Sparse Data Generation for Particle-Based Simulation of Hadronic Jets in the LHC

Breno Orzari [1] **Thiago R. F. P. Tomei** [1] **Maurizio Pierini** [2] **Mary Touranakou** [2 3] **Javier Duarte** [4]
Raghav Kansal [4] **Jean-Roch Vlimant** [5] **Dimitrios Gunopulos** [3]

Abstract

We develop a generative neural network for the generation of sparse data in particle physics using a permutation-invariant and physics-informed loss function. The input dataset used in this study consists of the particle constituents of hadronic jets due to its sparsity and the possibility of evaluating the network's ability to accurately describe the particles and jets properties. A variational autoencoder composed of convolutional layers in the encoder and decoder is used as the generator. The loss function consists of a reconstruction error term and the Kullback-Leibler divergence between the output of the encoder and the latent vector variables. The permutation-invariant loss on the particles' properties is combined with two mean-squared error terms that measure the difference between input and output jets mass and transverse momentum, which improves the network's generation capability as it imposes physics constraints, allowing the model to learn the kinematics of the jets.

1. Introduction

The Large Hadron Collider (LHC) at CERN (Evans & Bryant, 2008) is a proton-proton (pp) collider that allows fundamental physics research at the highest energy regimes Fast simulation of high energy physics (HEP) objects for particle physics data analysis has been a challenge for the last decades; the problem became exacerbated in the LHC environment. With the advent of machine learning (ML) techniques being applied to event reconstruction, jets classification, and other necessities of experimental particle physics (Guest et al., 2018; Albertsson et al., 2018), the usage of ML to build a generator of events can be seen as the next step for the current Monte Carlo based generators. Although it is a challenging task, it can be split into the generation of the distinct final-state objects of particle collisions. For the high energy pp collisions that take place at the LHC, the most common kinds of particles produced are *hadrons* that consist of quarks and gluons, such as the proton and neutron. Hadrons produced in these collisions usually appear in collimated groups, called *hadronic jets* (Marzani et al., 2019; Kogler et al., 2019).

We measure the jet physical properties by setting sensible detector elements around the pp collision region. Those elements allow the measurement of the energy and momenta (for charged particles) of the jet constituents. Those combined measurements are used to determine the jet characteristics through the particle-flow (PF) algorithm (Sirunyan et al., 2017; Aaboud et al., 2017). The jet can then be described as a sparse, unordered set of constituent particles where each particle is further characterized by its properties like energy, momentum[2], charge, particle type, among others. To deal with the sparsity of the data, one might use graph neural networks (Duarte & Vlimant, 2020; Shlomi et al., 2021) without assuming any specific ordering. For the same reason, energy flow networks (Komiske et al., 2019), physics-inspired permutation-invariant architectures, were introduced.

In general, the loss function used to train such models is the mean squared error (MSE). However, using such a loss term implies breaking the permutation invariance of the data. To avoid such behavior, in this paper we use a loss function based on a nearest-neighbor distance (NND) or Chamfer distance (Barrow et al., 1977; Fan et al., 2017).The aim of this work is to develop a generative neural network capable of producing sets of non-ordered particles as constituents of hadronic jets which correctly reproduce the par-

[1]SPRACE-Unesp, São Paulo, Brazil [2]European Organization for Nuclear Research (CERN) [3]National and Kapodistrian University of Athens, Athens, Greece [4]University of California San Diego, La Jolla, California, United States of America [5]California Institute of Technology, Pasadena, California, United States of America. Correspondence to: Breno Orzari <breno.orzari@cern.ch>.

Proceedings of the LatinX in AI (LXAI) Research workshop at ICML 2021.

[2]The coordinate system used in HEP experiments is as follows: origin set at the center of the local pp collision region; z axis along the beam direction, y axis vertically upward. 3D coordinates are usually given in terms of $\rho = (x^2 + y^2)^{1/2}$, azimuth ϕ and pseudorapidity $\eta = -\ln\tan(\theta/2)$, where θ is the polar angle.

ticles and jets' physical properties with high fidelity, using a permutation-invariant reconstruction error term in the loss function that, for this purpose, has not been applied so far.

This paper is organized as follows: we introduce the benchmark dataset and model in Section 2. The loss function and evaluation metric used are described in Section 3. We show performance on applications in Section 4. Conclusions and future steps are given in Section 5.

