Improving the Evaluation of Generative Models with Fuzzy Logic

Julian Niedermeier*, **Gonçalo Mordido***, **Christoph Meinel**
Hasso Plattner Institute
Prof.-Dr.-Helmert-Straße 2-3
14482 Potsdam, Germany
julian.niedermeier@student.hpi.uni-potsdam.de, goncalo.mordido@hpi.de, christoph.meinel@hpi.de

Abstract

Objective and interpretable metrics to evaluate current artificial intelligent systems are of great importance, not only to analyze the current state of such systems but also to objectively measure progress in the future. In this work, we focus on the evaluation of image generation tasks. We propose a novel approach, called Fuzzy Topology Impact (FTI), that determines both the quality and diversity of an image set using topology representations combined with fuzzy logic. When compared to current evaluation methods, FTI shows better and more stable performance on multiple experiments evaluating the sensitivity to noise, mode dropping and mode inventing.

Introduction

Accurate evaluation of a model's learning capabilities is of extreme importance to identify possible shortcomings in the model's behavior. When learning a discriminative, supervised task, this evaluation is often straightforward by comparing the model's predictions against ground-truth labels. For example, in an image classification task with labeled data, one can evaluate the agent's label prediction of an image on the test set to its real label.

However, in a generative, unsupervised task, the assessment of a model's capabilities is far more challenging. As an example, considering image generation with unlabeled data using generative adversarial networks (Goodfellow et al. 2014), a model would generate an image from random noise. How can one evaluate the quality of such an image? Moreover, how can one evaluate the diversity of the entirety of the generated set? Answering these questions is the focus of this work.

Our method builds on top of the topological representations created by UMAP's algorithm (McInnes, Healy, and Melville 2018). These topological features can be represented by a directed, weighted graph which first uses the k-nearest neighbors (KNN) algorithm to establish the connections between nodes. Then, such connections are weighted using principles of Riemannian geometry and fuzzy logic, representing the probability of the existence of each directed edge in the resulting graph.

We call our evaluation method Fuzzy Topology Impact (FTI), that has as basis the construction of two of the aforementioned graphs, one for the real and one for the fake data. Then, we analyze the impact that each sample of a given set has on the other set's graph to separately determine the quality and diversity of the fake data set. More precisely, quality is measured by the impact, on average, that a fake sample has on the real data graph, and diversity is measured inversely, by measuring the impact each real sample has on the fake data graph.

In the end, our method can be interpreted as the drop in the average probability of the existence of a connection in the real graph and fake graph, representing the quality and diversity of the fake data. We present the following contributions:

1. Retrieval of two interpretable metrics, which directly correlate to sample quality and diversity.
2. Contrarily to previous topology-based methods, our method can be seen as finer-grained approach due to the usage of fuzzy logic.
3. Thorough experimental discussion of existing evaluation methods, *i.e.* Inception Score (Salimans et al. 2016), Fréchet Inception Distance (Heusel et al. 2017), precision and recall assessment (Sajjadi et al. 2018), and improved precision and recall (Kynkäänniemi et al. 2019), showing the superiority of our approach.
4. Code for the reproducibility of the results is available at https://github.com/sleighsoft/fti.

Related Work

This work primarily focuses on the evaluation of image generation models targeting the evaluation of both image quality and diversity. In general, current approaches can be categorized into three different types: analysis of likelihoods (Theis, Oord, and Bethge 2015) and probability distributions (Heusel et al. 2017; Gretton et al. 2012), topological analysis of manifolds (Sajjadi et al. 2018; Kynkäänniemi et al. 2019; Khrulkov and Oseledets 2018), and classifier-based methods (Salimans et al. 2016; Gurumurthy, Ki-

*Equal contribution.

ran Sarvadevabhatla, and Venkatesh Babu 2017; Shmelkov, Schmid, and Alahari 2018). This work falls within the topological analysis category, where we propose a novel approach that improves existing metrics by following a finer-grained methodology. A description of the methods compared throughout this paper follows.

Inception score or IS (Salimans et al. 2016) analyzes the output distribution of a pre-trained Inception-V3 (Szegedy et al. 2016) on ImageNet (Deng et al. 2009) to measure both the quality and diversity of a fake image set. To this end, they use the Kullback-Leibler Divergence to compare the conditional probability distribution of a fake sample being classified as a given class as well as the marginal distribution of all samples across the existing classes. Higher IS should indicate that each fake sample is clearly classified as belonging to a single class and that all fake samples are uniformly distributed across all existing classes.

