An evaluation metric for generative models using hierarchical clustering

Gustavo Sutter P. Carvalho, Moacir A. Ponti

gustavo.sutter.carvalho@usp.br, moacir@icmc.usp.br

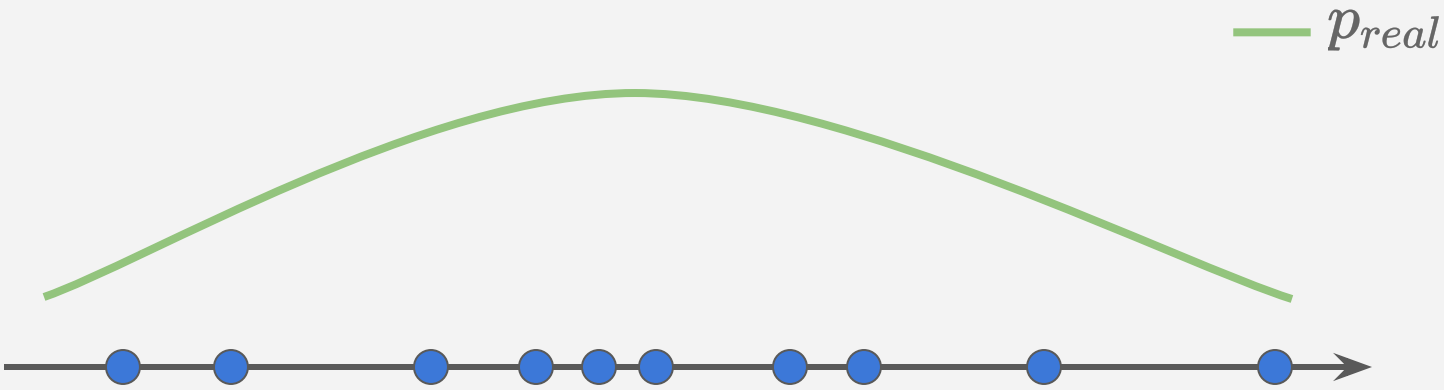
ICMC, Universidade de São Paulo (USP), São Carlos/SP

Agenda

- Generative modeling
- Dendrograms
- Proposed method
- Experiments and results
- Conclusion

