Estimating the Support Size of GANs for High Energy Physics Detector Simulation

Kristina Jaruskova
Czech Technical University in Prague
Prague, Czech Republic
jaruskri@fjfi.cvut.cz

Sofia Vallecorsa
CERN
Geneva, Switzerland
sofia.vallecorsa@cern.ch

Abstract

Generative Adversarial Networks (GANs) are nowadays able to produce highly realistic output, but a detailed evaluation of GANs results in scientific environment still remains an open problem. An analysis in [1] suggests that the GANs may not have good generalization properties even when the training appears successful. Authors in [2] presented a 3DGAN architecture for scientific simulations of High Energy Physics detectors which reached a high level of precision in terms of physical quantities. This work expands on the idea presented in [3] and uses the principle of the birthday paradox to make an estimate of the support size of the 3DGAN distribution as well as an estimate of the support size of the target distribution represented by the Monte Carlo data. The results suggest that the support size of the 3DGAN is substantially smaller than that of the training data while highlighting, at the same time, the role played by the definition of duplicate events.

1 Introduction

Different types of generative models - variational autoencoders, autoregressive models, generative adversarial networks (GANs) - were proved to be able to learn complex distributions and produce realistic samples. The original paper [4] suggests that the GANs can learn any target distribution if sufficiently large networks, training samples, and computation time are given. However, the theoretical analysis conducted in [1] showed that the training objective approaches its optimum value even though the generated distribution is far from the target distribution, specifically that the generated distribution has a low support size.

In [3], the authors proposed a test for estimating the support size based on the *birthday paradox*. The birthday paradox states that for a discrete uniform distribution with the support size N, a set of approximately $\sqrt{N}$ samples is needed to encounter duplicates with a probability exceeding 50 % [5]. The authors used this idea to estimate the support size of generated distributions for various well established generative models trained on widely used datasets such as CelebA, CIFAR-10 or LSUN [6, 7, 8]. This experiment provided an empirical evidence for their hypothesis presented in [1], the support sizes of the generated distributions were rather low, specifically for the models that produce images of high visual quality. It should be noted however that this procedure depends heavily on the chosen definition of duplicate events.

The level of desired precision of the generated images depends on the application of the generative models. Namely in the field of high energy physics (HEP) experiments, precise simulations of particle transport through matter are requested. The currently used Monte Carlo (MC) techniques possess a high degree of precision with the theoretical predictions but they are both time and resource intensive [9]. The LAGAN [10] and CALOGAN [11] models introduce the idea of using GAN for High Energy Physics shower simulation as two-dimensional images for a simplified calorimeter use case. Since then, there have been other demonstrations employing deep learning for HEP calorimeter

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

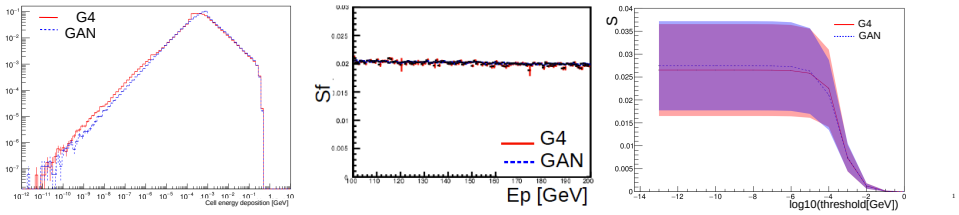


Figure 1: Example distributions validating the quality of the GAN generated images [2]. Left to right: the pixel intensity distribution, the sampling fraction (ratio between the total deposited energy and the input energy) as a function of the input energy and the image sparsity. GAN results are depicted in blue; Monte Carlo in red.

simulation [12, 13, 14]. In [15, 2], the authors explored the possibility of training a three-dimensional convolutional neural network, called the 3DGAN, to simulate the HEP detector response to a particle. The key point of this approach is the fact that this response can be interpreted as a 3D image.

Despite the number of GAN-based HEP applications, a consistent, somewhat standardised, performance validation approach has not yet been defined at the community level. In most cases, the GAN or, more generally, the generative models' output is compared to Monte Carlo data in terms of specific physics-inspired variables. However, an estimation of the generated support space and the identification of well-known problems related to GAN training (such as mode dropping or mode collapse) become even more important, in the context of scientific simulations. By exploring the possibility of implementing the idea of the birthday paradox for measuring the GAN support size, this work is intended as a contribution in this direction and one of the few examples (possibly the first to the best of our knowledge) attempting to give an estimate of the support space learned by a HEP GAN model and to make a direct comparison to Monte Carlo.

