Assessing Image and Text Generation with Topological Analysis and Fuzzy Logic

Gonçalo Mordido*, Julian Niedermeier*, Christoph Meinel
Hasso Plattner Institute
Potsdam, Germarny
goncalo.mordido@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. We propose a novel metric, called Fuzzy Topology Impact (FTI), that assesses both the quality and diversity of a generated set using topological representations combined with fuzzy logic. In our synthetic experiments, FTI consistently outperforms current evaluation methods in terms of stability and sensitivity to detect drops in quality and diversity in the generated set, both on image and text generation tasks. Moreover, FTI shows a high degree of correlation to human evaluation on unconditional language generation.

1. 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 model'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 [7], 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 [19]. 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.

Our method, Fuzzy Topology Impact (FTI), has as a 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.

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 [25], Fréchet Inception Distance [10], precision and recall assessment [24], and improved precision and recall [16], showing the superiority of our approach.

4. Code for the reproducibility of the results is available at `https://github.com/sleighsoft/fti`.

2. Related Work

This work primarily focuses on the evaluation of generative models targeting assessing both the quality and the diversity of the generated set. In general, current approaches can be categorized into three different types: analysis of likelihoods [32] and probability distributions [10, 8], topological analysis of manifolds [24, 16, 13, 20], and classifier-

*Equal contribution.

based methods [25, 9, 28]. 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 [25] analyzes the output distribution of a pre-trained Inception-V3 [31] on ImageNet [6] to measure both the quality and diversity of a generated set. 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 [10] builds upon the idea of using the Inception-V3 network, but simply to obtain feature representations. In contrast to IS, FID uses the real data distribution and retrieves a distance to the fake data distribution. Therefore, a lower FID is better since it indicates the fake distribution approximates the real one. Even though FID provides significant improvements over IS, like the detection of mode dropping where only identical samples 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*. [24] 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 samples, whereas recall quantifies the diversity in the generated 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, Kynkaanniemi *et al*. [16] 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 [29]'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 generated set that lie within the real data manifold, whereas recall is the fraction of points in the real 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 2(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.

3. Fuzzy Topology Impact

Following the method proposed by UMAP [19], we create a graph where each node represents the embeddings from a pre-trained model of each sample. 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.

3.1. 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 [19] for original lemmas and proofs), which is a requirement for the theoretical foundations from Laplacian eigenmaps [1, 2] 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 $p_{x_i,x_{i_j}} \in \mathbb{R}^+ : p_{x_i,x_{i_j}} \in [0,1]$:

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

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 set 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 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.

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.

3.2. 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 1 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:

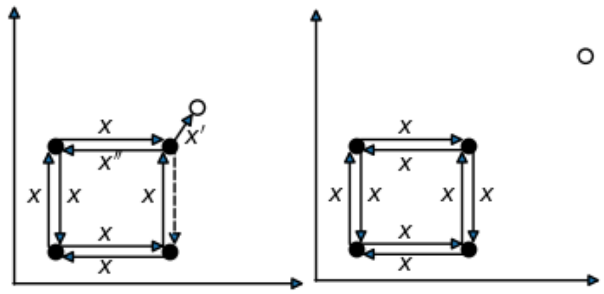
$$FTI(X, X', k) = \frac{\sum_{i=1}^{N'} \overline{P_G} - \overline{P_{G,x'_i}}}{N'}. \tag{7}$$

Algorithm 1 presents 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.

3.2.1 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 2). The top



(a) Effects of a new, realistic sample on the original graph ($k = 2$).

(b) Effects of a new, similar sample on the original graph ($k = 2$).

