A frequentist study of the false signals generated in the training of semi-supervised neural network classifiers using a Wasserstein Generative Adversarial Network as a data generator.

Benjamin Lieberman[1], Salah-Eddine Dahbi[1], Xifeng Ruan[1] and Finn Stevenson[1], Bruce Mellado[1,2]

[1]School of Physics and Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa
[2]iThemba LABS, National Research Foundation, PO Box 722, Somerset West 7129, South Africa

E-mail: benjamin.lieberman@cern.ch

Abstract. In resonance searches for new physics, machine learning techniques are used to classify signal from background events. When using machine learning classifiers it is necessary to measure the amount of background events being incorrectly labelled as signal events. In this research the $Z\gamma \rightarrow (\ell + \ell-)\gamma$ final state dataset focusing around 150 GeV centre of mass is used. A Wasserstein Generative Adversarial Network, WGAN, is used as a generative model and a semi-supervised Deep Neural Network, DNN, is used as a classifier. This study provides a methodology and the results of the measurement of false signals generated during the training of semi-supervised DNN classifiers.

1. Introduction

The Standard Model, SM, of particle physics was completed by the 2012 discovery of the Higgs boson, by the ATLAS and CMS collaboration [1, 2]. Since this discovery, developments in machine learning together with increasing luminosity at the Large Hadron Collider, LHC, has enabled searches beyond the SM, BSM. Searches for new bosons BSM is motivated by phenomena such as the matter-anti-matter asymmetry, Dark Matter, and the origin of the neutrino mass, which cannot be explained by the SM. In order to gain insight into these phenomena, machine learning classifiers are used to extract signal from background processes. The semi-supervised machine learning technique is able to reduce training biases by training models using a partially labeled dataset. In this analysis the machine learning method of the semi-supervised classification study, presented in Ref. [3], is investigated. This paper therefore proposes, implements and evaluates a methodology of scrutinising the success of semi-supervised machine learning classification using a Deep Neural network, DNN. This is achieved by quantifying the amount of error, in the form of fake signal events, caused by over-fitting during the training of semi-supervised models that confront side-bands and the signal regions.

When analysing local and global resonances, a extremely large dataset is necessary to overcome the "look elsewhere effect" [4]. This effect can be defined, in searches for resonances

within a given mass range, as the probability of observing a significant local excess of events, elsewhere within the range. To account for this phenomena, machine learning based data generators can be used in conjunction with traditional Monte Carlo (MC) generation to scale datasets. Generative models such as Generative Adversarial Networks (GAN) and Variational Auto-Encoders (VAE) are excellent examples of methodologies commonly used in industry to scale data [5]. Once trained generative models are able to generate events with excellent accuracy at scale with minimal computational resources. While a full evaluation of the different data generators is being conducted in a parallel study, the Wasserstein Generative Adversarial Network, WGAN, is used in this analysis.

In this study the Run 1 LHC data features and a 2HDM+S model, where S is a singlet scalar, is used Ref. [6, 7]. In this model the heavy scalar, H, decays predominantly into SS, Sh, where h is the SM Higgs boson. A possible singlet, S, candidate is reported in Ref. [8]. The model exposes multi-lepton anomalies, verified in Refs. [9, 10, 11, 12], astro-physics anomalies, when complemented by a Dark Matter candidate [13], as well as various other anomalies including the $g-2$ muon experiment reported by Fermilab [14, 15, 16]. A full review of anomalies can be found in Ref. [17].