Fréchet Inception Distance or FID (Heusel et al. 2017) builds upon the idea of using the Inception-V3 network, but this time to simply obtain feature representations. FID, in contrast to IS, uses the real data distribution and retrieves a distance to the fake data distribution. Therefore, a lower FID is better since it measures the cost of moving mass from one probability distribution onto another to make them identical. Even though FID provides significant improvements over IS, like the detection of intra-class mode dropping where only identical images of each class are generated, it also retrieves a single-valued metric. Therefore, it does not give a direct insight regarding the quality and diversity of the generated set.

To fix this, (Sajjadi et al. 2018) proposed to separate the evaluation into two distinct values, namely precision and recall, by using the relative probability densities of the real and fake distributions. For simplicity, we refer to this approach as Precision and Recall for Distributions (PRD). Thus, precision reflects the quality of generated images, whereas recall quantifies the diversity in the fake image set. Using Inception-V3's features, similarly to FID, for both real and fake samples, they use k-means clustering to group the totality of the samples and evaluate quality and diversity by analyzing the histograms of discrete distributions over the clusters' centers for the real and fake data. Precision and recall values are approximated by calculating a weighted F-Score with $\beta = 8$ and $\beta = \frac{1}{8}$, respectively.

Having concerns about how to appropriately choose β and reliability against mode dropping or truncation, (Kynkäänniemi et al. 2019) proposed to use non-parametric representations of the manifolds of both real and fake data. We refer to this approach as IMproved Precision And Recall (IMPAR). Instead of using Inception-V3, IMPAR uses VGG-16 (Simonyan and Zisserman 2014)'s feature representations. Moreover, instead of determining a set of clusters in the data, as proposed by PRD, IMPAR uses KNN to approximate the topology of the underlying data manifold by forming a hypersphere to the third nearest neighbor of each data point. Precision is then the fraction of points in the fake image set that lie within the real data manifold, whereas recall is the fraction of points in the real image set that lie within the generated data manifold.

Since IMPAR uses a binary overlapping approach to compare the real and fake data manifolds, it lacks into taking into consideration sample density. For example, when dealing with highly sparse data, big regions of the data space may intersect - think of a binary overlapping version of Figure 3(b). This may also be observed when using a high K. In this work, we propose a finer-grained, mathematical sound KNN approach based on fuzzy logic that is sensitive to different overlapping regions depending on the overall sample density.

Fuzzy Topology Impact

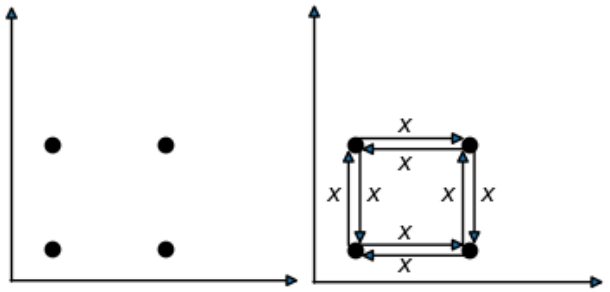
Following the method proposed by UMAP (McInnes, Healy, and Melville 2018), we create a graph where each node represents the embeddings from a pre-trained model of each image. The resulting weighted, directed graph is designed to maintain the topological representations of the embeddings using Fuzzy logic, with each weight representing the *probability of the existence* of a given edge. Then, we measure the drop in the average probability of existence that a new sample has in the original graph, which we call the Fuzzy Topology Impact (FTI). Following this principle, we separately analyze the quality, by calculating the impact that fake samples have in the real samples' graph, and diversity, by measuring the impact that real samples have in the fake samples' graph.

Topological Representation

We will now dive into the underlying properties used by UMAP that enable the data manifold approximation with a fuzzy simplicial set representation in the form of a weighted graph. The geodesic distance from a given point to its neighbors can be normalized by the distance of the k-th neighbor (or by a scaling factor σ), creating a notion of local distance that is different for each point. This notion aligns with the assumption that the data is uniformly distributed on the manifold with regards to a Riemannian metric (see (McInnes, Healy, and Melville 2018) for original lemmas and proofs), which is a requirement for the theoretical foundations from Laplacian eigenmaps (Belkin and Niyogi 2002; 2003) used to formally justify this manifold approximation.

