximum Mean Discrepancy metrics seeks to find the distance between two distribution in pixel space, one with real images, and one with generated samples. They show agreement with each other, but not with other metrics.
Literature review has shown that these are not good choices for metrics. This papers experiment confirms this.

6.1.2. Convolutional space distance metrics

Wasserstein Distance and Maximum Mean Discrepancy score far better when used in convolutional space. They agree with each other, as well as the inception score based metrics. In this papers literature review, they scored among the highest. However, far less has been written on this technique than on the inception based metrics, so their performance shows promise.

6.1.3. Inception score metrics

The Frechet Inception Distance and the Inception Score agree with each other and the convolutional space metrics. However, the Inception Score got a very low score in the literature review. This shows that the inception score might be a good metric to evaluate simple GANs like these in this experiment, but more elaborate study revealed many weaknesses. The Frechet Inception Distance scores very well in the literature review, and agrees with the other promising metrics in the experiment.

Overall, the Wasserstein Distance and Maximum Mean Discrepancy in convolutional space show great promise, and should be researched further. The Frechet Inception Distance is one of the most used, tested and written about GAN metrics, and this paper does nothing to disprove its quality. Since these three metrics all agree with each other, it is recommended to use the Frechet Inception Distance for now, as it is more proven.

6.2. Future Work

The Wasserstein Distance and Maximum Mean Discrepancy in convolutional space have shown to agree with the often use Frechet Inception Distance. Xu et al have gotten good results with them. [9] Overal, these are metrics that could use more study, as could the field of GAN metrics overall, as it is critical for the improvement of Generative Adverserial Neural Networks

Bibliography

[1] Ali Borji. Pros and Cons of GAN Evaluation Measures. *arXiv e-prints*, art. arXiv:1802.03446, Feb 2018.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1406.2661, Jun 2014.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv e-prints*, art. arXiv:1706.08500, Jun 2017.

[4] Daniel Jiwoong Im, He Ma, Graham Taylor, and Kristin Branson. Quantitatively Evaluating GANs With Divergences Proposed for Training. *arXiv e-prints*, art. arXiv:1803.01045, Mar 2018.

[5] H Kang. Collection of generative models in pytorch version. `https://github.com/znxlwm/pytorch-generative-model-collections`, 06 2018. Accessed: 25-05-2019.

[6] David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. *arXiv e-prints*, art. arXiv:1610.06545, Oct 2016.

[7] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *arXiv e-prints*, art. arXiv:1606.03498, Jun 2016.

[8] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv e-prints*, art. arXiv:1511.01844, Nov 2015.

[9] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv e-prints*, art. arXiv:1806.07755, Jun 2018.

[10] Zhiming Zhou, Weinan Zhang, and Jun Wang. Inception Score, Label Smoothing, Gradient Vanishing and -log(D(x)) Alternative. *arXiv e-prints*, art. arXiv:1708.01729, Aug 2017.