Firstly, the 3DGAN is briefly introduced. Then the implementation of the birthday paradox-based test is described, followed by the results of this test. Subsequently, a discussion on the results and possible improvements is presented. The last section contains a reflection on a possible broader impact of this work.

2 The 3D convolutional GAN

The 3DGAN is a 3D convolutional GAN architecture that simulates the calorimeter's energy response. The initial implementation of the 3DGAN [15] admitted conditioning of its input only by the primal energy E_p of the particle entering the detector and it also assumed a 90° angle θ between the direction of the particle and the detector surface. In [2], the 3DGAN is modified to produce images of higher granularity and the training is conditioned using both the incident angle θ, in the $60^\circ - 120^\circ$ range, and the primal energy E_p, in the $2 - 500$ GeV range. Each event in the dataset is represented by a three dimensional $51 \times 51 \times 25$ pixelized image and information about the primal energy E_p and the incident angle θ. The loss function was also modified to include these new features as well as physics-based constraints. Details on the architecture of the 3DGAN generator and discriminator networks can be found in [2].

3DGAN is trained on data simulated using the GEANT4 toolkit [16]. These are also the data that the images generated by the 3DGAN need to be compared to. In [2], the 3DGAN was compared to the GEANT4 data in terms of several physics quantities, such as the sampling fraction (ratio between the total deposited energy and the original particle energy), the image sparsity, as well as for the pixel intensities, which are reproduced in figure 1. As shown in figure 2, good agreement was also found for the energy shower profiles along the x, y, and z axes for different angles.

3 Birthday paradox test implementation for the 3DGAN

The birthday paradox introduced in the first section assumes a discrete uniform distribution over N possible outcomes. However, the GAN distribution is continuous so the occurrence of exact duplicates is not expected. Similarly to the authors in [3], we look for pairs of samples that are

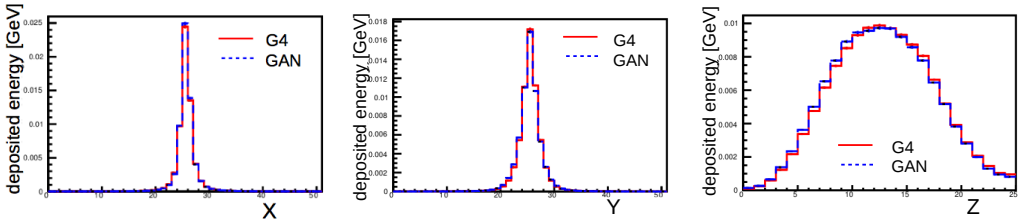


Figure 2: Shower shapes along the axes for a sample with the primal energy $E_p = 100 \pm 2.5$ GeV and the angle $\theta = 90° \pm 2.5°$. (blue) GAN. (red) Monte Carlo.

highly similar in terms of a predefined metric on selected features. Furthermore, the output images of the 3DGAN are conditioned by the primary particle energy E_p and the incident angle θ that vary over a given range. For the purposes of the following analysis, we may focus only on events with the primal energy $E_p = 100 \pm 2.5$ GeV and the angle $\theta = 90° \pm 2.5°$.

To perform the support size estimation using the birthday paradox idea, a definition of a pair of duplicates is needed. This paper includes two possible definitions. The first one is based on high-level features, namely the particle shower shapes along axes and the deposited energy. The second definition adds the Structural Similarity Index (SSIM) [17] as a representative of a pixel-based metric to the high-level features selection.

At first, a pair of duplicate images was defined using the particle shower shapes along the axes x, y, and z, an example is depicted in figure 2: these quantities represent the energy pattern deposited in the calorimeter sensors along the three directions. The difference between the shower shapes of two samples was measured by the *Jensen-Shannon divergence*

$$D_{JS}\left(P, Q\right) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right) \tag{1}$$

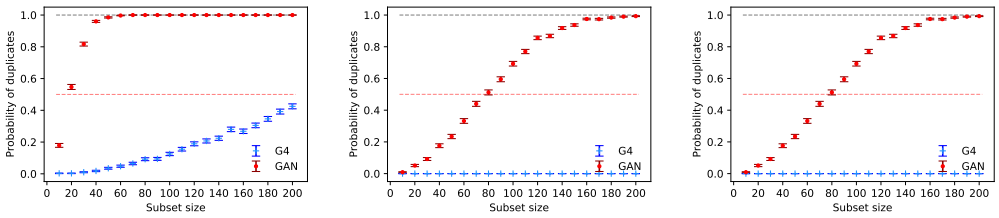
where $D_{KL}\left(P, Q\right) = P \cdot \log\left(\frac{P}{Q}\right) - P + Q$ refers to the unnormalized Kullback-Leibler divergence.

