Supplementary Material for Rethinking FID: Towards a Better Evaluation Metric for Image Generation

Sadeep Jayasumana Srikumar Ramalingam Andreas Veit Daniel Glasner
Ayan Chakrabarti Sanjiv Kumar
Google Research, New York
{sadeep, rsrikumar, aveit, dglasner, ayanchakrab, sanjivk}@google.com

A. Synthetic Experiment Details

In this section, we discuss the details of the experiment described in Section 3.2. As the reference distribution, we use an isotropic Gaussian distribution centered at the origin with a covariance matrix $\sigma^2\mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. The second distribution consists of four different equally-likely Gaussians, centered at the coordinates $(\lambda, 0), (0, \lambda), (-\lambda, 0), (0, -\lambda)$, and each with the covariance matrix $\tau_\lambda^2\mathbf{I}_2$. In Table 2, we show the distribution visualizations (first row), and the behavior of different distance metrics (remaining rows) with increasing values of λ. As λ increases, τ_λ is adjusted as described below so that the overall covariance matrix of the mixture-of-Gaussians distribution remains equal to $\sigma^2\mathbf{I}_2$. Trivially, the mean of the mixture-of-Gaussians is the origin. Therefore, as λ varies, both the mean and the covariance matrix of the mixture-of-Gaussians distribution remain equal to the reference distribution. Therefore, both FD and FD∞ estimated using Eq. (2) remain zero as λ increases. This is obviously misleading as the mixture-of-Gaussians distribution gets further and further away from the reference as λ increases. This error is a direct consequence of the incorrect normality assumption for the mixture-of-Gaussians distribution.

To see the relationship between τ_λ and λ that keeps the overall covariance matrix equal to $\sigma^2\mathbf{I}_2$, consider a mixture distribution consisting of 1-D PDFs $f_1, f_2, \dots, f_n$ with weights $p_1, p_2, \dots, p_n$, where each $p_i > 0$ and $\sum_i p_i = 1$. The PDF of the mixture distribution is then given by $f(x) = \sum_i p_i f_i(x)$. It follows from the definition of the expected value that, $\mu^{(k)} = \sum_i p_i \mu_i^{(k)}$, where $\mu^{(k)}$ and $\mu_i^{(k)}$ are the k^{th} raw moment of f and f_i, respectively. Recall also that variance is $\mu^{(2)} - \{\mu^{(1)}\}^2$. By applying the above result to x and y coordinates individually, we see that the overall covariance matrix of the above mixture of four Gaussians, when they are away from the mean by λ, is given by $(\tau_\lambda^2 + \lambda^2/2)\mathbf{I}_2$. Setting $\tau_\lambda^2 = \sigma^2 - \lambda^2/2$ therefore keeps the overall covariance matrix at $\sigma^2\mathbf{I}_2$ as we vary λ.

B. Multivariate Normality Tests

Fréchet Inception Distance (FID) hinges on the multivariate normality assumption. Since there is no canonical test, we show that the Inception features for a typical image dataset like COCO 30K do not satisfy this assumption using three different widely-accepted statistical tests: Mardia's skewness test [4], Mardia's kurtosis test [4] and Henze-Zirkler test [3].

The null hypothesis for all of the tests is that the sample is drawn from a multivariate normal distribution. Different tests use different statistics as described below.

Mardia's Skewness Test

For a random sample of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$, a measure of multivariate skewness is,

$$A = \frac{1}{6n} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big[(\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}) \Big]^3 . \tag{A}$$

Where $\hat{\mathbf{\Sigma}}$ is the biased sample covariance matrix, and $\bar{\mathbf{x}}$ is the sample mean.

Mardia [4] showed that under the null hypothesis that $\mathbf{x}_i$s are multivariate normally distributed, the statistic A will be asymptotically chi-squared distributed with $d(d+1)(d+2)/6$ degrees of freedom. Therefore, the normality of a given sample can be tested by checking how extreme the calculated A-statistic is under this assumption. For Inception embeddings computed on the COCO 30K dataset, this test rejects the normality assumption with a p-value of 0.0, up to machine precision.