2. Benchmark Dataset and Model

The input dataset[3] consists of lists of 100 particles that are constituents of hadronic jets. Where less than 100 particles are present in the jet, the list is filled with zero-padded entries for the remaining particles up to 100. Each particle is described by its momentum components as $\vec{p} = (p_x, p_y, p_z)$ in units of giga-electron volts (GeV). Approximately 177,000 examples of high-momentum jets originating from gluons[4] were generated, and this data is split 70% for training, 15% for validation and 15% for testing. A feature-dependent normalization is applied in the data such that the range of each particle feature is [0.0, 1.0].

A variational autoencoder (VAE) is used as the generative neural network model, composed of 3 convolutional layers and 2 dense layers in the encoder, a latent vector of dimension 20 as well as 2 dense layers and 3 convolutional layers in the decoder. After each layer the rectified linear unit (ReLU) activation function is applied, except after the last encoder layer, that is used for the reparameterization, and the last decoder layer, in which the sigmoid activation function is used to constrain the output of the network to be between [0.0, 1.0]. A schematic representation of the architecture used is depicted in figure 1.

3. Error Function and Evaluation Metric

In the training step, the main purpose of the VAE is to reconstruct the input data as closely as possible, while encoding the information in the latent vector for the generation of new data. The product of this step is referred to as the output dataset, which is composed of sets of reconstructed particles described by $\vec{\hat{p}} = (\hat{p}_x, \hat{p}_y, \hat{p}_z)$. To achieve this goal, the loss function can be written as a reconstruction error between input and output data, and a Kullback-Leibler (KL) divergence (D_{KL}) term, which forces the probability distribution of the latent vector values to take the form of a simple distribution, such as a standard Gaussian one. The loss function is then written as

$$L_{\mathrm{VAE}} = (1-\beta)L_{\mathrm{rec}} + \beta D_{\mathrm{KL}}, \tag{1}$$

where the β parameter (Higgins et al., 2017) is a relative weighting factor between L_{rec} and D_{KL}.

The reconstruction term is further composed by 3 distinct contributions. The first is an NND term that quantifies the difference between input and output particles properties in a permutation-invariant way, following the expression

$$D^{\mathrm{NND}}(\mathcal{J}_k, \hat{\mathcal{J}}_k) = \sum_{i\in\mathcal{J}_k} \min_{j\in\hat{\mathcal{J}}_k} D(\vec{p}_i, \vec{\hat{p}}_j) + \sum_{j\in\hat{\mathcal{J}}_k} \min_{i\in\mathcal{J}_k} D(\vec{p}_i, \vec{\hat{p}}_j), \tag{2}$$

where D is the squared Euclidean distance, $\mathcal{J}$ and $\hat{\mathcal{J}}$ represent the *kth* input and output jets, respectively, and the indices i and j represent the *ith* and *jth* particles inside a given jet. Summing over all the jets in the dataset, we have

$$L^{\mathrm{NND}} = \sum_k D^{\mathrm{NND}}(\mathcal{J}_k, \hat{\mathcal{J}}_k). \tag{3}$$

The other two terms measure the difference between input and output jets' p_{T} and invariant mass, thus forcing the network to learn the jets characteristics as well. They are written as

$$L^{\mathrm{J}} = \sum_k \left[\gamma_{p_{\mathrm{T}}} \mathrm{MSE}(p_{\mathrm{T}k}, \hat{p}_{\mathrm{T}k}) + \gamma_m \mathrm{MSE}(m_k, \hat{m}_k)\right], \tag{4}$$

where $p_{\mathrm{T}k}$ ($\hat{p}_{\mathrm{T}k}$) and m_k ($\hat{m}_k$) are the *kth* input (output) jet p_{T} and mass, respectively, and $\gamma_{p_{\mathrm{T}}}$ and γ_m are weights that can be optimized. Here it would be important to highlight that, since the dataset is composed of the particles features only, the calculation of the jets characteristics from the particles properties needs to be performed for every input/output jet. However, to properly compute these quantities, the normalization applied in the dataset has to be inverted, because the sum of the particles' momenta in absolute GeV scale are necessary.

At the generation step, a set of random values sampled from a Gaussian distribution is fed into the VAE latent vector, and the network output will be referred to as the generated dataset. The earth mover's (Wasserstein) distance (EMD) (Rubner et al., 2000) has been also proposed as a permutation-invariant metric to compare unordered points in sets (Fan et al., 2017), however its use directly in the loss function has been shown to be computationally costly. Hence, we use EMD to measure the generating capabilities of the network by calculating the distance between the 1D probability distributions of input and generated jets mass, p_{T}, energy, η and ϕ. The sum of these 5 quantities, referred to as $\mathrm{EMD}_{\mathrm{sum}}$, is used to measure the network's performance.

