Estimating the Support Size of GANs for High Energy Physics

Kristina Jaruskova[1], Sofia Vallecorsa[2]

[1]Czech Technical University in Prague, jaruskri@fjfi.cvut.cz, [2]CERN, sofia.vallecorsa@cern.ch

Can the birthday paradox idea help to estimate support size of GANs in High Energy Physics applications?

Simulations of particle transport through matter are fundamental to High Energy Physics (HEP) research. GANs offer a fast alternative to standard Monte Carlo (MC) methods but a unified performance validation approach has not yet been defined at the community level.

This work expands on the idea presented in [1]. We adapted the birthday paradox test to 3DGAN [2], a model simulating calorimeters output, to estimate its support size. We explored different duplicate-events definitions proving that this step is also the strongest limitation of the method.

Birthday paradox

P(*2 people in a room born the same day*) > 0.5

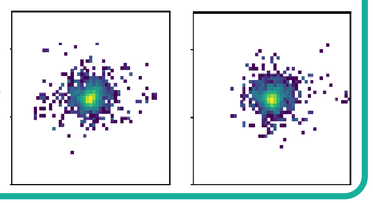
How many people need to be in the room to satisfy the inequality?

- A year with N **days** $\rightarrow \sqrt{N}$ **people** is needed
- $N \approx$ support size of a discrete uniform distr.

How many GAN samples are needed to find any duplicates with probability of 0.5?

- **(The answer)$^2 \approx$** estimate of support size
- Images as samples from a multivariate continuous distribution

GAN duplicates = images “similar enough”



3DGAN

Convolutional GAN architecture [2]

- Simulates 3D output (51x51x25) of high granularity EM calorimeter
- Output interpreted as an image
  - Energy measurement as pixel intensity
- Remarkable agreement to Monte Carlo

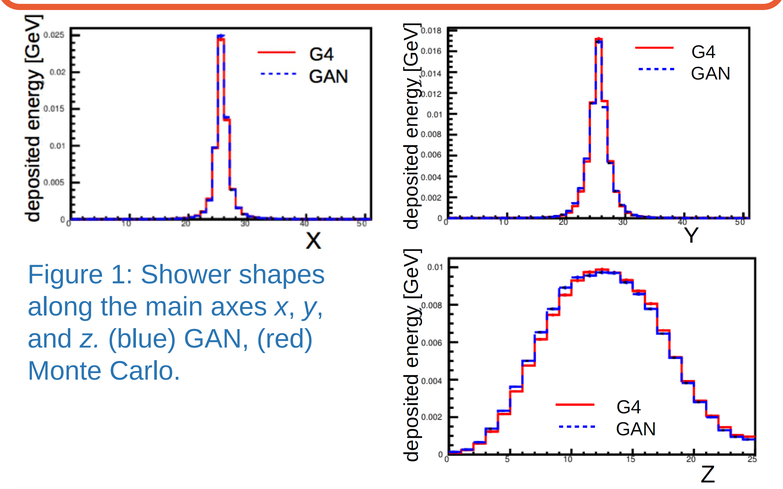


Figure 1: Shower shapes along the main axes *x*, *y*, and *z*. (blue) GAN, (red) Monte Carlo.

GAN duplicates

Metrics of similarity

- Jensen-Shannon divergence between shower shapes (Figure 1)
- Difference in total deposited energy
- SSIM (Structural Similarity Index) [3]

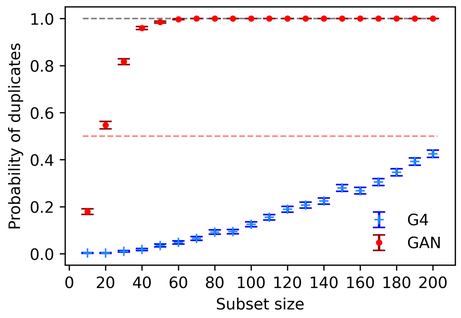
Definition of duplicates

- Compute the selected metrics on training data.
- Define threshold as 0.02-quantile (or 0.98-q.).
- Combine the threshold condition for all metrics.

Estimates of support size

Duplicates defined by the similarity in shower shapes:

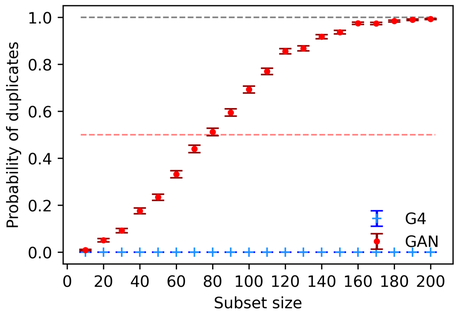
- GAN: 400
- MC data: over 4000



a. Shower shapes

Duplicates in terms of the similarity in shower shapes, dep. energy and SSIM:

- GAN: 6 400
- MC data: not approaching probability of 0.5



b. Shower shapes, dep. energy, SSIM

Figure 2: Probabilities of encountering duplicates for different duplicates definitions. (G4) Monte Carlo dataset.

Conclusions

- **3DGAN has significantly smaller subset size.**
- **Limitation: Estimates of the support size depend strongly on the definition of duplicate events which is subjective and varies with the use case.**
- **Need to understand the effect of the way GAN data is used in the experimental workflow.**

[1] S. Arora, Y. Zhang (2017)
[2] G. R. Khattak et al. (2019)
[3] Zhou Wang et al. (2004)

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