When combining the aforementioned principles with Riemannian geometry, most concretely by connecting each data point using 1-dimensional simplices, we achieve a weighted, directed, k-neighbor graph that represents the approximated manifold. The weight values of the resulting graph are computed using fuzzy logic, which inherently describes the probability of the existence of each edge.

Given N embeddings, $X = \{x_1, \ldots, x_N\}$, and the $k \in \mathbb{N}$ nearest neighbors under the euclidean distance $d \in \mathbb{R}^+$ of each $x_i \in X$, $\{x_{i_1}, \ldots, x_{i_k}\}$, we have the following graph G: $G = (V, E)$, where V represents the embeddings X and E forms a set of directed edges, $E \subseteq \{(x_i, x_{i_j}) \mid j \in \mathbb{N} : j \in [1, k] \wedge i \in \mathbb{N} : i \in [1, N]\}$. Each directed edge $e_{x_i, x_{i_j}} \in E$, is associated with the following weight or probability of existence $p_{x_i, x_{i_j}} \in \mathbb{R}^+ : p_{x_i, x_{i_j}} \in [0, 1]$:



(a) Original samples. (b) Original weighted, directed graph.

Figure 1: Given a set of original samples represented as filled circles (a), we generate a weighted, directed graph using $k = 2$ (b). Since all samples' closest neighbors are at the same distance, the same weight is shared among all edges.

$$p_{x_i,x_{i_j}} = \exp\left(\frac{-d(x_i, x_{i_j})}{\sigma_i}\right), \tag{1}$$

where $\sigma_i \in \mathbb{R}_*^+$ represents the scaling factor associated with x_i such that:

$$\sum_{j=1}^{k} \exp\left(\frac{-d(x_i, x_{i_j})}{\sigma_i}\right) = \log_2(k). \tag{2}$$

Thus, the existence probability associated with each embedding's connections are scaled such that the cardinality of the resulting fuzzy is fixed: $\sum_{j=1}^{k} p_{x_i,x_{i_j}} = \log_2(k)$. Note that $\log_2(k)$ was chosen through an empirical search by the original UMAP's implementation and we re-use this value. Such scaling standardizes the weights of the resulting graph while still maintaining the notion of local connectivity by the usage of individual scaling factors for each embedding.

The resulting graph is weighted and directed, with the corresponding weights representing the probability of existence of the directed connection between a point and respective neighbors. Figure 1 provides a simple illustration of these principles.

Note that there are several differences between our final graph and UMAP's. While we use a directed graph, UMAP combines disagreeing weights to represent the probability of at least one of the edges existing to form an undirected graph. Contrarily to UMAP, we set the local connectivity to 0, meaning that the weight of each sample's closest neighbor is not set to 1.0. This was done to mitigate the influence of outliers in the retrieved impact. Moreover, each node in the graph represents each sample's embeddings from a pre-trained model instead of the sample itself. We found using the embedding information to be more stable in our experiments. Finally, instead of finding a low dimensional representation from the resulting graph, we use the inherent topological information to evaluate generative models, which is described next.

Impact Evaluation

Considering the previously described graph G, we can calculate the average probability of existence of the directed edges by:

$$\overline{P_G} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} p_{x_i,x_{i_j}}}{N \times k}. \tag{3}$$

The proposed evaluation metric is to simply retrieve the average drop of $\overline{P_G}$ when adding a new sample x'_i to the original graph. To achieve this, we modify each weight in the following way:

$$p^{x'_i}_{x_i,x_{i_j}} = \begin{cases} 0, & \text{if } j = k \wedge d(x_i, x_{i_k}) > d(x_i, x'_i) \\ e^{\frac{-d(x_i,x_{i_j})}{\sigma'_i}}, & \text{if } j \neq k \wedge d(x_i, x_{i_k}) > d(x_i, x'_i) \\ p_{x_i,x_{i_j}}, & \text{otherwise.} \end{cases} \tag{4}$$