[3] 100 particles dataset

[4] The dataset also contains simulated jets that originate from W bosons, Z bosons, top quarks (t), and light quarks (q), that were not used for this work, but will be used for future research.

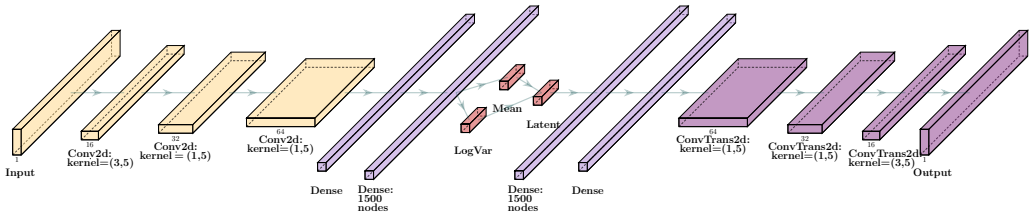

Figure 1. Schematic representation of the VAE architecture used in this study. For each convolutional (dense) layer the kernel size (number of neurons) is specified.

4. Applications

The architecture is implemented in PyTorch, and trained using the early stopping technique (Raskutti et al., 2014; Matet et al., 2017) at first, with a patience of 15 epochs, and at a second stage it was trained for 1,500 epochs (approximately 4–6 times the number of epochs as before) to test the effect of a longer training period. The optimizer used is Adam (Kingma & Ba, 2015) with a learning rate of 0.0001. The parameter β was set to 0.9998, due to the large values of L_{rec} compared to D_{KL}, but there is room for further optimization. The hyperparameter $\gamma_{p_{\text{T}}}$ was set to 1.0 and remained unchanged, while γ_m was set as 1.0 at first, and was increased to 10.0 later, since, it showed improvements in the generative capabilities of the network.

Every 50 epochs, each of the trained models was set to generate around 26,000 jets for the evaluation of the generating properties of the network. The EMD metric was calculated and the best model for each training section was selected as the one that showed the smallest EMD_{sum}. Table 1 shows the results of four distinct models trained as described above: using early stopping with γ_m = 1.0 (ES1) and γ_m = 10.0 (ES10); training for more epochs with γ_m = 1.0 (ME1) and γ_m = 10.0 (ME10).

Table 1. Best EMD_{sum} values for distinct models: early stopping with γ_m = 1.0 (ES1) and γ_m = 10.0 (ES10); training for more epochs with γ_m = 1.0 (ME1) and γ_m = 10.0 (ME10)

MODEL	EMD_{sum}
ES1	0.0119
ME1	0.0090
ES10	0.0085
ME10	**0.0062**

Based on the EMD_{sum}, we observe a large improvement in the VAE generation of hadronic jets when increasing both the γ_m parameter and the number of epochs. The reason for the former is that the largest contribution to the evaluation metric comes from the distinction between input and generated jets mass, and, making the MSE on the jets mass term more important in the error, resulted in a better generation of the jets mass.

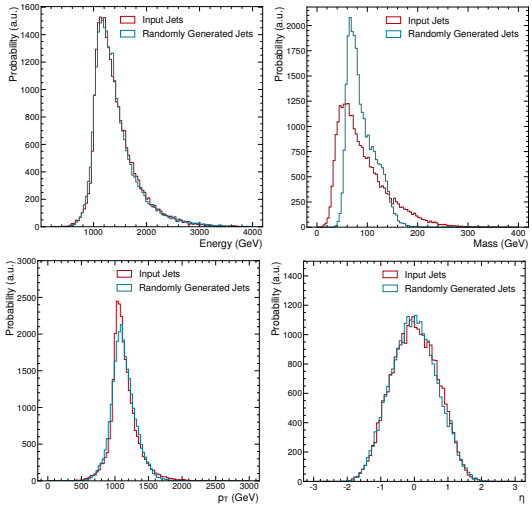

Figure 2. Input (red) and generated (blue) jets variables for the best performing model (ME10). Top: energy (left), mass (right). Bottom: p_{T} (left), η (right).

Figure 2 shows the distributions of some input and generated jets variables for the best model performance (ME10). Although the input mass distribution is not reproduced by the generated jets, all of the other generated jet variables have similar input and generated distributions.

