WIND: WASSERSTEIN INCEPTION DISTANCE FOR EVALUATING GENERATIVE ADVERSARIAL NETWORK PERFORMANCE

P. Dimitrakopoulos[1], *G. Sfikas*[1,2], *Christophoros Nikou*[1]

[1]Dept. of Computer Science & Engineering, University of Ioannina, Greece
[2] CERTH/ITI, Thessaloniki, Greece

ABSTRACT

In this paper, we present Wasserstein Inception Distance (WInD), a novel metric for evaluating performance of Generative Adversarial Networks (GANs). The proposed metric extends on the rationale of the previously proposed Fréchet Inception Distance (FID), in the sense that GAN performance is quantified in terms of data and model distribution divergence. We extend FID by relaxing the Gaussian hypothesis of the related inception features and extend it for non-Gaussian, multimodal distributions. Gaussian Mixture Models (GMMs) are used to model data and model inception features, and the Wasserstein distance is employed as a pdf matching metric. We show that the proposed WInD metric inherits the desirable features of FID and correlates well with actual GAN performance. Furthermore, WInD can correctly evaluate cases were data and model distribution erroneously would appear as well peforming using FID. Numerical experiments on synthetic and real datasets validate our claim.

Index Terms— Generative Adversarial Networks, Fréchet Inception Distance, Gaussian Mixture Models, Probability distribution distance, Earth Mover's distance

1. INTRODUCTION AND RELATED WORK

GANs are powerful generative models that since their introduction in 2014 [1] have found numerous applications in various learning-related tasks. Training is defined as finding a Nash equilibrium of an appropriate game, where both players are neural networks. The generator network aims to be able to produce data that look as close as possible to the training data, effectively capturing the true data distribution. The discriminator network on the other hand aims to be able to discern actual training data from data produced by the generator.

An important factor that makes evaluating and comparing performance of GANs difficult is that they do not define or rely on a likelihood term. A data fit cannot be evaluated numerically in these terms as done for explicit models (e.g. Gaussian Mixture Models [2]). Hence, effort has been put into gauging GAN performance with an objective, quantitative measure. Two proposed measures have proven to be the most popular in the recent literature: Inception Score (IS) [3] and Fréchet Inception Distance (FID) [4]. Both of these metrics measure GAN performance by means of a pre-trained classifier, meant to be used as a deep feature extractor. Semantic features are extracted as activations of this network (the Inception v3 classifier [4], hence the term "Inception" as part of the acronym of both metrics) and subsequently are used to model distributions for true and model data. IS is a metric that numerically attempts to codify variability of class-conditional labels along with variability of sample data as a single score. Out of the two metrics, IS has however been found to behave too poorly for meaningful evaluation [5]. Disadvantages include that optimizing IS was found to lead to adversarial examples, it is sensitive to small network weight changes, and does not constitute a proper distance. FID was found to be more reliable than IS, while pertaining to a number of disadvantages. FID quantifies performance in terms of affinity of the data and model distributions. The two distributions are estimated by fitting Gaussian distributions on the respective Inception feature embeddings of the data. Subsequently, a Fréchet distance measures the divergence between the two distributions, and a lower score means that the synthetic data are closer to the true, original data.

Our contribution is a new metric that follows FID in its rationale, in particular with respect to gauging performance in terms of distance of the true and model distributions. However, in contrast to FID, we drop the assumption that Inception features are Gaussian-distributed, and model the related distributions as GMMs. Thus, multimodal distributions can be better modelled than using simple, unimodal Gaussians. GMMs are estimated using the Expectation-Maximization algorithm (EM), and distances are computed with the Wasserstein (or otherwise known as Earth Mover's Distance, EMD [6, 7]) distance, a metric approriate to measure distances between Finite Mixture Models. With tests on synthetic and real data (MNIST, CIFAR-10, CelebA, BBBC038v1) we show that the proposed metric correlates well with perceived affinity of sets of data. Also, we show that it is robust to hyperparameter choice, as well as it is capable to differentiate between

[1]This research has been partially co-financed by the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call OPEN INNOVATION IN CULTURE, project *Bessarion* (T6YBΠ-00214).

objectively divergent sets when FID cannot.