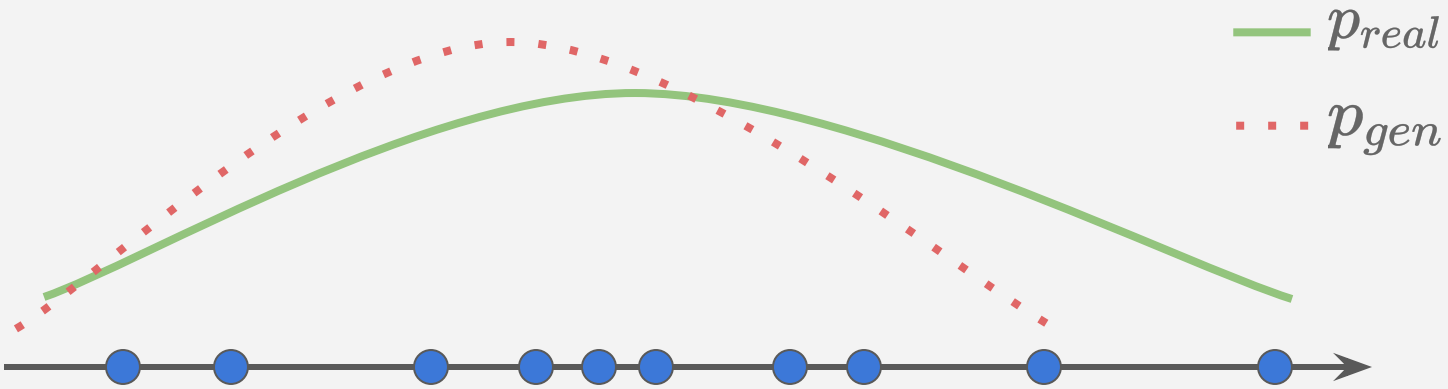
Generative models

- Aim to estimate the generative process of a set of data points



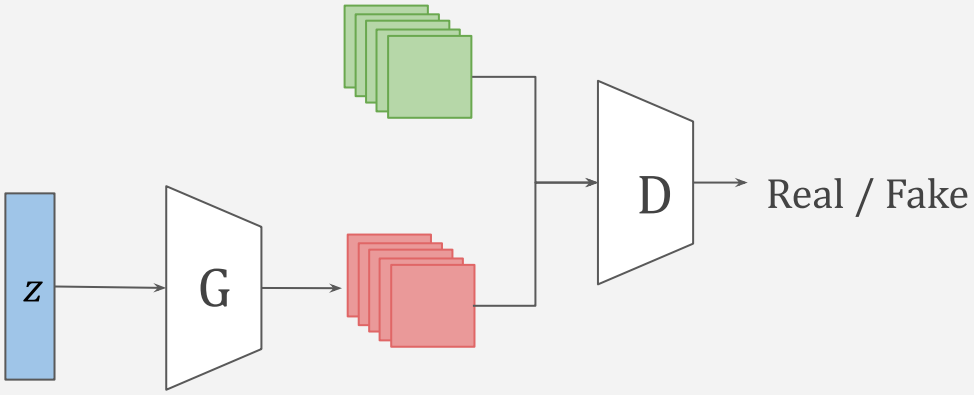
Generative models

- Aims to estimate the generative process of a set of data points



Generative Adversarial Networks (GAN)

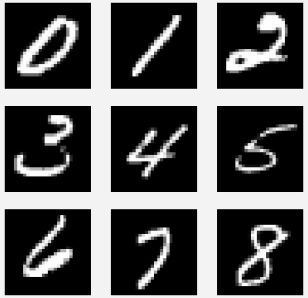
- *GAN* is an implicit generative model which uses two separate neural networks to estimate the distribution p_{gen}



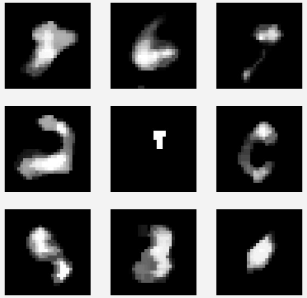
Challenges when training GANs

Low quality samples

- Blurred images, without structure or in the worst case just noise.



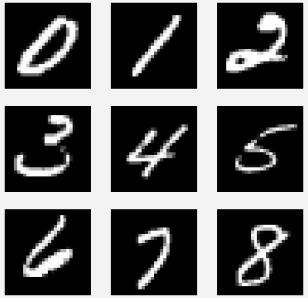
p_{real}



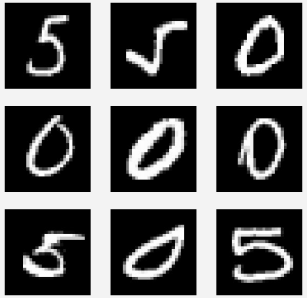
p_{gen}

Mode collapse

- The generator only learns to create a subset of the modes present on the dataset.



p_{real}



p_{gen}

Evaluation metrics for GANs

Inception Score

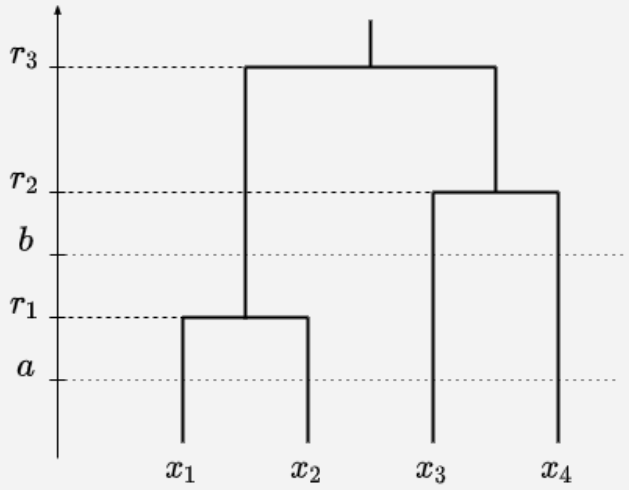
- Uses the probability distribution $p(y|x)$ produced by an Inception-v3 pre-trained on ImageNet.
- Based on entropy of $p(y|x)$ being low and that entropy of $p(y) = \int p(y|x)dx$ being high

Fréchet Inception Distance

- Uses the activations of the last convolutional layer of the Inception-v3.
- Assumes that real and fake data follow a normal distribution and computes the Fréchet distance between the two.