In the next step, a subset of 5000 samples from the GEANT4 dataset was taken and the divergences between all samples in this subset were calculated for energy distribution along all three axes. Subsequently, the α-quantile of the computed divergences was found for each direction. Finally, two samples are marked as duplicates if the divergences of the energy distributions along axes are below the α-quantile values for all three axes.

To generate an estimate of the support size, the divergences between s randomly selected samples are calculated and it is examined if there is at least one pair of duplicates in this subset. The randomized selection of s samples was repeated 1000 times for both GEANT4 and GAN data to obtain the probability of encountering duplicates in a subset of size s. Figure 3a depicts the probabilities for the GEANT4 and the GAN generated data with $\alpha = 0.02$. The probability of encountering duplicates exceeds 0.5 for a subset of 20 samples generated by the GAN which indicates that the support size is ≈ 400. To reach the 0.5 probability for the GEANT4 data, more than 200 samples are needed.

The divergences between the shower shapes reflect mainly the spatial information about the energy deposition. An additional restriction of the absolute difference between the total energy deposited in the calorimeter sensor was included. The threshold for this measure of dissimilarity between two samples was determined using the GEANT4 data in the same manner as the thresholds for the divergences between the shower shapes. The probabilities of finding a duplicate in a subset of size s are depicted in figure 3b. The additional criterion of the total deposited energy significantly changed the results. The number of the GAN generated data needed to achieve the 0.5 probability increased to 80 which yields the number $\approx 6\,400$ as an estimate of the support size of the GAN. For the GEANT4 data, the probabilities are not approaching the value of 0.5 for the examined subset sizes.

The previous approach used only high-level physical features to assess the similarity of two samples. Both the choice of features and the threshold have major influence on the estimated support size. However, the samples have teh form of a 3D image, so we introduce a pixel-based metric, the Structural Similarity Index (SSIM). The SSIM ranges from 0 to 1 and satisfies $\mathrm{SSIM}\left(x, y\right) = 1$



(a) Threshold definition based on shower shapes.

(b) Threshold definition based on shower shapes and the total deposited energy.

(c) Threshold definition based on shower shapes, total deposited energy and SSIM values.

Figure 3: Probabilities of encountering duplicates for sets of different sizes (denoted as subset size). The first subset size for which the probability of 0.5 (red dashed line) is exceeded gives the estimate of the support size.

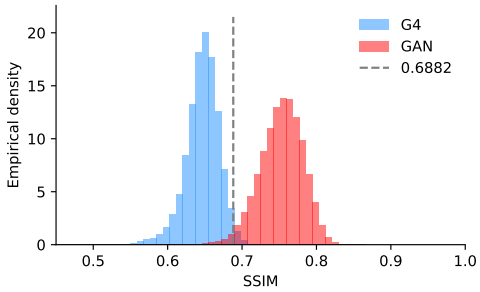


Figure 4: Histogram of the SSIM values for pairs of images from the GEANT4 dataset and the GAN generated dataset. The dashed line marks the 0.98-quantile computed on the GEANT4 data.

if and only if $x = y$. Based on the results presented in [2], the value of the SSIM parameter L that represents the dynamic pixel range was set to $L = 10^{-4}$. Figure 4 shows the SSIM values calculated for random pairs of images in the GEANT4 and GAN datasets. The dashed line, representing the 0.98-quantile computed on the GEANT4 data, shows how most of the GAN pairs have an SSIM above the threshold, thus they appear more similar than the GEANT4 pairs in terms of the SSIM. Nevertheless, we included the SSIM-based selection in the definition of duplicate events. Results are shown in figure 3c: since most of the GAN pairs have SSIM above the SSIM-based threshold, the effect of the SSIM criterion on GAN data is negligible, while slightly reducing the corresponding probability for GEANT4 samples.

4 Results

Although preliminary, these results clearly show that the 3DGAN produces images that are significantly more similar in terms of the shower shapes and the total deposited energy than the images produced by the MC simulation (GEANT4). However, they also suggest that the estimate of the subset size depends strongly on the way the duplicates are defined.

Figure 5 shows the two-dimensional energy projections of the two closest duplicates in the GAN sample, selected according to the 0.02-quantile thresholds. The two events clearly exhibit differences, in particular in the distribution of small energy deposits scattered at the edges, and it is likely that the two events would not be flagged as duplicates by a visual inspection. This fact suggests that the definition of duplicate events is a choice to be made on a use-case basis, depending on the simulated sample specific requirements (in terms of detector resolution, for example).

