CAP6610 Project Report 4

Justin Ho
justinho@ufl.edu

April 4, 2023

1 Introduction

As a reminder from the last report, I was able to successfully create a Conditional Variational Auto-Encoders written using TensorFlow 2.0 for the CIFAR-10 dataset. Now that all my models are working, I can almost begin experimentation and analysis. The only thing left is to pick out the metrics that I want to use to evaluate these models and to design the evaluation steps to be used for my experiment. In this report, my objective is to provide a rough framework that I will be using to evaluate the images generated by the models, and an overview of the metrics I will use to evaluate the models. I did a brief cursory overview of the metrics primarily used to evaluate generative models in report 1 that is worth revisiting now that the experiment is more defined.

2 Revisiting Generative Metrics

Before the advent of automatic metrics, generative techniques were primarily evaluated by human evaluators to see if the images generated were aligned with human perceptions of quality. This technique was simple but had the downside of being very expensive and also being unable to scale particularly well [3]. To counter this, automatic metrics based on neural techniques were created. These techniques have been shown to correlate with human judgment, and are computationally inexpensive, allowing them to practically replace humans in the evaluation process [1]. The objective nature of automatic metrics also offers an attractive solution to counteract issues of potential bias that can occur in human annotators. The two standard metrics that are used are Inception Score and Fréchet inception distance.

2.1 Inception Score

Inception Score (IS) is an automatic metric created in 2016 that aims to score the images produced by generative models based on two primary factors: the variety and the quality of the images [5]. Variety refers to the diverseness in the class of images generated and quality refers to alignment with human judgment (features such as clarity, discernability, etc). IS utilizes the pre-trained Inception classifier, to generate outputs that correspond to probability distributions of what that classifier thinks an image is. This process is repeated for every single class and summed together to obtain a marginal distribution. This marginal distribution is utilized as a metric to judge the diversity of a generative model while the class distribution is used to judge image quality [2]. To compare two distributions, KL divergence is used to compare the difference between the two distributions, and the more different a class distribution is from the marginal distribution, the higher of a score we should give the generative model. Lastly, we take the exponential of this value to make the metric more sensitive to improvements [5].

2.2 Fréchet Inception Distance

Fréchet Inception Distance (FID) is another automatic Inception based metric introduced in 2017 that improves upon the weaknesses of IS and aims to provide a more comprehensive metric that is less susceptible to problems that plague IS such as no sensitivity to mode collapse, bias to high definition images, etc [3]. FID works by first using the Inception network to project samples from real and generated images into an embedding. Next, the assumption is made that the distribution is of a multi-variate Gaussian, and with that, the two embeddings are then compared using Fréchet distance for similarity [4]. If the generated embeddings and the real embeddings overlap, the FID score will be small, and vice versa.

3 A brief description of evaluation steps

In terms of the implementation of these two metrics, the maintainers of TensorFlow provide a convenient out-of-the-box implementation of both IS and FID in their TF-GAN package. In terms of the evaluation process, I first will split the CIFAR-10 data into train and test as the test data is needed for evaluation purposes with FID. I will then train 4 GANs and 4 VAEs on the training data. The models that will be tested will use the following architectures: transposed convolution, upscaling convolution, conditional transposed convolution, and conditional upscaling convolution. In terms of model size, all GANs and VAEs respectively will be very similar if not the same size to prevent bias from being introduced. Lastly, all 8 models will generate 200 to 400 random images and will be processed using both FID and IS.

4 Conclusion and Next Steps

The past weeks were really interesting as I spent the majority of my time reading and thinking about how I would structure this experiment more than coding. In terms of the next steps, because the code is pretty much all but done, all that is left for me to do is to pick a universal ”size” to use for all the models, train the models, evaluate them and obtain the results. After this, all that will be left for me to do is write the final paper.

References

[1] Hamed Alqahtani, Manolya Kavakli, and Dr. Gulshan Kumar Ahuja. An analysis of evaluation metrics of gans. 07 2019.

[2] Shane Barratt and Rishi Sharma. A note on the inception score, 2018.

[3] Ali Borji. Pros and cons of GAN evaluation measures. *CoRR*, abs/1802.03446, 2018.

[4] 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. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[5] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016.