The remainder of the paper is structured as follows. In section 2 we present briefly the related, previously proposed Fréchet Inception Distance and in the subsequent section 3 we present the proposed Wasserstein Inception Distance. In section 4 we experimentally evaluate and compare the proposed metric, and close with section 5.

2. FRÉCHET INCEPTION DISTANCE

GAN numerical evaluation with FID is performed in terms of computing the distance between two distributions, each of which are represented in practice by two corresponding finite datasets: one dataset corresponds to the "real" data on which the GAN is trained, and the other dataset corresponds to the data the trained GAN produces. Given the two datasets, the FID is defined as the Fréchet distance [8] of the distributions of the Inception features [4] of the two sets. Inception features are defined as a special type of deep features, i.e. activations of a particular intermediate layer of a pretrained neural network. Deep features are known to be effective as powerful semantic descriptions [9, 10]. Inception features are extracted by keeping activations of the last pooling layer of the Inception v3 network pretrained on ImageNet [4]. Hence, each datum is coded as a vector in $\Re^{2048}$. In effect, the data are assumed to be Gaussian-distributed, of which parameter estimation is straightforward: Subsequently, first and second moments for the two distributions are computed: μ_x, Σ_x and μ_g, Σ_g. their parameters correspond to statistics computed over sample embeddings from the data distribution (x) and the model distribution (g). The FID is then computed as the Fréchet distance over the two distributions, defined as:

$$d_{FID}(x,g) = ||\mu_x - \mu_g||^2 + Tr[\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}}] \quad (1)$$

Lower scores correspond to better GAN performance, as lower distance corresponds to higher semantic affinity.

3. PROPOSED MEASURE

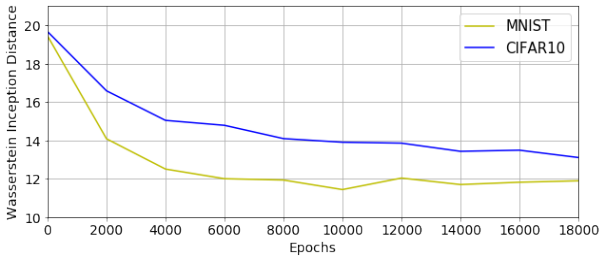


Fig. 1. WInD as a function of training iteration number. As GAN training progresses, data of better quality are produced; this is reflected on WInD values, which progressively attain lower values (hence better quality data).

The FID has been found to correlate well with "human judgement" [4] of affinity between two datasets, however it is characterized by a number of shortcomings [11]. In this work, we follow a rationale similar to the one set by FID to evaluate GAN performance, in the sense that we also cast the problem as one of comparing the distribution of the data versus that of the model. However, we focus in particular on the assumption of the FID that the compared distributions are Gaussian; we believe that such an assumption is overly simplistic, and indeed real data are more often more complex than a simple Gaussian [2], or even any unimodal distribution.

Therefore, we relax the assumption that embeddings from the data and model distributions are Gaussian-distributed. A more powerful way to model any generic distribution is through the use of finite mixture models (FMM) [2]. In contrast to the Gaussian, FMMs are multimodal and can hence model more diverse types of distributions. Perhaps the most often used case of FMM is the Gaussian Mixture Model (GMM) [12, 13]. A K-kernel GMM is multimodal, parameterized probability density function, defined as a weighted sum of K Gaussians. Its parameters are K tuples of means $\mu^1, \mu^2, \cdots, \mu^K$ and covariances $\Sigma^1, \Sigma^2, \cdots, \Sigma^K$, with one tuple corresponding to each Gaussian kernel, and a non-negative scalar controlling each kernel's weight. Further, the K weights $\pi^1, \pi^2, \cdots, \pi^K$ sum up to unity, $\sum_{j=1}^{K} \pi^j = 1$. Formally, a GMM is defined as:

$$p_{GMM}(x) = \sum_{j=1}^{K} \pi^j \mathcal{N}(\mu^j, \Sigma^j) \quad (2)$$

Estimating parameters of a GMM can be performed with the Expectation-Maximization algorithm (EM) [13]. While EM does not provide a global optimum as in the case of estimating Gaussian parameters, it is well-known to converge to a local optimum, and each of its updates are defined in closed form. For the case of application of GMM, the updates always guarantee that the next estimate adheres to sum-to-unity and positive-definite constraints for kernel covariance matrices.

