Assessing Generative Models via Precision and Recall

M. S. M. Sajjadi [1,2,3] O. Bachem [3] M. Lučić [3] O. Bousquet [3] S. Gelly [3]

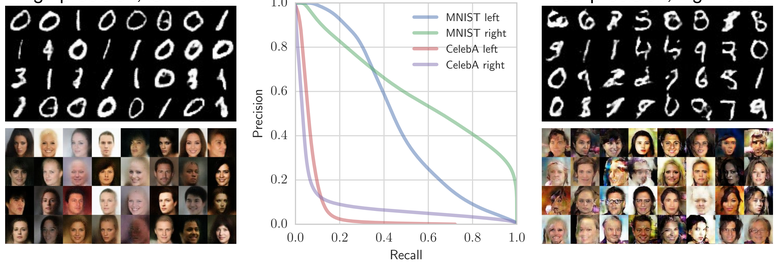
[1] Max Planck Institute for Intelligent Systems [2] Max Planck ETH Center for Learning Systems [3] Google Brain

msajjadi.com

Overview

- **Evaluating generative models is a challenging problem.** Inception Score (IS) and Fréchet Inception Distance (FID) correlate with the perceptual sample quality, but it is easy to find models with the same IS or FID that have highly different characteristics as these are only one-dimensional scores (Figure below).
- **We propose** a score that disentangles *precision* (quality of generated samples) from *recall* (proportion of target distribution that is covered by the generator).
- **Our contributions**: A novel definition of precision and recall for distributions (PRD) with desirable properties, an efficient algorithm to compute it, and we demonstrate that PRD distinguishes mode dropping from mode inventing on real world data sets (image and text data).

High precision, low recall



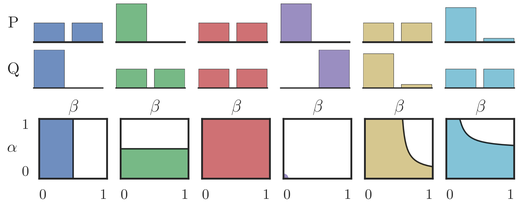
Low precision, high recall

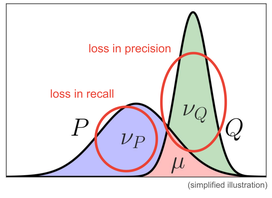
- **Figure above:** Models with similar FID (MNIST: 32/29, CelebA: 65/62) but highly different characteristics. The PRD curves (middle) successfully distinguish between precision and recall.

Definition: Precision and Recall for Distributions (PRD)

- **Goal:** Evaluate distribution $\mathbb{Q}$ w.r.t. a reference distribution $\mathbb{P}$ in terms of precision (*how much of* $\mathbb{Q}$ *is covered by* $\mathbb{P}$) and recall (*how much of* $\mathbb{P}$ *is covered by* $\mathbb{Q}$).
- **Key idea:** Decompose $\mathbb{P}$ and $\mathbb{Q}$ into mixtures of a shared component μ and noise distributions ν_P (loss in recall) and ν_Q (loss in precision).
- **Formal definition:** For $\alpha, \beta \in (0, 1]$, the distribution $\mathbb{Q}$ has precision α at recall β w.r.t. $\mathbb{P}$ if there exist distributions μ, ν_P, ν_Q such that

$$\mathbb{P} = \beta\mu + (1-\beta)\nu_P \quad \text{and} \quad \mathbb{Q} = \alpha\mu + (1-\alpha)\nu_Q.$$

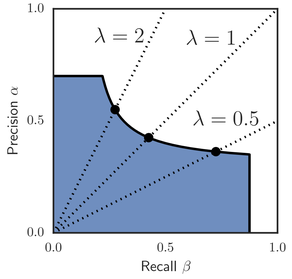




- **Toy examples** for distributions (left, top row) and their respective PRD curves (left, bottom row).

Algorithm

- **How to compute** $\mathrm{PRD}(\mathbb{Q}, \mathbb{P})$**?** There is an infinite number of ways to decompose $\mathbb{P}$ and $\mathbb{Q}$. It is therefore not immediately clear how to compute the optimal set of precision α and recall β.
- **Key insight:** We can fix $\alpha = \lambda\beta$ and iterate over different values for λ.

