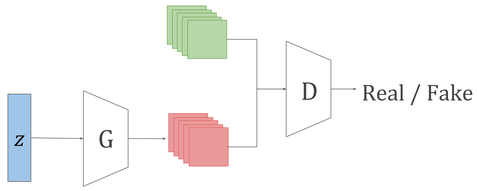
An evaluation metric for generative models using hierarchical clustering

Gustavo Sutter P. Carvalho, Moacir A. Ponti
ICMC, Universidade de São Paulo (USP), São Carlos/SP
gustavo.sutter.carvalho@usp.br, moacir@icmc.usp.br

Introduction

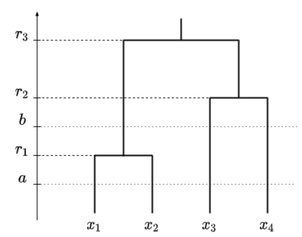
- Generative modeling aims to estimate the generative process of a set of data points
- GANs are an implicit generative model which uses two separate neural networks to estimate the true data distribution



- Optimizing two networks at the same time is hard, often resulting in challenges:
  - Low sample quality
  - Mode collapse
- Contrary to classification or regression it isn't straightforward to define an evaluation metric
  - We propose a novel metric, focusing on mode collapse detection

Dendrograms

- Representation of the hierarchical clustering of the dataset
- It's a function that maps every distance *r* to a the set of clusters at that point.
- For example: $\theta(b) = \{\{x_1, x_2\}, \{x_3\}, \{x_4\}\}$



- Carlsson and Mémoli (2010) have demonstrated that a dendrogram *(X,θ)* is equivalent to an ultrametric space *(X, u)*:

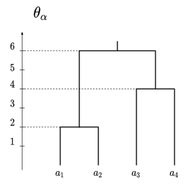
$$u(x_i, x_j) = \min\{r | x_i \text{ and } x_j \text{ belong to the same cluster}\}$$

Distance between dendrograms

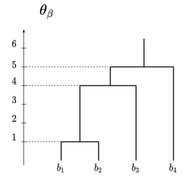
- Because of this equivalence we can use the Gromov–Hausdorff distance
- The true distance is inefficient, but Costa (2017) proposes a approximation:

$$\hat{d}_{\mathcal{GH}}(X_\alpha, X_\beta) = \max_i |u_\alpha(i) - u_\beta(i)| \qquad u_\alpha(i) \leq u_\alpha(i+1) \quad u_\beta(i) \leq u_\beta(i+1)$$

- For example:



$$((u_\alpha)) = \begin{pmatrix} 0 & 2 & 6 & 6 \\ 2 & 0 & 6 & 6 \\ 6 & 6 & 0 & 4 \\ 6 & 6 & 4 & 0 \end{pmatrix}$$

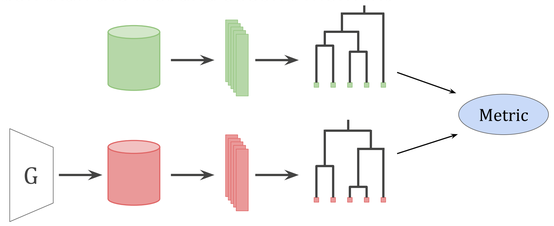


$$((u_\beta)) = \begin{pmatrix} 0 & 1 & 4 & 5 \\ 1 & 0 & 4 & 5 \\ 4 & 4 & 0 & 5 \\ 5 & 5 & 5 & 0 \end{pmatrix}$$

$$\hat{d}_{\mathcal{GH}} = max_i \left| \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2 \\ \dots \\ 6 \\ 6 \\ 6 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \dots \\ 5 \\ 5 \\ 5 \end{bmatrix} \right| = |6 - 4| = 2$$

Proposed method

- If the generated data follows a distribution similar to the real data, their clustering must also be similar
- Our hypothesis is that the dendrogram captures more about the distribution than the first and second moment

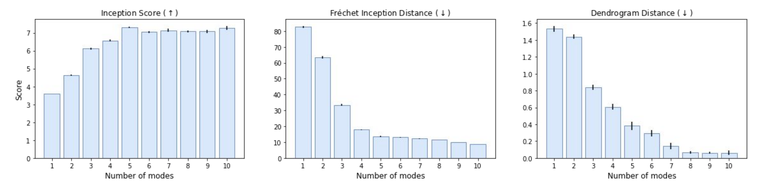


- We used a relaxation of the distance proposed by Costa (2017), using the mean instead of the maximum value

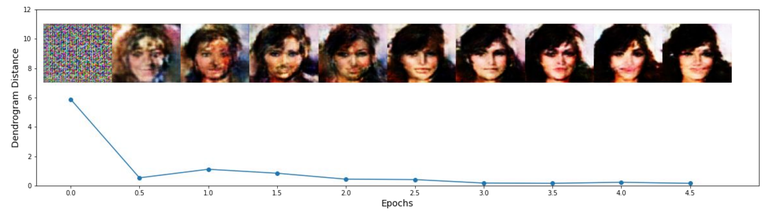
$$\text{DD}(X_{\text{real}}, X_{\text{gen}}) = \frac{1}{N} \sum_{i=1}^{N} |u_{\text{real}}(i) - u_{\text{gen}}(i)|$$

Experiments

- We simulated a class-imbalanced scenarios by sampling subsets of the CIFAR-10 dataset
- Compared our results with Inception Score and Fréchet Inception Distance



- We also investigated the behaviour of our metric while the generative model is training
- Trained a SAGAN on CelebA



Conclusions and future work

- Our metric is competitive when compared to other state of the art approaches, even producing better results on mode collapse detection
- As it still work in progress there are things to be addressed
  - Compare to more metrics (Wasserstein distance, Mode Score, Kernel MMD)
  - Test on different datasets
  - Experiments on sample efficiency