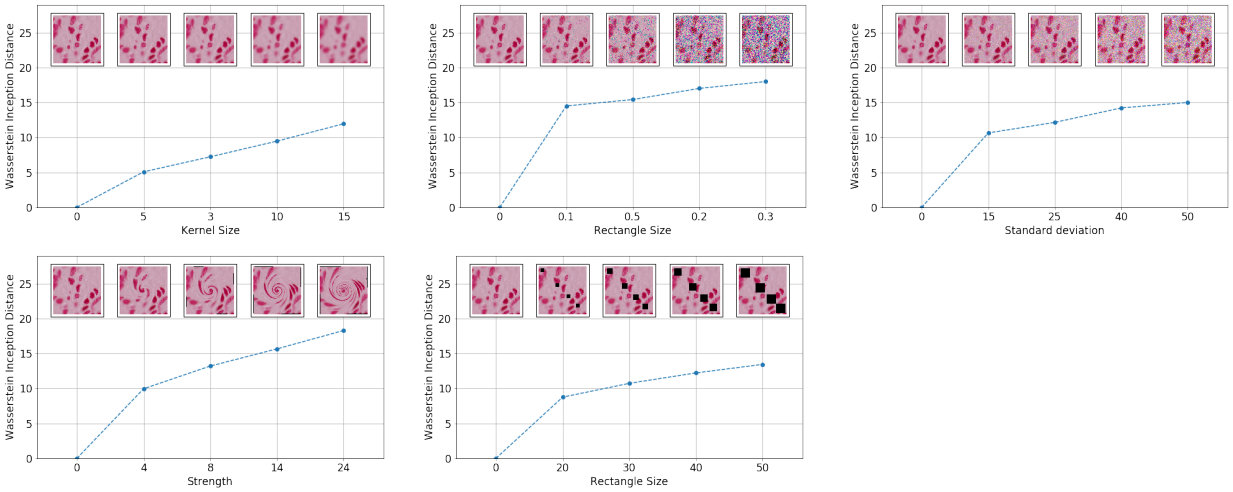
After having computed pdf parameters for the data and model distributions ($\{\pi_x^j, \mu_x^j, \Sigma_x^j\}_{j=1}^K$ and ($\{\pi_g^j, \mu_g^j, \Sigma_g^j\}_{j=1}^K$) respectively), the two pdfs are to be compared w.r.t a distance measure. The Fréchet metric (eq. 1) is unusable in this case, as it is applicable when the pdfs can be completely defined w.r.t. to their first two moments, which is not the case for a FMM. A generalization of the Fréchet metric for GMMs comes in the form of the Wasserstein distance. The Wasserstein distance between two finite mixture models with J and K kernels respectively, is defined as:

$$d_{EMD}(x,g) = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} f^{jk} d^{jk}}{\sum_{j=1}^{J}\sum_{k=1}^{K} f^{jk}}, \quad (3)$$

where f^{jk} denotes the *flow* from kernel j to kernel k. The flow f^{jk} for each kernel pair is a non-negative figure that is

Table 1. Numerical results for tests on the BBBC038v1 cell microscopy imaging dataset. BL: Blur, SP: Salt & Pepper, GN: White Gaussian Noise, SW: Swerl, OC: Occlusion. Proposed WInD distance as a function of comparing with degraded versions of a cell image distribution. WInD increases as the level of degradation increases.

	BL		SP		GN		SW		OC	
Noise intensity	WInD	FID	WInD	FID	WInD	FID	WInD	FID	WInD	FID
1	5.08 ± 0.36	65	9.96 ± 0.03	126	19.64 ± 0.13	129	9.96 ± 0.03	126	8.75 ± 0.51	84
2	7.23 ± 0.02	32	13.2 ± 0.09	208	12.15 ± 0.06	169	13.20 ± 0.09	208	10.73 ± 0.04	126
3	9.47 ± 0.08	115	15.66 ± 0.10	287	14.22 ± 0.09	225	15.66 ± 0.10	287	12.22 ± 0.10	168
4	11.95 ± 0.01	174	18.29 ± 0.01	384	15.00 ± 0.07	252	18.29 ± 0.01	384	13.44 ± 0.03	195