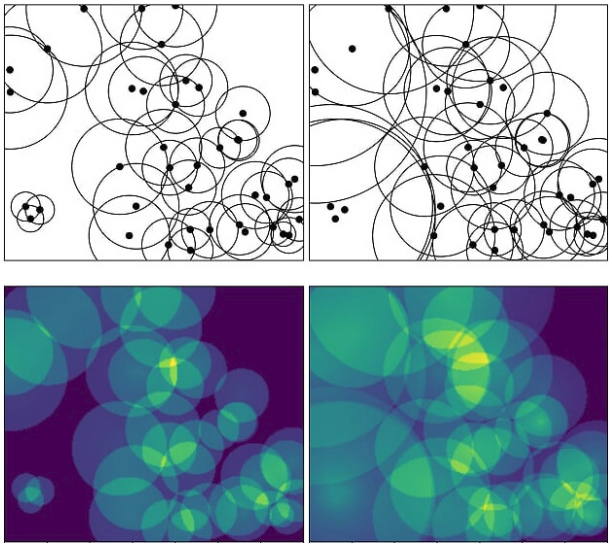
Figure 1. Original samples are represented by filled circles whereas new samples are shown as empty circles. New samples that are the k closest neighbor to a given original point will affect the weights of all directed edges from such point (a). Outlier samples, *i.e.* new samples that are not a closest 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 $\frac{impact}{N'}$

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.



(a) $k = 2$.

(b) $k = 3$

Figure 2. 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.

3.2.2 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:

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

4. Experimental Results

We tested our proposed metric, FTI, on several synthetic experiments that simulate a drop in quality and/or diversity of the generated set. Such experiments cover both image (Sections 4.1 and 4.2) and text (Section A.1) generation. Moreover, we evaluated FTI in terms of correlation to human evaluation on language generation (Section 4.3). All the experiments are performed using the default k of each method, particularly $k = 3$ for FTI and IMPAR.

4.1. Synthetic Experiments

We first tested our approach alongside IS, FID, PRD, and IMPAR on three image datasets: Fashion-MNIST [34], CIFAR-10 and CIFAR-100 [15]. The performed experiments evaluate the sensitivity to noise (Section 4.1.1), mode dropping (Section 4.1.2) as well as mode addition and mode invention (Section 4.1.3). 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.

4.1.1 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 3 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.

4.1.2 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 4 shows the comparison results. Note that IS is excluded from this

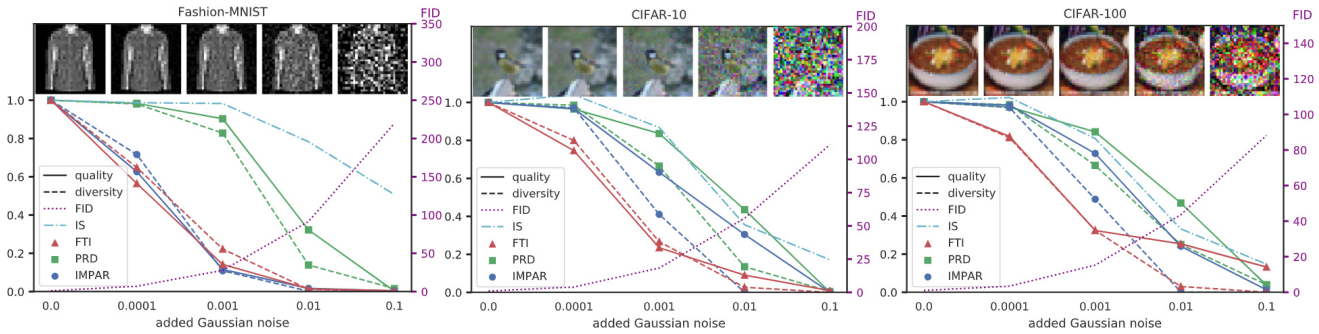


Figure 3. Results for added Gaussian noise on Fashion-MNIST, CIFAR-10 and CIFAR-100. All metrics are normalized by their respective values obtained on unaltered test images, *i.e.* without added noise.