Hence, if a new sample x'_i is part of the k closest neighbors of an original sample x_i, we remove the connection to the original k'th furthest neighbor, *i.e.* $p^{x'_i}_{x_i,x_{i_k}} = 0$, and update the weight values of the original $k-1$ nearest neighbors according to Eq. 1 and the new σ'_i satisfying Eq. 5. On the other hand, if x'_i is not a k cloest neighbor to any original sample x_i, the original weight values remain unchanged. Figure 2 illustrates these scenarios.

$$\sum_{j=1}^{k-1}\left(\exp\left(\frac{-d(x_i, x_{i_j})}{\sigma'_i}\right)\right) + \exp\left(\frac{-d(x_i, x'_i)}{\sigma'_i}\right) = \log_2(k). \tag{5}$$

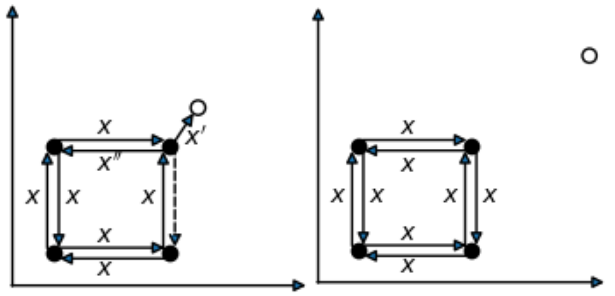
Thus, the drop of average probability of existence of the original connections by a new sample x'_i can be described as:

$$\overline{P_{G,x'_i}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} p^{x'_i}_{x_i,x_{i_j}}}{N \times k}. \tag{6}$$

Finally, having X as the original set used to generate G with k nearest neighbors, and N' new samples $X' = \{x'_1, \ldots, x'_{N'}\}$, FTI can be defined as the average drop of probability of existence of the original connections:

$$FTI(X, X', k) = \frac{\sum_{i=1}^{N'} \overline{P_G} - \overline{P_{G,x'_i}}}{N'}. \tag{7}$$

Algorithm 1 provides a more practical view of the proposed method. Note that the presented pseudo-code is optimized for visualization, not performance. The function $SmoothDistApprox$ executes a binary search that satisfies Equation 5 for the distances passed as argument, similarly to UMAP.



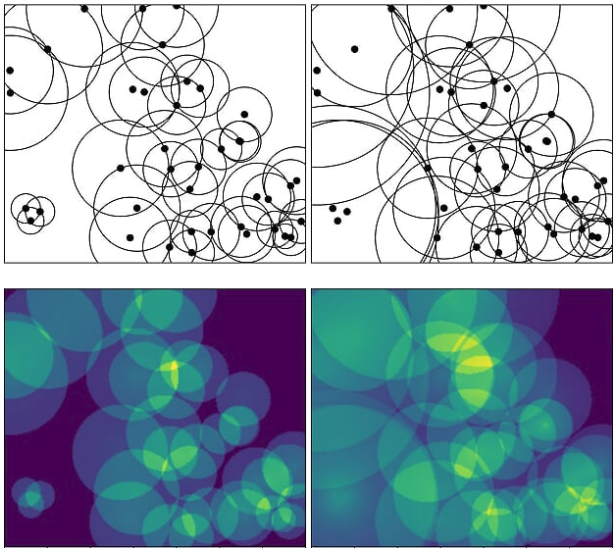
(a) Effects of a new, realistic sample on the original graph. (b) Effects of a new, similar sample on the original graph.

Figure 2: Original samples are represented by filled circles whereas new samples are shown as empty circles. New samples that are the k closest neighbor to a given original point will affect the weights of all directed edges from such point (a). Outlier samples, *i.e.* new samples that are not a closest k neighbor to any original point, cause no impact in the original graph (b).

Algorithm 1 Fuzzy Toplogy Impact. G represents the original graph and $dist$ a dictionary with the euclidean distances of each sample's nearest neighbors.

