

Accelerating HEP research: A Deep Learning framework

Francesco Vaselli

Pisa, 22/02/2022

Introduction

The need for fast and reliable computing methods in the physical sciences, notably in the high-energy field, has fueled much of the technological progress of modern times. Despite this, the average physicist has generally little time to spend on innovating and improving its computing toolkit, and has to content himself with boilerplate solutions. This is especially true in highly specialized fields, such as HEP, which pursue a wide variety of research directions.

Specifically, in recent years, *machine learning* techniques have been massively adopted by scientific collaborations around the world. In particular, the paradigm known as *deep learning*, which leverages multiple layers of *artificial neurons* (theorized by [10]) trained through the use of a *loss function* and *backpropagation*, has achieved a wide range of successful usages. However, such tools remain geared towards the necessities of industry; much work remains to be done to enable the use of this technologies in hard sciences.

As a physicist with a keen interest in this type of applications, while still retaining useful domain knowledge, I am convinced that I could achieve significant results by pursuing the following research directions.

Research topics

What follows is a series of deep learning tools addressing the necessities of Run 3/ High Luminosity LHC: the HL-LHC will produce more than 250 inverse femtobarns of data per year and will be capable of collecting up to 4000 inverse femtobarns, making it possible to set new constraints on a whole set of processes regarding the Higgs boson, such as its coupling to fermions of the second generation or the Higgs self-coupling. The vast amount of data will require efficient and fast *simulation techniques*, improvements in both *tracking* and *vertex reconstruction*, as well as reliable *trigger* and *real time* analysis frameworks.

Flash sim

This is the topic that I am currently pursuing with my Master thesis. The key idea is to directly generate a high level analysis format, such as CMS NANO AOD [9], training on fully simulated events. As a benchmark to evaluate the performance of such a simulation, the search data for the decay of Higgs to muon pairs in the VBF channel has been chosen. This kind of analysis requires only a limited number of muons and jets features to be simulated while still depending upon proper handling of correlations, so it is a good benchmark for a first prototype of this deep learning based approach. The goal of this kind of simulation, that we call *flash sim*, is to generate the full detector response (simulation and reconstruction) in a negligible time compared to a full simulation, hence enabling the generation of future large datasets at low computing cost.

Generative Adversarial Networks (GANs) [3] and *Variational Autoencoders* (VAEs) [6] are deep generative models which showed remarkable results in the field of computer vision, and have already been extensively investigated by the collaboration at CERN (see [1] and [7]); despite this, there is still a limited literature regarding behavior in low dimensionality as in our case, e.g. [5]. These models are also prone to *mode collapse* and failure to converge, and necessitate careful and domain-informed tuning to result in successful HEP applications.

A novel and promising approach is the one offered by Normalizing Flows [8], a family of methods for constructing flexible learnable probability distributions, often with neural networks, which allow us to surpass the limitations of simple parametric forms to represent complex high-dimensional distributions. In this case, a simple multivariate source of noise, for example a standard i.i.d. normal distribution, $X \sim \mathcal{N}(\mathbf{0}, I_{D \times D})$, is passed through a vector-valued invertible bijection, $g : \mathbb{R}^D \rightarrow \mathbb{R}^D$, to produce the more complex transformed variable $Y = g(X)$, and we can compose such bijective transformations to produce even more complex distributions, i.e. $Y = (g_{(0)} \circ g_{(1)} \circ \dots \circ g_{(L-1)})(X)$.

The main challenge is in designing parametrizable multivariate bijections that have closed form expressions for both g and g^{-1} , a tractable Jacobian whose calculation scales with $O(D)$ rather than $O(D^3)$, and can express a flexible class of functions. Recent advancements have demonstrated the suitability of *spline transforms* (see [2]), and my Master thesis showed promising results in the generation of samples which provide convincing distributions as well as correct correlations and conditioning on the simulation ground truth.

Still, lots of potential application and improvements of this architecture for both generation and inference (see [4]) could be investigated.

Deep vertex and tracking

Trigger and real time analysis

References

- [1] Anja Butter, Tilman Plehn, and Ramon Winterhalder. How to gan lhc events. *SciPost Physics*, 7(6), Dec 2019.

- [2] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows, 2019.
- [3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
- [4] Stephen R. Green and Jonathan Gair. Complete parameter inference for gw150914 using deep learning, 2020.
- [5] Felix Jimenez, Amanda Koepke, Mary Gregg, and Michael Frey. Generative adversarial network performance in low-dimensional settings, 2021-04-20 04:04:00 2021.
- [6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.
- [7] Sydney Otten, Sascha Caron, Wieske de Swart, Melissa van Beekveld, Luc Hendriks, Caspar van Leeuwen, Damian Podareanu, Roberto Ruiz de Austri, and Rob Verheyen. Event generation and statistical sampling for physics with deep generative models and a density information buffer, 2021.
- [8] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.
- [9] Andrea Rizzi, Giovanni Petrucciani, and Marco Peruzzi. A further reduction in CMS event data for analysis: the NANOAOB format. In *European Physical Journal Web of Conferences*, volume 214 of *European Physical Journal Web of Conferences*, page 06021, July 2019.
- [10] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.