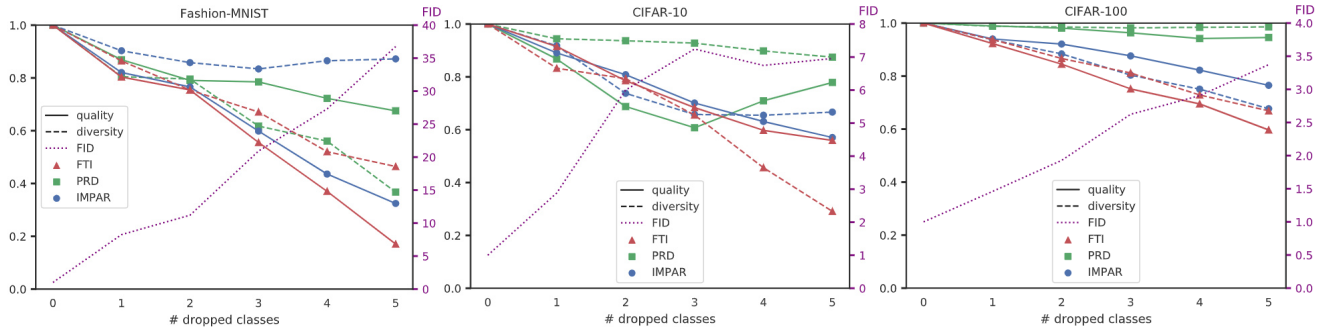


Figure 4. Mode dropping results on Fashion-MNIST, CIFAR-10 and CIFAR-100. Metrics are normalized by their respective values on zero dropped classes.

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 on mode dropping across all datasets.

4.1.3 Mode Addition & Invention

Inspired by Sajjadi *et al*. [24]'s experimental setup, we evaluated 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. Note that, since the cardinality of the test set changes, we do not normalize FTI by the number of original connections in this experiment. Thus, we evaluate 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 5 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 addition and mode invention across all data sets.

4.2. Experiments on StyleGAN and StyleGAN2

To assess our method generated images with high-fidelity, we studied the current state-of-the-art architecture for unconditional generative image modeling as of the writing of this work: StyleGAN2 [12]. We manipulated the quality and diversity of the generated set by using different pre-trained StyleGAN2 models trained with different truncation ψ, as this hyperparameter is known to provide a tradeoff between image quality and diversity [16, 18, 3, 11, 14]. The different models were trained on LSUN-Church [35], LSUN-Horse [35], and FFHQ [11]. We also performed identical experiments for its predecessor, StyleGAN [11], on LSUN-Bedroom [35], LSUN-Cat [35], and FFHQ. All images were rescaled to 256x256 for all the experiments.

Figures 6 and 7 shows the effect of the different models on all the previously compared evaluation metrics. Ideally, as truncation ψ increases, the diversity of the generated set also increases, whereas quality assessment is expected to decrease. We observe that FTI correctly responds to the increase in truncation, with a more pronounced change in quality and diversity when compared to the compared methods across all the datasets. It can be further observed that PRD fails to exhibit a meaningful response to the change of truncation values.

To empirically study the effects of using a different k in our method, we varied the number of neighbors when calculating the k-nearest neighbors algorithm and re-conducted

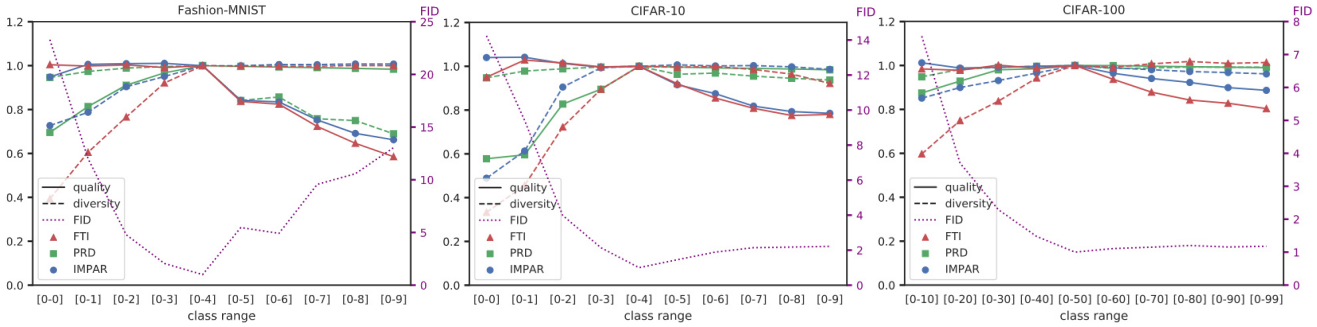


Figure 5. Mode invention experiment on Fashion-MNIST, CIFAR-10, and CIFAR-100. Metrics are normalized by their respective values for [0-4], [0-4], and [0-50] class ranges, respectively.