1.1. $Z\gamma$ Monte Carlo Dataset

In this study, the simulated $Z\gamma$ background, is considered. $Z\gamma$ contributes to 90% of the total backgrounds in the production of the Higgs like heavy scalar decaying to $Z\gamma$ ($pp \to H \to Z\gamma$) events, where $Z \to e^+e^-$ or $Z \to \mu^+\mu^-$. The $Z\gamma$ SM MC samples used in this analysis have been generated using `Madgraph5` [18] and the detector level simulation is performed using `Delphes`(v3) [19]. The focus of the analysis is around the centre of mass of 150GeV (122GeV$< m_{\ell\ell\gamma} <$ 178GeV). The kinematic features used in the study are $Z\gamma$ invariant mass, $m_{\ell\ell\gamma}$, missing transverse energy, E_T^{miss}, missing transverse energy azimuthal angle, $\Phi_{E_T^{miss}}$, $\Delta R_{\ell\ell}$ ($\Delta R \equiv \sqrt{(\Delta\eta_{ll})^2 + (\Delta\phi_{ll})^2}$), the number of jets, N_j, number of central jets, N_{cj}, and the transverse momentum, $Pt_{\ell 1|\ell 2|\gamma}$, azimuthal angle, $\Phi_{\ell 1|\ell 2|\gamma}$, and pseudo-rapidity, $\eta_{\ell 1|\ell 2|\gamma}$, for each of the leptons and the photon respectively. The feature distributions are shown in Figure 2.

1.2. Wasserstein Generative Adversarial Network with gradient penalty

The GAN training strategy is an interaction between two competing neural networks. The generator, G, model maps a source of noise to the input space. A discriminator, model receives either a generated sample or a MC data sample and must distinguish between the two. The generator is trained to output data of sufficient quality to fool the discriminator into believing it is MC data. The discriminator is simultaneously trained to distinguish MC from generated data. An improved methodology of the GAN, described in Ref. [20], is the Wasserstein GAN, WGAN, which adopts the Wasserstein distance, $W(q,p)$, defined as the minimum cost of transporting mass in order to transform the distribution q into the distribution p. The discriminator is replaced in the improved model with a critic, C, and the gradients are controlled using a gradient penalty, GP, which penalizes the norm of the critic gradients with respect to the input. An overview of the WGAN with gradient penalty is shown in Figure 1. The generator loss, L_g, function is defined as

$$L_g = \min \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[C(\tilde{x})], \tag{1}$$

where $\tilde{x} = G(z)$ and z is the latent space noise and $\mathbb{P}_g$ is the generator model distribution, $\tilde{x}$. The critic loss, L_c, with gradient penalty is defined as

$$L_c = \max \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[C(x)] - \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[C(\tilde{x})] + \lambda * GP, \tag{2}$$

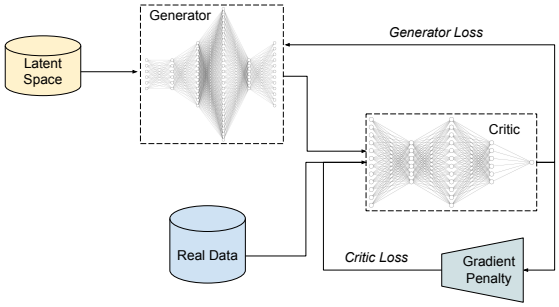

Figure 1. Systematic diagram of WGAN with gradient penalty

where λ is the gradient penalty coefficient, $\mathbb{P}_r$ is the MC data distribution, and the gradient penalty, GP, is defined as

$$GP = \underset{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}{\mathbb{E}}[(||\nabla_{\tilde{x}}C(\tilde{x})||_2 - 1)^2] \tag{3}$$

2. Methodology

The methodology used in this analysis, to quantify fake signal generated in the training of DNN classifiers, can be broken down into three components. These are the WGAN data generator, the semi-supervised DNN and the background rejection scan. In order to conduct a $3\,\sigma$ frequentest analysis of fake signal generated, the semi-supervised DNN model must be trained and evaluated on independent datasets more than $5{\cdot}10^4$ times. For each run, iteration of training and evaluating the DNN, the WGAN is therefore used to generate a statistically distinct dataset.

2.1. Data Generation using a WGAN with gradient penalty

The machine learning based data generator used in this analysis, is the WGAN with gradient penalty, which is implemented in Python using pytorch [21]. The model is trained and optimised in order to reproduce the $Z\gamma$ MC events and accurately as possible. To this end the quality of the generated events must be evaluated in terms of the feature distributions as well as the event-wise feature correlations. In order for the generated events to successfully mimic high energy physics events, both the feature distributions and correlations must match the MC data.