Require: X, the original set of samples; X', the new set of samples; k, the number of neighbors

$impact \leftarrow 0$
for each $x'_i \in X'$ **do**
  $p^X \leftarrow 0$
  $p^{X'} \leftarrow 0$
  $count \leftarrow 0$
  for each $x_i \in X$ **do**
    if $d(x_i, x'_i) < d(x_i, x_{i_k})$ **then**
      $count \leftarrow count + 1$
      $p^X \leftarrow p^X + p_{x_i, x_{i_k}}$
      del $dists[(x_i, x_{i_k})]$
      $p^{x'_i}_{x_i, x_{i_k}} \leftarrow 0$
      $dists[(x_i, x'_i)] \leftarrow d(x_i, x'_i)$
      $\sigma'_i \leftarrow SmoothDistApprox(dists, k)$
      for $j = 1, \ldots, k-1$ **do**
        $p^X \leftarrow p^X + p_{x_i, x_{i_j}}$
        $p^{x'_i}_{x_i, x_{i_j}} \leftarrow \exp\left(\frac{-d(x_i, x_{i_j})}{\sigma'_i}\right)$
        $p^{X'} \leftarrow p^{X'} + p^{x'_i}_{x_i, x_{i_j}}$
      end for
    end if
  end for
  $impact \leftarrow impact + p^X - p^{X'}$
end for
return $\frac{impact}{N'}$



(a) $k = 2$. (b) $k = 3$

Figure 3: Visualization of the impact of new points given randomly distributed original points using 2 (a) and 3 (b) neighbors. Warmer colors indicate higher impact than cooler colors, with the darkest color indicating no impact.

Number of neighbors The open cover of the manifold is computed by finding the k-nearest neighbors of each original sample. Therefore, using smaller k values promote a more detailed local structure, whereas larger k values induce a larger, global structures. In another words, a higher number of neighbors leads to the resolution of which the topology is approximated to become more diffused, spreading high impact over larger regions.

To visualize such effect of using different number of neighbors in the overall impact, we analyze one toy example with 40 random original samples (Figure 3). The top row shows the original samples in a 2-dimensional space with the radius to the k-th nearest neighbor, while the bottom row presents the impact a new sample would have at any given (x,y)-coordinate.

Quality and Diversity We introduced FTI as the drop in the average probability of existence in the original graph. If we consider the real data as the original sample set R and the generated data as the new sample set G, we can derive both the quality and diversity of the generated data by calculating the bi-directional impact between both sets.

More specifically, quality can be defined as the impact that, on average, a fake sample has on the real data graph. In contrast, diversity is defined as the impact that, on average, a real sample has on the fake data graph. The two metrics are then defined as follows:

$$quality = FTI(R, G, k) \qquad diversity = FTI(G, R, k) \tag{8}$$

Experimental Results

We tested our approach alongside IS, FID, PRD, and IMPAR using three datasets: Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton, and others 2009). The performed experiments evaluate the sensitivity to noise as well as the sensibility to mode dropping and mode inventing. Throughout our experimental setup, we used the training images and testing images of each dataset as real and generated samples, respectively. The embeddings used by our approach were calculated using Inception-V3 due to lower runtime than VGG-16. Since the different compared metrics have different ranges, we analyze the results using their respective ratios.

Noise Sensitivity

To test the sensitivity of the different methods against different amounts of noise, we incrementally added Gaussian noise to the test images of each dataset. Ideally, all methods should show signs of deterioration and, while quality should decrease faster than diversity when little noise is added, both metrics should degrade. Figure 4 shows the comparison results.

We observe that FID is very sensitive to noise with distances growing by an order of magnitude even at almost imperceptible noise amounts. IS is barely perturbed by the noise on Fashion-MNIST and, unexpectedly, shows an increase on CIFAR-10 and CIFAR-100, as well as constant behavior at early noise stages on Fashion-MNIST. Similarly, PRD shows little sensitivity from low to mid noise amounts and then rapidly drops as noise increases. Even though IMPAR and FTI show similar performance, IMPAR shows a faster decrease in diversity over quality, which we argue is not ideal for this experiment. Finally, FTI shows the most levels of sensitivity which we directly link to the fine-grained property of our method.

Mode Dropping

We further simulated mode collapse by first defining a constant window that includes samples from only half of the classes of the different datasets as the real sample set. On the other hand, the test set window slides through the remaining classes, one class at a time, dropping samples from a class represented in the real sample set while adding samples from one unseen class. Ideally, all methods should show a proportional decrease with the number of real classes dropped. Moreover, quality is affected by adding samples from fake classes while diversity is also affected as real classes are removed from the test set. Figure 5 shows the comparison results. Note that IS is excluded from this experiment as it uses a pre-trained classifier on all classes.