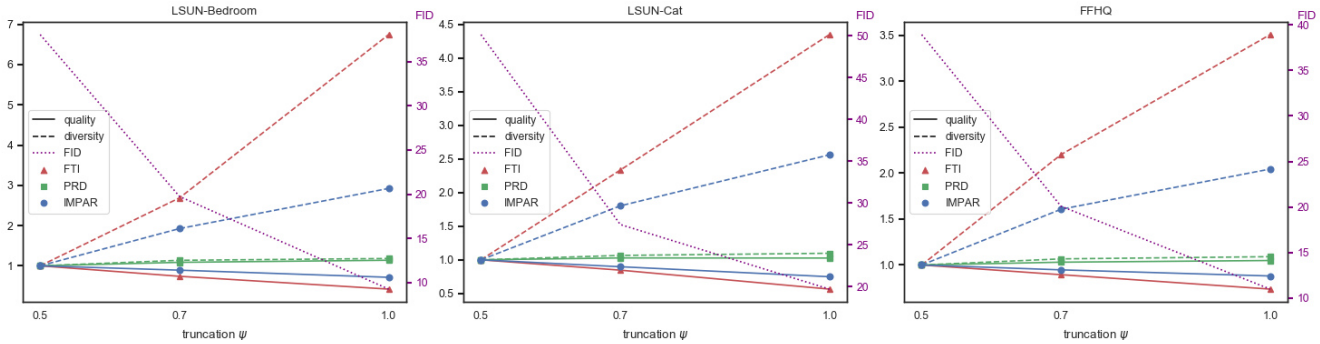


Figure 6. Truncation variation of pre-trained StyleGAN models on LSUN-Bedroom, LSUN-Cat and FFHQ. Metrics are normalized by their respective values at the lowest truncation value.

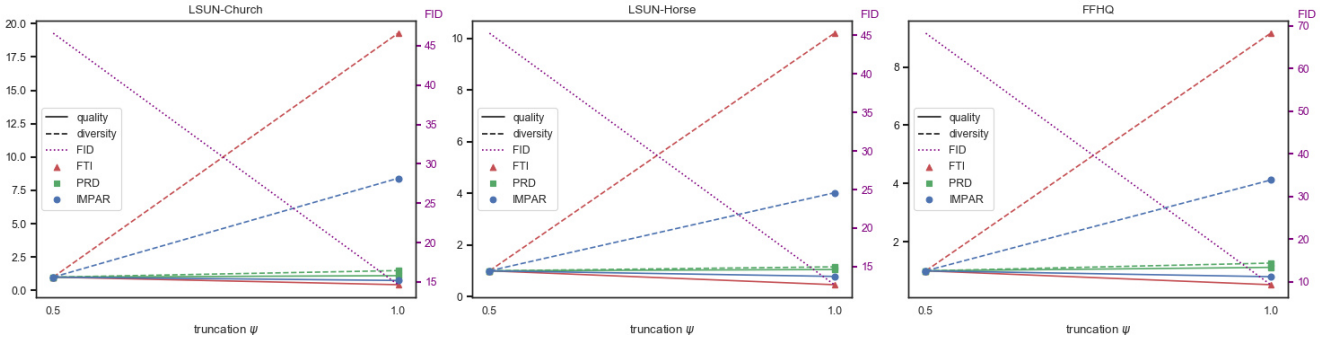


Figure 7. Truncation variation of pre-trained StyleGAN2 models on LSUN-Church, LSUN-Horse and FFHQ. Metrics are normalized by their respective values at the lowest truncation value.

the previous experiment in Figures 8 and 9. In general, we observe that the change in the ratio of diversity assessment is higher when using fewer neighbors. This result further supports our previous observations in Section 3.2.1.

4.3. Correlation with Human Evaluation

We further evaluated the correlation of our metric to human evaluation on language generation. To achieve this, we use the human scores assigned to 10 different language models presented by Cífka *et al*. [5]. In their work, they generated 200 sentences from a variety of autoencoders as well as a language model and asked 3 human evaluators to assess them on a 5-point Likert scale, ranging from gibberish (1) to human-generated sentences (5). In the end, each model was assigned a score representing the average of its sentence median.