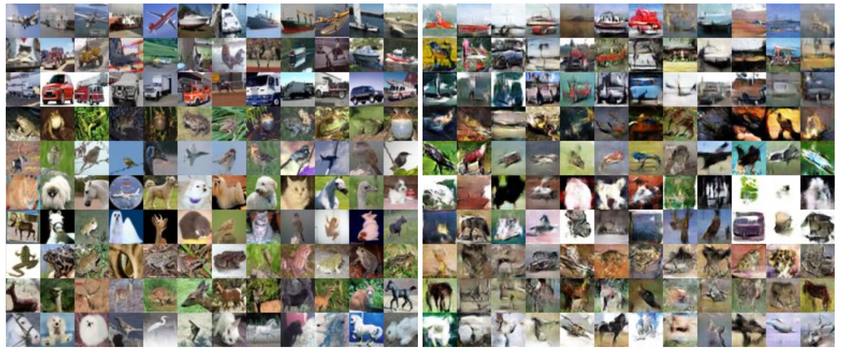


- **We define** $\alpha(\lambda) = \sum_{\omega\in\Omega} \min\left(\lambda\mathbb{P}(\omega), \mathbb{Q}(\omega)\right)$ and $\beta(\lambda) = \sum_{\omega\in\Omega} \min\left(\mathbb{P}(\omega), \mathbb{Q}(\omega)/\lambda\right)$.
- **For a resolution** m, let $\Lambda = \left\{\tan\left(\frac{i}{m+1}\cdot\frac{\pi}{2}\right) \mid i = 1, 2, \ldots, m\right\}$. We then have: $\mathrm{PRD}(\mathbb{Q}, \mathbb{P}) = \{(\alpha(\lambda),\ \beta(\lambda)) \mid \lambda \in \Lambda\}$.
- **Visualization of the PRD algorithm** (left). For each precision-recall ratio λ, we compute the optimal pair of precision and recall via the equation above and add a point to the curve. This gives rise to a PRD curve which shows the relationship between the distributions in terms of precision and recall.

Application to Generative Models

- **Applying the PRD algorithm** in the original space is infeasible as $\mathbb{P}$ and $\mathbb{Q}$ are both only defined through a finite set of samples in the common case of evaluating generative models.
- **We follow this Procedure:**

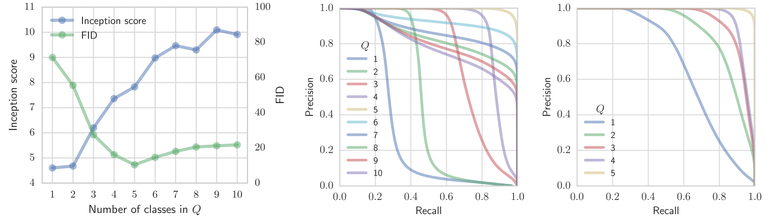
1. **Feature extraction:** Embed both real and generated samples into the *Pool3* layer of a pre-trained Inception network, yielding 2048-dimensional feature vectors.
2. **Clustering:** Cluster the union of both distributions in feature space. The cluster assignment histogram for each one of $\mathbb{P}$ and $\mathbb{Q}$ captures the characteristics of the distribution.
3. **PRD algorithm:** Finally, apply the PRD algorithm on these one-dimensional discrete distributions.



- **Clustered samples** from CIFAR-10 (above) show that the cluster assignments in feature space are meaningful. Each row is sampled from real (left) and generated (right) images of the same cluster.

Disentangling Mode Drop from Mode Inventing

- **Setting:** $\mathbb{P}$ contains first 5 classes, $\mathbb{Q}_i$ has first i classes from CIFAR-10 dataset.
- **Inception Score** (left, blue) linearly increases as we add more classes.
- **FID** (left, green) drops in both cases, but they are not distinguished (e.g., $\mathbb{Q}_4 \approx \mathbb{Q}_6$).
- **PRD** (middle) clearly shows a drop in recall for $i < 5$ and a drop in precision for $i > 5$.
- **Applying PRD on NLP** (right) shows similar behavior. The MultiNLI dataset used here consists of sentences from 5 topics. The embedding is based on a BiLSTM model.



Large-Scale Evaluation

- **Inspecting PRD for 800 generative models** (7 GAN variants, VAE). To summarize each PRD curve into a pair of values for visualization, we show the F_β scores (left):

$$F_\beta = (1+\beta^2)\frac{p \cdot r}{(\beta^2 p) + r}.$$

- **A glance at different models** trained on Fashion-MNIST (right) confirms that the results correlate well with the perceived precision and recall of the samples.
- **In this experiment, GANs generally have a higher precision and lower recall than VAEs.** This follows the *folklore* that GANs produce higher-quality samples than VAEs but they often collapse to generating only part of the dataset.