5. Summary and Outlook

We presented a generative neural network suited for the generation of sparse data, showing an application of a VAE for the generation of hadronic jets, just like the ones that are collected as data at the LHC. Distinct techniques for the training step of the VAE were applied, and the best model showed a value of the EMD_{sum} metric of 0.0062. Although the comparison of input and generated jets mass histograms shows a difference in its distributions, other relevant variables showed good performance. There is still a lot of room for improvement through the optimization of the network hyperparameters. In order to improve the agreement and obtain a more accurate jet generator, we plan

to study the addition of normalizing flows (Kobyzev et al., 2020) in the latent space to learn a better posterior that what we obtain from Gaussian sampling.

Acknowledgements

B. O. and T. T. are supported by grant #2018/25225-9, SãoPaulo Research Foundation (FAPESP). B. O. was also partially supported by grants #2018/01398-1 and #2019/16401-0, São Paulo Research Foundation (FAPESP). M. T. and M. P. are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 772369). J-R. V. is supported by the U.S. Department of Energy (DOE), Office of Science, Office of High Energy Physics under Award No. DE-SC0011925, DE-SC0019227, and DE-AC02-07CH11359. J. D. and R. K. are supported by the DOE, Office of Science, Office of High Energy Physics Early Career Research program under Award No. DE-SC0021187 and by the DOE, Office of Advanced Scientific Computing Research under Award No. DE-SC0021396 (FAIR4HEP).

References

Aaboud, M. et al. Jet reconstruction and performance using particle flow with the ATLAS Detector. *Eur. Phys. J. C*, 77(7):466, 2017. doi: 10.1140/epjc/s10052-017-5031-2.

Albertsson, K. et al. Machine Learning in High Energy Physics Community White Paper. *J. Phys. Conf. Ser.*, 1085(2):022008, 2018. doi: 10.1088/1742-6596/1085/2/022008.

Barrow, H. G. et al. Parametric correspondence and Chamfer matching: Two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (KJCAI)*, volume 2, pp. 659, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc. URL https://www.ijcai.org/Proceedings/77-2/Papers/024.pdf.

Duarte, J. and Vlimant, J.-R. Graph neural networks for particle tracking and reconstruction. In Calafiura, P., Rousseau, D., and Terao, K. (eds.), *Artificial Intelligence for High Energy Physics*. World Scientific Publishing, 2020. doi: 10.1142/12200. Submitted to *Int. J. Mod. Phys. A*.

Evans, Lyndon, e. and Bryant, Philip, e. LHC Machine. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001.

Fan, H., Su, H., and Guibas, L. J. A point set generation network for 3D object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463, 6 2017. doi: 10.1109/CVPR.2017.264.

Guest, D., Cranmer, K., and Whiteson, D. Deep Learning and its Application to LHC Physics. *Ann. Rev. Nucl. Part. Sci.*, 68:161, 2018. doi: 10.1146/annurev-nucl-101917-021019.

Higgins, I. et al. β-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, 2015.

Kobyzev, I., Prince, S., and Brubaker, M. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1, 2020. doi: 10.1109/tpami.2020.2992934.

Kogler, R. et al. Jet Substructure at the Large Hadron Collider: Experimental Review. *Rev. Mod. Phys.*, 91(4):045003, 2019. doi: 10.1103/RevModPhys.91.045003.

Komiske, P. T., Metodiev, E. M., and Thaler, J. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP*, 01:121, 2019. doi: 10.1007/JHEP01(2019)121.

Marzani, S., Soyez, G., and Spannowsky, M. *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, volume 958. Springer, 2019. doi: 10.1007/978-3-030-15709-8.

Matet, S. et al. Don't relax: early stopping for convex regularization, 2017.

Raskutti, G., Wainwright, M. J., and Yu, B. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15:335, 2014. URL http://jmlr.org/papers/v15/raskutti14a.html.

Rubner, Y., Tomasi, C., and Guibas, L. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40:99, 11 2000. doi: 10.1023/A:1026543900054.

Shlomi, J., Battaglia, P., and Vlimant, J.-R. Graph Neural Networks in Particle Physics. *Mach. Learn.: Sci. Technol.*, 2:021001, 7 2021. doi: 10.1088/2632-2153/abbf9a.

Sirunyan, A. M. et al. Particle-flow reconstruction and global event description with the CMS detector. *J. Instrum.*, 12(10):P10003, 2017. doi: 10.1088/1748-0221/12/10/P10003.