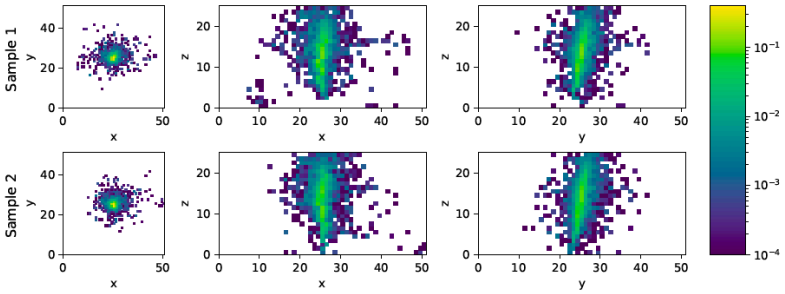


Figure 5: 2D energy shower projections of the 2 closest duplicates, selected according to the 0.02-quantile thresholds.

5 Conclusion and broader impact

The definition of common validation strategies, agreed at the community level, represents an important step toward bringing generative model prototypes to production-level quality and speeding their integration in the experiments simulation frameworks. An estimation of the generated support space and the identification of well-known problems related to GAN training (such as mode dropping or mode collapse) are even more important in the context of studies investigating the possibility to replace Monte Carlo simulation with generative models for High Energy Physics applications.

In this work,we presented the results of a study estimating the size of the generated support space of 3DGAN, a model trained to reproduce high-granularity calorimeters output. We have applied the birthday paradox test, exploring different practical implementations of the duplicate-events definition, based on high-level variables (such as energy deposition profiles) and pixel-wise quantities more commonly used for image analysis (such as the SSIM). Our findings highlight a significant difference between the 3DGAN and the target distribution support sizes. We are continuing our investigations in order to better characterise this support size difference and understand how it could affect the practical use of GAN generated samples in experimental applications. At the same time, we want to draw attention to the strongest limitation of this approach, namely choosing the duplicate-events definition, which has a large impact on the test result and, once more, proves the challenges related to validating generative models output.

References

[1] S. Arora and Y. Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs). *ArXiv e-prints*, 2017.

[2] G. R. Khattak, S. Vallecorsa, F. Carminati, and G. M. Khan. Particle detector simulation using generative adversarial networks with domain related constraints. *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 28-33, 2019.

[3] S. Arora and Y. Zhang. Do GANs actually learn the distribution? An empirical study. *ArXiv e-prints*, 2017.

[4] I. J. Goodfellow et al. Generative Adversarial Networks. *ArXiv e-prints*, 2014.

[5] Brink, D. A (probably) exact solution to the Birthday Problem. *Ramanujan J* **28**, 223–238, 2012.

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

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Citeseer*, 2009.

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

[9] The HEP Software Foundation. A Roadmap for HEP Software and Computing R&D for the 2020s, *Computing and Software for Big Science*, 2019.10.1007/s41781-018-0018-8.

[10] de Oliveira, L. and Paganini, M. and Nachman, B. Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis, 2017.

[11] Paganini, Michela and de Oliveira, Luke and Nachman, Benjamin. CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev. D*, 97, 1,014021, 2018. 10.1103/PhysRevD.97.014021.

[12] Salamani, D. and others. Deep Generative Models for Fast Shower Simulation in ATLAS. *Proceedings, 14th International Conference on e-Science*. 2019. 10.1109/eScience.2018.00091.

[13] Di Sipio, R. and Giannelli, M. F. and Haghighat, S. K. and Palazzo, S. DijetGAN: A Generative-Adversarial Network Approach for the Simulation of QCD Dijet Events at the LHC. 2020. 10.1007/JHEP08(2019)110.

[14] Chekalina, Viktoria and Orlova, Elena and Ratnikov, Fedor and Ulyanov, Dmitry and Ustyuzhanin, Andrey and Zakharov, Egor. Generative Models for Fast Calorimeter Simulation: the LHCb case. 2019. 10.1051/epjconf/201921402034.

[15] G. R. Khattak, S. Vallecorsa, and F. Carminati. Three dimensional energy parametrized generative adversarial networks for electromagnetic shower simulation. *25th IEEE International Conference on Image Processing (ICIP)*, 3913–3917, 2018.

[16] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, vol. A506, 250–303, 2003.

[17] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.