Fréchet Inception Distance: Reliability and Robustness in Evaluating Adversarial Images (Research Proposal)

Cody Fang

August, 2023

Introduction

With the prevalence and advancement of generative models in computer vision such as Generative Adversarial Networks (GAN) and the Diffusion models, the outputs of these models are also growing increasingly more complex. This necessitates a precise evaluation metrics that aids in determining the performance of such models, more specifically, measuring the realism of a generated image in comparison with real photographs.

One of the metrics that claims to perform this task is the Fréchet Inception Distance (FID). It is calculated by computing a similarity metric (Fréchet Distance) between the probability distributions of features (extracted with the Inception-v3 Neural Network) from both real and generated images. However, this metric has several flaws, most prominently that it has been discovered to be sensitive to minor and unnoticeable image manipulations (resulting images referred to as "adversarial images"). Generative models that produce the best FID score could simply be repeating FID sensitive patterns that do not correlate with perceived realism. Thus, FID could be inadequate in assessing model performance, and its ubiquitous usage further contributes to the significance of this issue.

Objective

This project aims to investigate and test the sensitivity of Fréchet Inception Distance score to adversarial images, and find image modifications that produce substantial increments to the metric. Note that this experiment will not attempt to rectify this issue fully, but rather to find insights and patterns that may lead to further investigations.

Related Work

In current literature, there is compelling criticism of the FID score, and some alternatives have been proposed[1]. However, there is only little empirical examination of the FID score and its limitations. One such research found that image resizing and the supposedly inconsequential compression steps required to perform this have resulted in surprising increases in FID[2]. A specific example includes an identical photograph compressed using both JPEG

100 and 75 algorithms, with researchers observing a subsequent 20 point increase in FID despite both images appearing identical.

Others have also challenged FID's underlying assumptions[3], and the suitability of using the Inception-v3 Neural Network in its latent feature extraction step. A particular research concluded that FID rates images of the human face contrary to human intuition as a result of inadequate representation of domain-specific features[4]. By extension, we see that unpredictable behaviour in the FID algorithm is theoretically possible.

However, these works are limited in the range of image operations tested. Furthermore, similar experiments have only been performed with facial recognition data sets. As a result, the severity of the problem with FID on a broader scope is largely unclear, and a more comprehensive investigation could be benefitial.

Methodology

To test for a wide range of potential use cases for the FID as well as to ensure comprehensive coverage of features, the following data sets are included in this experiment.

- MNIST (handwritten of numerical digits)
- CIFAR-100 (collection of animals and objects)
- CelebA (images of human faces)

We will first perform a variety of minor image processing operations on selected samples that does not change the perceived realism of images. Examples of this include such as resizing, rotating, adjusting gamma, contrast. In addition, we will experiment with adding low levels of structured noise. The observed differences in FID scores will be compared to identify operations to which FID is most sensitive.

Subsequently, we look more extensively into three or four image processing effects that leads to the greatest FID disturbance, and apply these to all data sets to quantify its overall impact. We also identify patterns in FID behaviour in response to adversarial images. We will furthermore apply different combinations of these image effects to test for compounding effects on FID score.

Expected Results and Analysis

In our results, we expect to find bias in the FID score when comparing processed and unprocessed images. At the very least, previously demonstrated results should be replicated. The inconsistencies found will also be verified with human visual judgement.

We analyse the effect of image processing on FID score across different data sets by simply computing both the mean FID score, and the mean FID score difference. Comparing these on a data set by data set basis (given that the images are real) gives an accurate indication of factors contributing to FID sensitivity. There is potentially scope for running these modifications on images generated from a Stable Diffusion Model. If time allow, we plan to do preliminary reading and research to investigate potential causes and potentially to propose a modified FID metric that avoids the weaknesses we find.

Impact

Exposing uncertainties in the FID score will raise awareness about its drawbacks as an objective evaluation metric of generated images. Hence, it will encourage model developers to reconsider using this metric as definitive proof of performance, but rather as a reference score with room for improvement in its discerning capabilities.

References

[1] Min Jin Chong and David A. Forsyth. "Effectively Unbiased FID and Inception Score and where to find them". In: *CoRR* abs/1911.07023 (2019). arXiv: 1911.07023. URL: http://arxiv.org/abs/1911.07023.

[2] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. "On Buggy Resizing Libraries and Surprising Subtleties in FID Calculation". In: *CoRR* abs/2104.11222 (2021). arXiv: 2104.11222. URL: https://arxiv.org/abs/2104.11222.

[3] George Stein et al. *Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models.* 2023. arXiv: 2306.04675 [cs.LG].

[4] Shaohui Liu et al. "An Improved Evaluation Framework for Generative Adversarial Networks". In: *CoRR* abs/1803.07474 (2018). arXiv: 1803.07474. URL: http://arxiv.org/abs/1803.07474.