We observe that FID almost linearly increases for Fashion-MNIST and CIFAR-100, but stagnates for CIFAR-10 at 3 dropped classes. PRD detects a change in the number of modes for Fashion-MNIST but does not capture mode dropping for CIFAR-10, as its quality first decreases and then increases unexpectedly, and CIFAR-100 where its decrease of both quality and diversity is negligible. IMPAR's diversity fails to detect a decrease in diversity on Fashion-MNIST, even showing an increase on CIFAR-10 when all classes are dropped. Overall, FTI is the most stable approach showing sensitivity to mode drop across all datasets.

Mode Inventing

We replicated (Sajjadi et al. 2018)'s experimental setup to evaluate a different variant of mode collapse and inventing which sheds more light on the importance of using two separate metrics to measure quality and diversity independently. The window of the real set is identical to the last experiment, however, instead of a sliding window for the testing set, we simply add one class at a time, without dropping any class. This way, this experiment measures mode addition until all real classes are present in the test set, and mode invention for additionally added classes. Ideally, the quality remains constant during the mode dropping phase, while diversity increases with each added class. In the mode invention phase, diversity should remain constant whereas quality should decrease as the added classes are not part of the real sample set. Figure 6 shows the comparison results.

On FID, we observe signs of sensitivity to mode collapse, as shown in the previous experiment, however, on CIFAR-10 and CIFAR-100, it fails to punish mode inventing with the overall distance remaining almost constant. Hence, we verify that FID's single-value is unclear with regards to image quality and diversity, as seen on Fashion-MNIST, reinforcing the importance of a separate analysis of quality and diversity. Nevertheless, PRD's quality and diversity behave contradictory to what is expected. Moreover, on CIFAR-10, PRD's diversity stays constant which is also seen on CIFAR-100 for both quality and diversity. IMPAR assigns the same diversity to the class range [0-3] as it does to [0-4] for CIFAR-10 and it lacks to disentangle quality and diversity measures for CIFAR-100. In conclusion, and once more, we see the expected behavior on FTI for this experiment, successfully detecting mode invention and mode dropping across all data sets.

Conclusion and Future Work

Accurately evaluating the performance of machine-generated content is of utmost importance. This work provides an in-depth look at four existing metrics on several experiments using three different datasets and multiple experiments concerning sensitivity to noise and detection of mode dropping and mode inventing. We propose a novel method that evaluates the quality and diversity of generated images using topological representations and fuzzy logic. The experimental results show the overall superiority of the proposed method as well as shortcomings of current approaches.

Since our method simply uses embedding information, it is not limited to image generation tasks solely. Thus, it would be interesting to test the effectiveness of our approach outside image generation, such as text generation tasks. In the future, we plan to extend our evaluation to real-world scenarios to solidify the proposed metrics.

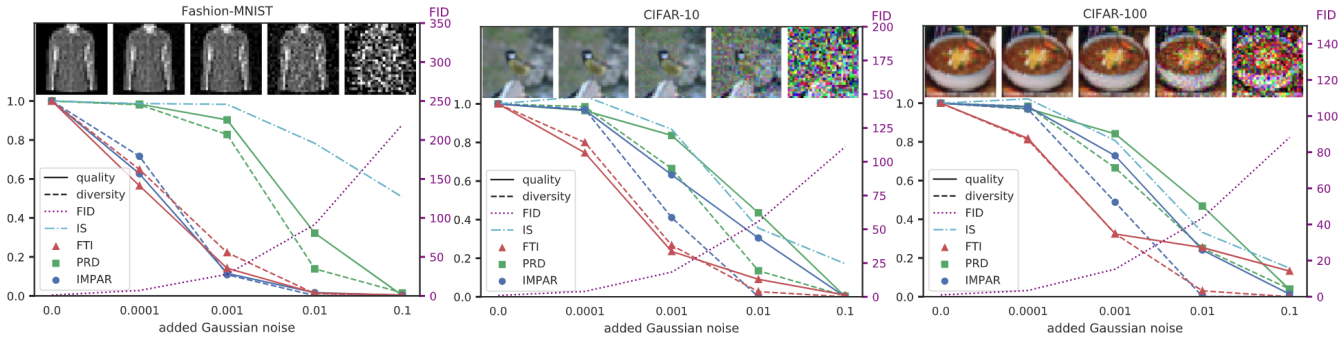


Figure 4: Results for added Gaussian noise on Fashion-MNIST, CIFAR-10 and CIFAR-100. All metrics are normalized by their respective values obtained on unaltered test images, *i.e. no added noise*.