The optimisation of the model is therefore achieved by minimising the difference in feature distributions and correlation between the generated data and MC training data. The final model architectures used for the WGAN are summarised in Table 1. The final hyper-parameters used are latent dimension of 18 (equal to number of features), learning rate of $6 \cdot 10^{-5}$, batch size of 512 and gradient penalty coefficient, λ, of 0.001. The Critic is trained five times for each Generator training iteration in order to optimised the WGAN training.

2.2. Semi-Supervised DNN Training

The DNN model used, is a replica of the optimised model used in Ref. [3]. The model is trained on all of the features except the invariant mass, $m\ell\ell\gamma$, which is used to define the signal and background regions. The dataset is divided into two training samples, namely mass-window or signal region, (144GeV to 156GeV), and side-band or background region, (132GeV to 144GeV and 156GeV to 168GeV). As both samples comprise of pure $Z\gamma$ background, we do not expect the DNN response to find any separation between samples.

Table 1. Critic and generator final model architectures.

Model	Layer(s)	Number of nodes	Activation function
Critic	Input Layer	18 (Number of features)	ReLu
	Hidden Layers	[256, 512, 256]	ReLu
	Output Layer	1	Linear
Generator	Input Layer	18 (Latent Space)	BatchNorm, ReLu
	Hidden Layers	[256, 512, 1024, 512, 256]	BatchNorm, ReLu
	Output Layer	18 (Number of features)	BatchNorm, ReLu

2.3. Quantification of Fake Signal using background rejection scan

For each of the generated datasets that the DNN is trained on, a DNN response, in range $(0, 1)$, is produced. In order to evaluate local and global fake signals, a background rejection scan of the response distribution is used.

The background rejection scan is implemented by extracting batches of events, from the response distribution, to be analysed. The batches extracted make up 50, 60, 70, 80 and 90% of the total events. Each batch is taken starting from the maxima, 1, of the response distribution and moving towards the minima, 0. Once a batch of data is extracted, the events are mapped to their corresponding invariant mass. Each batch's invariant mass distribution is fit with an exponential function, $f(x)$, which exposes the distribution of background events. A second fit, using the exponential function with an added Gaussian, $g(x)$, is applied with the Gaussian centred at the center of mass, 150 GeV, and σ equal to the resolution of the dataset, 2.4. The Gaussian therefore is able to represent any signal events found within the mass window. As there are no signal events within the analysis dataset, any signals found can be assumed to be generated within the training of the DNN. The significance of signal found for each batch can be calculated using Equation 4.

$$\sigma = \frac{\int_a^b [g(x) - f(x)]\, dx}{\sqrt{\int_a^b [f(x)]\, dx}} \tag{4}$$

where a and b are the minima, 132 GeV, and maxima, 168 GeV, of the invariant mass respectively.

3. Results and discussion

The WGAN with gradient penalty converged after 1200 epochs. The feature distributions and corresponding correlations, generated by the model, are visually compared to that of the MC data in Figures 2 and 3 respectively. A visual analysis clearly exposes that all the generated feature distributions describe the MC data well except the $\Phi_{\ell 1|\ell 2|E_T^{miss}}$ which need further improvement. The Kolmogorov-Smirnov score and bin-wise relative difference scores are used to measure the difference between generated and MC feature distributions. The final model is found to generate data with an average Kolmogorov-Smirnov score of 0.074 and average bin-wise relative difference of 0.021. The event-wise feature correlation of the generated data is shown to excellently mimic that of the MC data using the Spearmann correlation and mean correlation difference. The final model achieves a Spearmann correlation score of 0.825 and mean correlation difference of 0.023. As the feature distributions and correlations of the generated data are of sufficient quality, the pre-trained model can be used in the frequentest study.