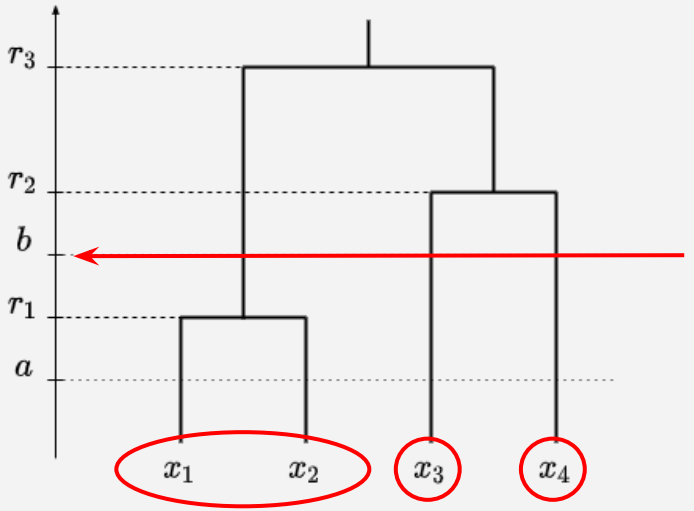
Dendrograms

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



Dendrograms

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



Dendrograms as ultrametric spaces

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

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

- Allows us to use methods from ultrametric spaces, such as the Gromov–Hausdorff distance.

Distance between dendrograms

- The exact Gromov–Hausdorff distance is expensive to compute, but *Costa (2017)* provides an approximation

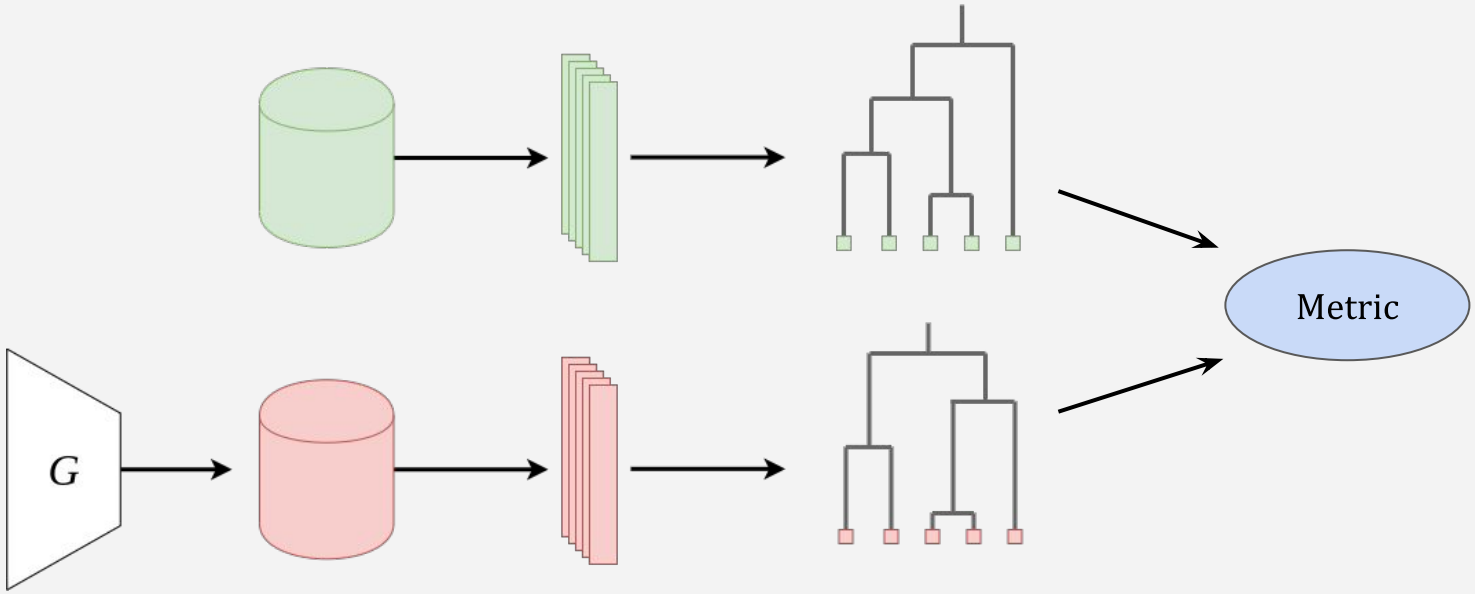
$$\hat{d}_{\mathcal{GH}}(X_\alpha, X_\beta) = \max_i |u_\alpha(i) - u_\beta(i)| \quad , \quad \begin{matrix} u_\alpha(i) \leq u_\alpha(i+1) \\ u_\beta(i) \leq u_\beta(i+1) \end{matrix}$$