We assessed each of the 10 language models using the previously discussed metrics by generating 10k sentences per model and comparing it to 10k real samples provided by the authors. Since we are now dealing with textual data, we get the embedding representation of each sentence by feeding them to the Universal Sentence Encoder or USE [4]. Moreover, we compared all metrics against the forward and reverse cross-entropy results provided by Cífka *et al*. [5], which were calculated using a pre-trained language model trained on English Gigaword [23].

Methods	**Forward CE**	**Reverse CE**	**FD**	**PRD**	**IMPAR**	**FTI**
Pearson (r)	0.606	0.440	0.879	0.702	0.860	**0.910**
Spearman (p)	0.697	0.491	**0.952**	0.879	0.915	0.927

Table 1. Absolute Pearson and Spearman correlations to human evaluation on language generation. FTI shows the highest Pearson correlation out of all methods while outperforming all methods except FID on Spearman correlation.

To evaluate the assessment assigned to each model by each evaluation metric, we calculated the Pearson and Spearman correlations to human evaluation. Results are shown in Table 1. FTI shows a higher Pearson correlation than all the other metrics. Regarding Spearman correlation, FTI is only outperformed by FID. Nevertheless, our metric shows a high degree of ranking correlation to human evaluation.

A visualization of the rankings of the different models obtained by using each evaluation metric is further illustrated in Figure 10. It is easy to observe that the recently proposed topological metrics, as well as FID, show a higher resemblance to human-assessed ranking than more traditional metrics, *i.e.* Forward CE and Reverse CE. Additional experiments on evaluation of sentence quality and diversity can also be found in the Appendix.

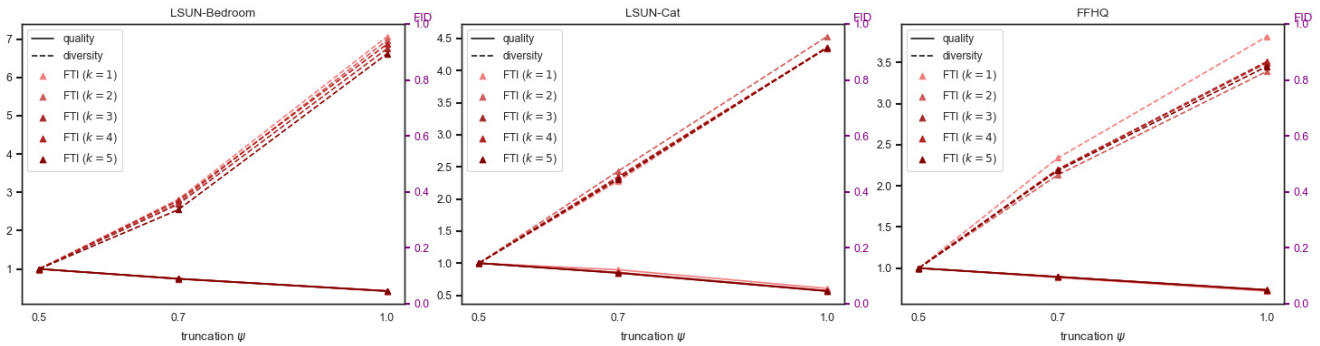


Figure 8. Truncation variation of pre-trained StyleGAN models using a different number of neighbors k on LSUN-Bedroom, LSUN-Cat and FFHQ. Results are normalized by their respective values at the lowest truncation value.