Mardia's Kurtosis Test

For a random sample of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, a measure of multivariate kurtosis is,

$$B = \sqrt{\frac{n}{8d(d+2)}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\,\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\right]^2 - d(d+2)\right\}. \tag{B}$$

It was shown in [4] that, under the null hypothesis that $\mathbf{x}_i$s are multivariate normally distributed, the statistic B will be asymptotically standard normally distributed. For Inception embeddings computed on the COCO 30K dataset, this test also rejects the normality assumption with a p-value of 0.0, To intuitively understand the confidence of the outcome: this Mardia's test places the test statistics $19,023$ standard deviations away from the mean in a normal distribution. This indicates the test's extreme confidence in rejecting the normality of Inception embeddings.

Henze-Zirkler Test

The Henze-Zirkler test [3] is based on a functional that measures the distance between two distributions and has the property that, when one of the distributions is standard multivariate normal, it is zero if and only if the second distribution is also standard multivariate normal. The Henze-Zirkler test has been shown to be affine invariant and to have better power performance compared to alternative multivariate normal tests.

The Henze-Zirkler test's p-value for Inception embeddings of COCO 30K is again 0.0 up to the machine precision. Therefore, the Henze-Zirkler test also rejects the normal assumption on Inception embeddings with overwhelmingly high confidence.

C. CMMD Kernel

Here we explain in more detail why we prefer the Gaussian RBF kernel, rather than the third-degree polynomial kernel used in KID. The Gaussian kernel is *characteristic* and therefore yields a *metric* [1, 2], whereas, the polynomial kernel is not characteristic and does not a yield a metric. To explain this intuitively, the kernel in KID captures moments only up to 3 degrees. It therefore fails to distinguish between distributions that differ in higher orders (just like FD in Table 2 fails to capture differences beyond 2 degrees). The Gaussian RBF kernel has an infinite-dimensional feature map and is therefore able to capture infinitely-high order interactions. This can be clarified with the Taylor series expansion $\exp(-\|\mathbf{x}-\mathbf{y}\|^2) = \exp(-2)\sum_{i=0}^{\infty}\frac{2^i}{i!}\left(\mathbf{x}^T\mathbf{y}\right)^i$ (CLIP embeddings are l^2-normalized): the Gaussian kernel is an infinite series of *all* polynomial kernels. It is therefore much richer than the 3rd degree polynomial. It also has no additional computational overhead over the polynomial kernel.

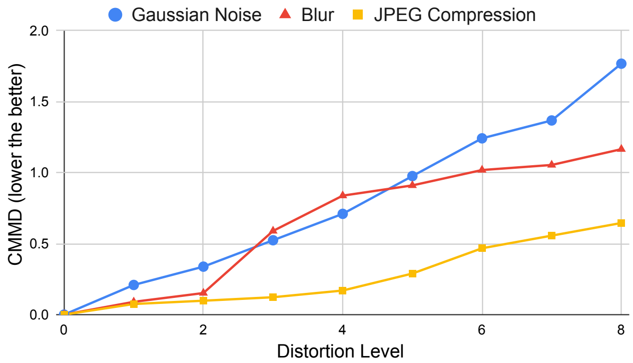

Figure A. *CMMD's behavior under different distortions to images.*

D. Behavior of CMMD under Image Distortions

Figure A shows that CMMD increases with increased levels of image distortions such as Gaussian blue, Gaussian noise and JPEG compression artifacts. This further confirms that CMMD accurate measures the image quality.

References

[1] Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *NeurIPS*. Curran Associates, Inc., 2008. 2

[2] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, 2012. 2

[3] Norbert Henze and Bernd Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, 1990. 1, 2

[4] K. V. Mardia. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika*, 1970. 1, 2