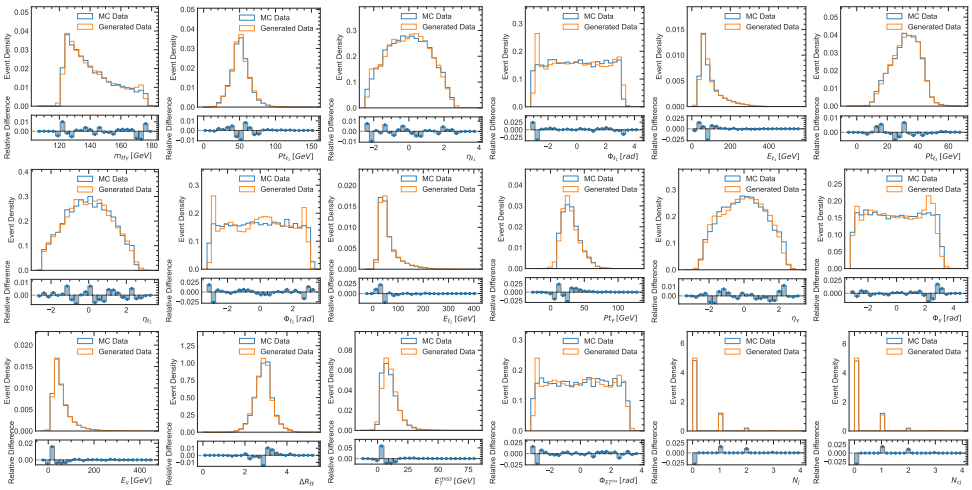

Figure 2. Final feature distribution comparison of MC data and WGAN generated data.

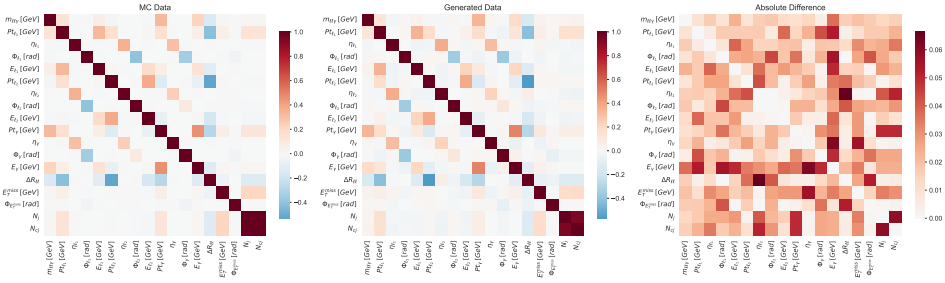

Figure 3. Feature correlation heatmaps. Left: MC training data. Centre: WGAN generated data. Right: Difference between MC and generated data.

A single run of the frequentest study consists of; WGAN generated dataset, training the DNN on the given dataset, mapping the DNN response to invariant mass distribution and applying fits to quantify significance of fake signal for each background rejection. To initially evaluate the extent of false signal, the study is repeated 500 times. The result of these runs is shown in Figure 4. It is important to note that measured significance reflect not only the over-training of the DNN but also the success of the WGAN in synthesising realistic events.

4. Conclusions

The analysis has shown that, using the above methodology, the extent of fake signal generated in the training of machine learning classifiers can be quantified. The generated dataset show clear evidence that a the WGAN can be used to synthesise high energy physics events. The significance distributions produced in Figure 4, demonstrates that for 500 runs the significance measured is reasonable, maximum 2σ. In order to complete this analysis the study needs to be repeated for more than 50000 runs. This will allow the significance to be integrated over to quantify the extent of the influence of local and global excesses.

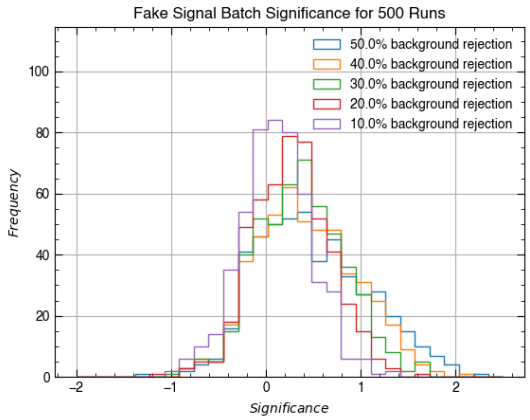

Figure 4. Signal significance measured for DNN response event batches on 500 runs of the frequentest study.