found as the result of a constrained optimization process. In particular, the terms f_{jk} are defined as the optima for eq. 3 given the following constraint:

$$\sum_{j=1}^{J} f^{jk} \leq \pi^j, \sum_{j=1}^{J} f^{jk} \leq \pi^k, \forall j,k$$

$$\sum_{j=1}^{J}\sum_{k=1}^{K} f^{jk} \min\{\sum_{j=1}^{J} \pi^j, \sum_{k=1}^{K} \pi^k\}, \tag{4}$$

which covers also the more general case where the distributions have a non-balanced weight sum. Constraints are linear, hence flow can be computed using a linear programming technique. Metrics d^{jk} are defined as ground distances between kernel j and k individually; we use Fréchet measures as ground distances, since our mixture model is Gaussian they can be readily computed in closed form.

4. NUMERICAL EXPERIMENTS

We have run experiments on various synthetic and real datasets in order to measure quantitatively the suitability of the proposed distance. In a nutshell, with our experiments we validate that the proposed WInD metric is indeed suitable as a measure to gauge GAN performance, it is robust to the choice of metric hyperparameters, and it is overall a more suitable metric than FID. In all cases, k-means is used to initialize

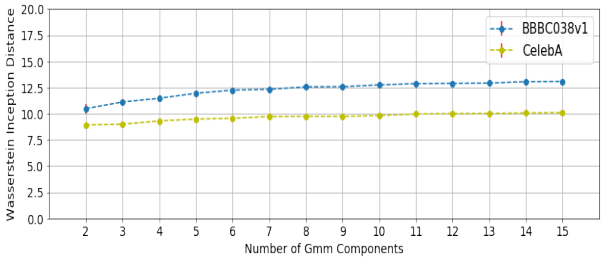


Fig. 2. Effect of choice of K = number of Gaussian kernels on WInD. The true data distribution is compared to the distribution of degraded data. WInD magnitude is relatively stable w.r.t K, and tends towards a constant value as K increases.

the model parameters for the Gaussian Mixtures, and mixture covariance invertibility is guaranteed by adding $\epsilon = 10^{-6} I$ to covariance matrices. Unless otherwise specified, diagonal covariances are used and the number of kernels is set to $K = 5$. Training the GMMs is performed with scikit-learn [14]. As with FID, for all data we use the Inception v3 network and the last pooling layer to produce the inception embeddings, effectively mapping each given image to a vector in $\Re^{2048}$.

Proof-of-concept test, use with GAN: We have run a first set of numerical trials over the $BBBC038v1$ dataset, comprising 729 cell microscopy images [15]. We have degraded images from the above datasets using each of the follow-

WInD = 4.9 ± 0.11 WInD = 10.3 ± 0.05 WInD = 9.2 ± 0.37 WInD = 10.6 ± 0.32

Fig. 3. Comparing dissimilar distributions (each distribution marked with different colour). FID is erroneously zero for all of the above examples.

Table 2. Effect of number of EM runs on statistics of proposed metric. WInD magnitude is relatively stable and practically independent of the specific EM fit.

# of EM runs	WInD
1	13.12
2	13.44 ± 0.03
5	13.30 ± 0.15
10	13.30 ± 0.12
20	13.25 ± 0.16

ing noise schemes, at different levels of noise intensity: a) (Arithmetic-mean) Blur, at convolution windows 3×3, 5×5, 10×10, 15×15. b) Salt and Pepper, at levels 0.01, 0.05, 0.2, 0.3. c) Additive White Gaussian Noise, at $\sigma = 15$, 25, 40, 50. d) Swerl deformation, at strength = 4, 8, 14, 24. e) Occlusion, at occlusion window size = 20, 30, 40, 50. These datasets are meant to simulate GAN outputs where a GAN is trained on the original, un-degraded dataset. We would require from any metric of GAN performance, that degraded outputs score worse (higher distance) than undegraded output (Note that this experiment follows the logic of the analogous experiment presented in [4] to test FID). Results are shown in Table 1, where the proposed WInD is shown to increase consistently as degradation levels increase. We have also used the proposed metric with a real GAN trial on real data (MNIST, CIFAR-10 datasets [4], Fig. 3). As GAN training progresses, more realistic samples are produced. This is reflected on the reported WInD values, which progressively decrease, corresponding to better estimated sample quality.

