An evaluation metric for generative models using hierarchical clustering

Gustavo Sutter P. Carvalho
ICMC, Universidade de São Paulo
gustavo.sutter.carvalho@usp.br

Moacir A. Ponti
ICMC, Universidade de São Paulo
moacir@icmc.usp.br

Abstract

We present a novel metric for generative modeling evaluation that uses divergence between dendrograms computed from training and generated data. Our approach, which borrows theoretical foundations from clustering analysis, is validated by sampling from real datasets and also on samples generated by a GAN during training, with results comparable to state-of-the-art metrics.

1 Introduction

Generative modeling techniques have became largely studied after the introduction of Variational Autoencoders [8] and Generative Adversarial Networks [5], followed by variations of both of them. Although several metrics where proposed to evaluate such models, e.g. the Inception Score (IS) [11] and the Fréchet Inception Distance (FID) [7], there still problems to be solved [13] [1] [6] and new metrics to be investigated.

2 Proposed Method

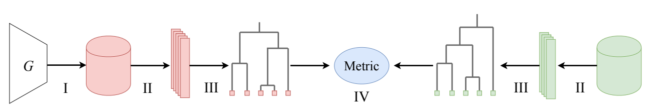

Figure 1: Step-by-step representation of our method. (I) The set of generated images is produced by the generator. (II) Representations are extracted from the images. (III) A dendrogram is built using each set. (IV) The metric is computed as the divergence between the dendrograms.

We investigate the problem of evaluating generative models through the lens of hierarchical clustering, as illustrated in Figure 1, by creating dendrograms for real and fake data, and using the divergence between dendrograms as a measure of quality and diversity of generated data. The motivation arises from the natural assumption that if the generated data distribution is similar to the real distribution, the clustering of both distributions will capture the similarity of their representations. Our hypothesis is that the dendrogram captures more about the distributions than the first and second moment.

The similarity between clusters is quantified by using results from Carlsson and Mémoli [2], which demonstrates a dendrogram is equivalent to a ultrametric space, allowing to employ the Gromov-Hausdorff distance. Although it provides theoretical guarantees, the exact distance is computationally prohibitive, therefore we use an approximation which was applied successfully for concept drift detection on data streams by Costa [3].

Given two dendrograms θ^{real} and θ^{fake}, constructed from the real data X_{real} and the fake data X_{real}, the *Dendrogram Distance* (DD) can be computed using the agglomerative distances of each

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

dendrogram, represented as d, as follows:

$$DD(X_{real}, X_{fake}) = \frac{1}{N}\sum_{i=1}^{N} |d_i^{real} - d_i^{fake}| \tag{1}$$

As other metrics for generative learning, we use a neural network to extract a better representation for the data points, in our case the output of the global average pooling operation after the last convolution of a Inception-V3 [12] pre-trained on ImageNet [4].

3 Experiments

We use two different evaluation schemes to check whether our metric is effectively capable of judging how well the generator approximates the real data distribution: (i) evaluate the model capability of detecting mode collapse, which is done via sampling on real datasets, (ii) investigate how our metric correlates with sample quality during training.

In order to check if our metric is detecting mode collapse in the generated samples it is necessary to have a controlled environment where the presence of mode collapse is already known. Therefore we simulate a class-imbalanced scenario, by sampling from one class at a time. This way we are able to check if the metric gives a worse score for sets where there is less class diversity compared to a ground-truth set, which reflects the true class distribution of the datasets.

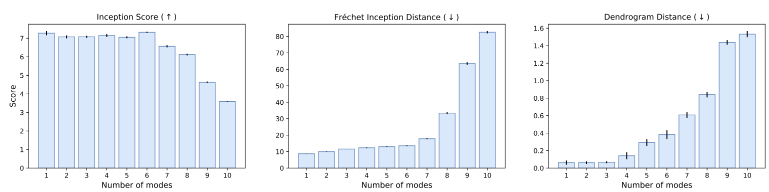

Figure 2: Comparison between metrics with increasing modes: IS (left), FID (middle), DD (right).

Figure 2 shows results with CIFAR-10 [9], where the ground-truth set has 10 classes while the other vary from one to ten classes. Our results confirm the metric is able to detect the growing presence of classes as the value goes up as the number of classes get closer to the ground-truth/complete real dataset. Note that the IS failed to capture it. Also, our metric grows more smoothly than the FID.

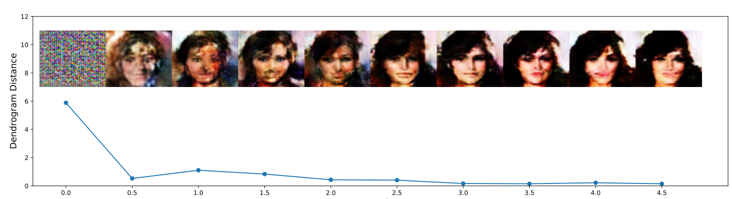

Figure 3: The behavior of our metric while the generative model is training.

To understand the metric behavior during training we trained a SAGAN [14] on CelebA [10], computing our metric every half epoch. The results in Figure 3 demonstrate that the Dendrogram Distance is capable of capturing the overall image quality at each step, being an indicator for the model's convergence.

4 Conclusions and Future Work

Our metric is promising, being competitive to other state-of-the-art approaches, which opens new possibilities for the study of generative model behavior in particular when concerning mode collapse. As a work in progress, there are studies yet to be conducted, namely analyzing the effect of the sample size, the behavior on other datasets and the feasibility of using DD as a auxiliary objective when training GANs.

References

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

[2] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of machine learning research*, 11(Apr):1425–1470, 2010.

[3] Fausto Guzzo Da Costa. Employing nonlinear time series analysis tools with stable clustering algorithms for detecting concept drift on data streams. 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[6] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020.

[7] 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 *Advances in neural information processing systems*, pages 6626–6637, 2017.

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

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[13] 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 preprint arXiv:1806.07755*, 2018.

[14] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019.