References

[1] Aad G *et al.* (ATLAS) 2012 *Phys. Lett. B* **716** 1–29 (*Preprint* 1207.7214)

[2] Chatrchyan S *et al.* (CMS) 2012 *Phys. Lett. B* **716** 30–61 (*Preprint* 1207.7235)

[3] Dahbi S E, Choma J, Mokgatitswane G, Ruan X, Lieberman B, Mellado B and Celik T 2021 *International Journal of Modern Physics A* ISSN 1793-656X URL http://dx.doi.org/10.1142/S0217751X21502419

[4] Gross E and Vitells O 2010 *The European Physical Journal C* **70** 525–530 URL

[5] Otten S, Caron S, de Swart W, van Beekveld M, Hendriks L, van Leeuwen C, Podareanu D, de Austri R R and Verheyen R 2019 Event generation and statistical sampling for physics with deep generative models and a density information buffer URL https://arxiv.org/abs/1901.00875

[6] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B and Reed R G 2015 (*Preprint* 1506.00612)

[7] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B, Reed R G and Ruan X 2016 *Eur. Phys. J. C* **76** 580 (*Preprint* 1606.01674)

[8] Crivellin A, Fang Y, Fischer O, Kumar A, Kumar M, Malwa E, Mellado B, Rapheeha N, Ruan X and Sha Q 2021 (*Preprint* 2109.02650)

[9] von Buddenbrock S, Cornell A S, Fadol A, Kumar M, Mellado B and Ruan X 2018 *J. Phys. G* **45** 115003 (*Preprint* 1711.07874)

[10] Buddenbrock S, Cornell A S, Fang Y, Fadol Mohammed A, Kumar M, Mellado B and Tomiwa K G 2019 *JHEP* **10** 157 (*Preprint* 1901.05300)

[11] von Buddenbrock S, Ruiz R and Mellado B 2020 *Phys. Lett. B* **811** 135964 (*Preprint* 2009.00032)

[12] Hernandez Y, Kumar M, Cornell A S, Dahbi S E, Fang Y, Lieberman B, Mellado B, Monnakgotla K, Ruan X and Xin S 2021 *Eur. Phys. J. C* **81** 365 (*Preprint* 1912.00699)

[13] Beck G, Kumar M, Malwa E, Mellado B and Temo R 2021 (*Preprint* 2102.10596)

[14] Sabatta D, Cornell A S, Goyal A, Kumar M, Mellado B and Ruan X 2020 *Chin. Phys. C* **44** 063103 (*Preprint* 1909.03969)

[15] Abi B *et al.* (Muon g-2) 2021 *Phys. Rev. Lett.* **126** 141801 (*Preprint* 2104.03281)

[16] Aoyama T *et al.* 2020 *Phys. Rept.* **887** 1–166 (*Preprint* 2006.04822)

[17] Fischer O *et al.* 2021 *Unveiling hidden Physics Beyond the Standard Model at the LHC* (*Preprint* 2109.06065)

[18] Alwall J, Frederix R, Frixione S, Hirschi V, Maltoni F, Mattelaer O, Shao H S, Stelzer T, Torrielli P and Zaro M 2014 *Journal of High Energy Physics* **2014** URL

[19] de Favereau J, Delaere C, Demin P, Giammanco A, Lemaître V, Mertens A and Selvaggi M (DELPHES 3) 2014 *JHEP* **02** 057 (*Preprint* 1307.6346)

[20] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A 2017 Improved training of wasserstein gans URL https://arxiv.org/abs/1704.00028

[21] Paszke A and Gross S 2019 *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc.) pp 8024–8035