Effect of hyperparameter choice: The proposed WInD metric, while it assumes a non-Gaussian distribution of inception features, it introduces the number of mixture kernels as a hyperparameter. Furthermore, the resulting fit is dependent on the initial parameters of the EM algorithm [2]. However, in practice we have observed that the aforementioned parameters are largely inconsequential to the metric value. We have run tests on the $BBBC038v1$ cell microscopy dataset, and the $CelebA$ dataset. $CelebA$ comprises approximately $200k$ portrait images in total [16], out of which we have used a random subset of $2,500$ images. In Fig. 3, we show the effect of choosing different values for K when evaluating WInD. The true data distribution is evaluated versus degraded (blurred with a constant kernel of size 15×15) versions of the same set. In both cases, the value for K does not change the WInD value dramatically. Furthermore, as K increases, it tends to stabilize around a fixed value. We must assume that this effect is related to the GMM fit, and the distribution of data; in particular, if some big enough K is chosen that effectively overclusters the data, in terms of a GMM fit and consequently in terms of Wasserstein distance, any newly added kernel will not contribute in practice anything to the fit. Hence, WInD should remain relatively constant. In Fig. 2, for a trial on the $BBBC038v1$ set, WInD values are apparently again relatively stable with respect to the number of different EM initializations/executions employed.

WInD is a better metric than FID: We argue that due to the Gaussianity assumption that underlies FID, datasets that can be more or less clearly different in distribution will be marked as similar or identical with FID, as long as they have the same or similar first moments. As a metric of GAN performance, this translates to a metric that will tag inapropriate / bad quality data samples as realistic. The proposed WInD metric is not characterized by the same problem, because it does not assume Gaussian statistics for the inception feature distributions. An experiment on synthetic data is shown in Fig. 3, where unlike WInD, FID marks the distributions erroneously as similar (with a value of zero). EM covariances are set to full/non-diagonal. Note that EM kernel number is fixed to 5 in all cases, which obviously does not correpond to the true number of clusters. However, WInD is robust to this, and is correctly computed as a non-zero value.

5. CONCLUSION

We have proposed and tested WInD, a novel metric for the evaluation of GANs. The metric combines and extends the rationale of FID and Inception features with a stronger non-Gaussian modeling of sample and model distributions. Our results show that the proposed WInD metric is suitable as a metric of GAN performance, and furthermore that it can be more suitable than FID.

6. REFERENCES

[1] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet, "Are GANs created equal? a large-scale study," in *Advances in neural information processing systems*, 2018, pp. 700–709.

[2] Geoffrey McLachlan and David Peel, *Finite mixture models*, John Wiley & Sons, 2004.

[3] 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*, 2016, pp. 2234–2242.

[4] 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*, 2017, pp. 6626–6637.

[5] Shane Barratt and Rishi Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[6] Hayit Greenspan, Guy Dvir, and Yossi Rubner, "Region correspondence for image matching via EMD flow," in *CVPR 2000 Workshop on Content-Based Access of Image and Video Libraries*, 2000, pp. 27–31.

[7] Thilo Stadelmann and Bernd Freisleben, "Fast and robust speaker clustering using the earth mover's distance and mixmax models," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[8] DC Dowson and BV Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[9] George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, Giorgos Sfikas, and Basilis Gatos, "An alternative deep feature approach to line level keyword spotting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] G.Sfikas, A.Noula, J.Patacas, D.Ioannidis, and D.Tzovaras, "Building thermal output determination using visible spectrum and infrared inputs," in *International Conference on Energy and Sustainable Futures (ICESF)*, September 2019.

[11] Ali Borji, "Pros and cons of GAN evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[12] Giorgos Sfikas, Constantinos Constantinopoulos, Aristidis Likas, and Nikolas P Galatsanos, "An analytic distance metric for gaussian mixture models with application in image retrieval," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 835–840.

[13] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nat Methods*, vol. 9, no. 7, pp. 637, 2012.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.