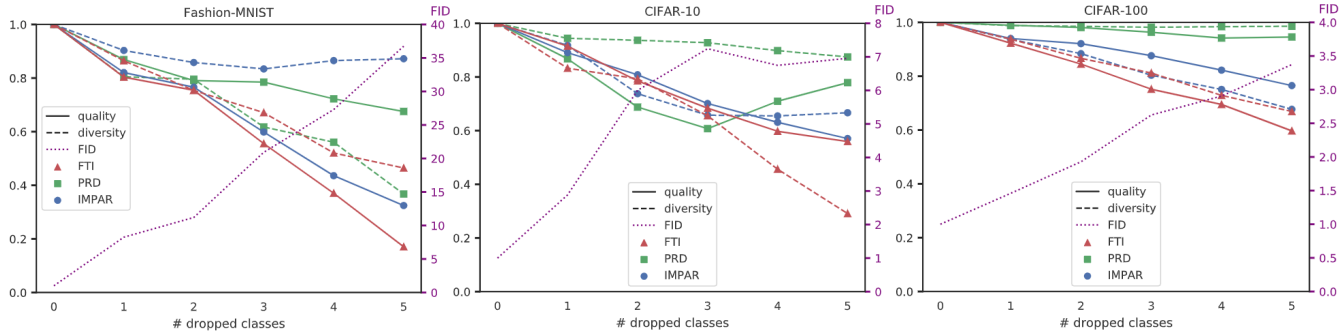


Figure 5: Mode dropping results on Fashion-MNIST, CIFAR-10 and CIFAR-100. Metrics are normalized by their respective values on zero dropped classes.

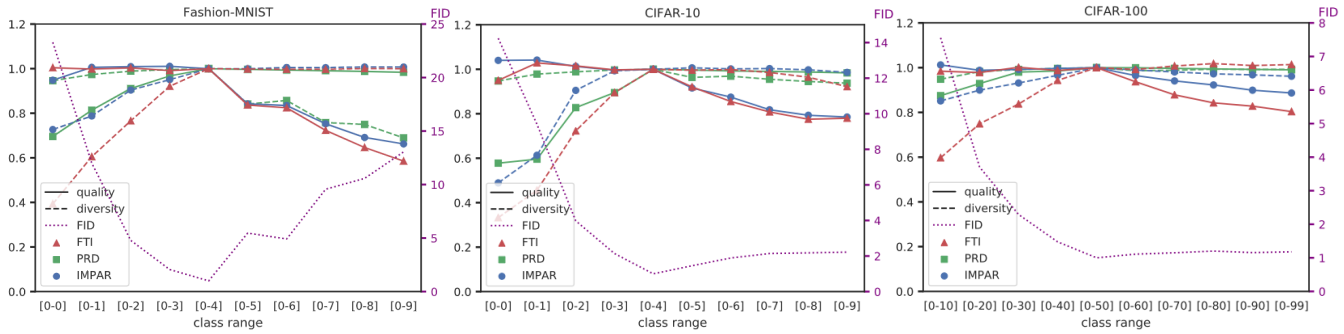


Figure 6: Mode invention experiment on Fashion-MNIST, CIFAR-10, and CIFAR-100. Metrics are normalized by their respective values for [0-4], [0-4], and [0-50] class ranges, respectively.

References

Belkin, M., and Niyogi, P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, 585–591.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15(6):1373–1396.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13(Mar):723–773.

Gurumurthy, S.; Kiran Sarvadevabhatla, R.; and Venkatesh Babu, R. 2017. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 166–174.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Khrulkov, V., and Oseledets, I. 2018. Geometry score: A method for comparing generative adversarial networks. *arXiv preprint arXiv:1802.02664*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Sajjadi, M. S.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 5228–5237.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2018. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–229.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Theis, L.; Oord, A. v. d.; and Bethge, M. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.