- “The greatest difference between the sorted distances”

Proposed method: dendrogram distance

- If 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**

Proposed method: dendrogram distance



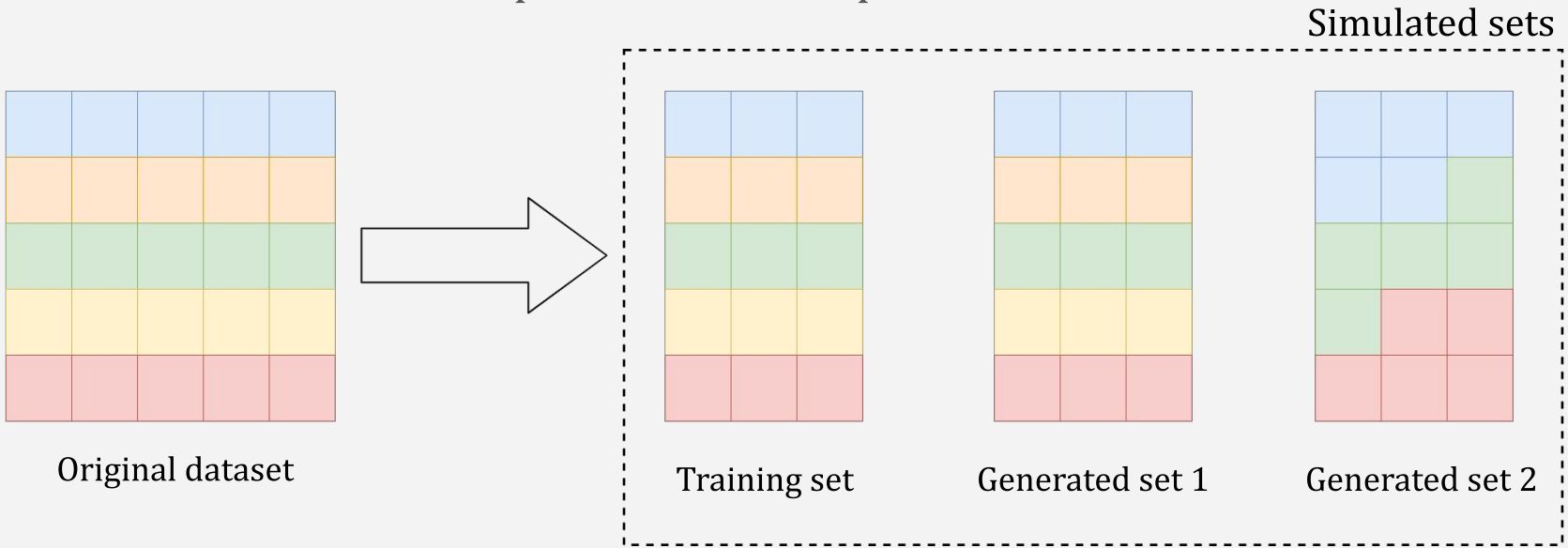
Proposed method: dendrogram distance

- The metric used in our experiments

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

Experiments with real data: mode collapse

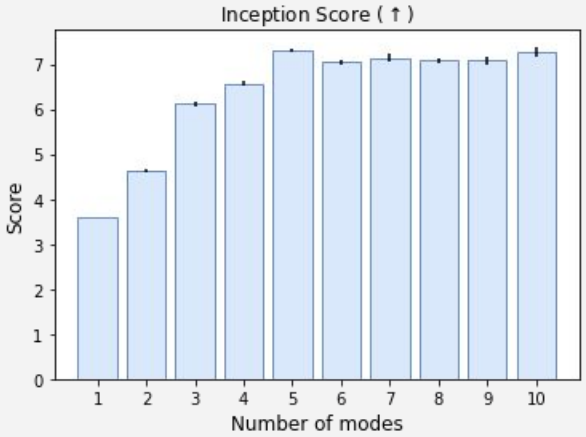
- How to check if metric captures mode collapse?