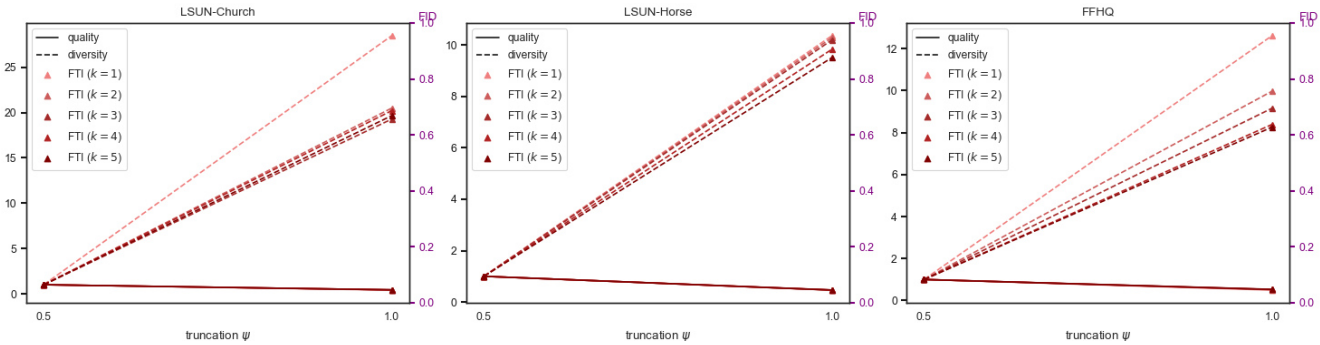


Figure 9. Truncation variation of pre-trained StyleGAN2 models using a different number of neighbors k on LSUN-Church, LSUN-Horse and FFHQ. Results are normalized by their respective values at the lowest truncation value.

5. Conclusion

Accurately evaluating the performance of machine-generated content is of utmost importance. More specifically, assessing the generated data both in terms of quality and diversity may help in improving the generation process by shedding some light on where a specific generation system is lacking. Stimulating sample diversity while maintaining high-quality samples is an active and important area of research in generative models [17, 21, 22, 30, 26].

On top of the proposal of a novel and effective evaluation metric, this work provides an in-depth look at the performance of recently proposed metrics on several synthetic experiments as well as in terms of correlation to human evaluation. With a wide range of the performed experiments, which utilize both image and textual data, we showed the overall superiority of our method, as well as the shortcomings of current approaches.

In the future, we plan to extend this study to finer-grained textual representations, such as contextualized word embeddings. This would enable a thorough comparison of the current methods, as well as our own, for conditional language generation, such as machine translation and text summarization. In the end, an ideal evaluation metric should be able to be applied in different contexts and data types, ranging from unconditional to conditional data generation tasks.

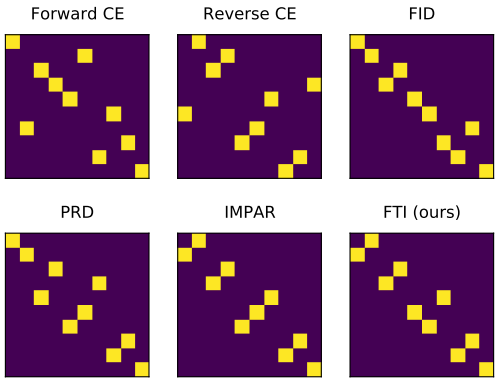


Figure 10. Ranking assessment of the 10 language models with different metrics. Diagonal representations show a higher correlation to human evaluation.

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for

dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[5] Ondřej Cífka, Aliaksei Severyn, Enrique Alfonseca, and Katja Filippova. Eval all, trust a few, do wrong to none: Comparing sentence generation models. *arXiv preprint arXiv:1804.07972*, 2018.

[6] 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.

[7] 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.

[8] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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

[10] 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.

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. *arXiv preprint arXiv:1802.02664*, 2018.

[14] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[16] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*, 2019.

[17] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in neural information processing systems*, pages 1498–1507, 2018.

[18] Marco Marchesi. Megapixel size image creation using generative adversarial networks. *arXiv preprint arXiv:1706.00082*, 2017.

[19] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[20] Gonçalo Mordido and Christoph Meinel. Mark-evaluate: Assessing language generation using population estimation methods. *arXiv preprint arXiv:2010.04606*, 2020.

[21] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators. *arXiv preprint arXiv:1807.11346*, 2018.

[22] G. Mordido, H. Yang, and C. Meinel. microbatchgan: Stimulating diversity with multi-adversarial discrimination. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3050–3059, 2020.

[23] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*, AKBC-WEKEX '12, page 95–100, USA, 2012. Association for Computational Linguistics.

[24] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pages 5228–5237, 2018.

[25] 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.

[26] Jonathan Sauder, Ting Hu, Xiaoyin Che, Goncalo Mordido, Haojin Yang, and Christoph Meinel. Best student forcing: A simple training mechanism in adversarial language generation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4680–4688, Marseille, France, May 2020. European Language Resources Association.

[27] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

[28] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2018.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.

[31] 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.

[32] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[33] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[34] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[35] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.