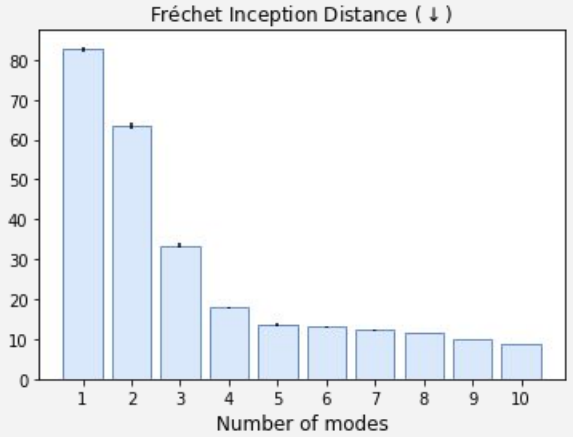


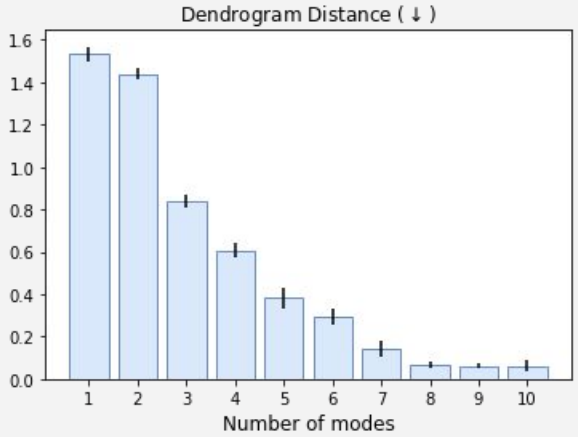
Experiments with real data: mode collapse

- CIFAR-10
- Used output of last convolutional layer of a pre-trained Inception-V3 as data representation
- Inception Score (IS) and Fréchet Inception Distance (FID) were also computed for comparison

Experiments with real data: mode collapse

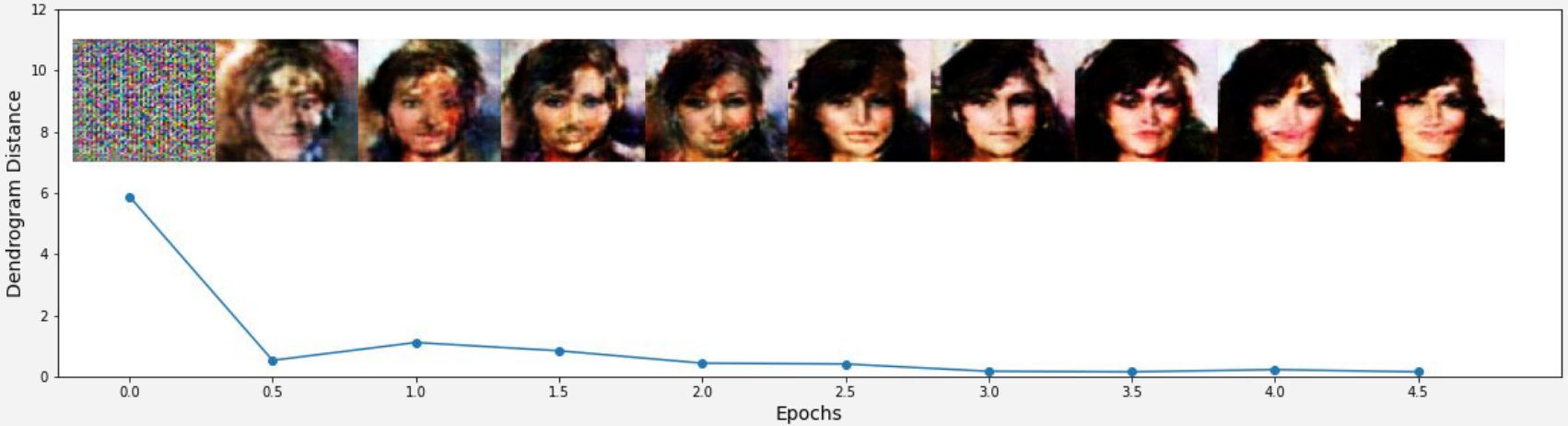






Experiments with real data: metric during training

- Observed how metric evolves during the training procedure
- Trained SAGAN using CelebA dataset



Conclusion

- 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

Thanks.