Evaluation and Optimisation of a Generative-Classification Hybrid Variational Autoencoder in the Search for Resonances at the LHC

Finn Stevenson[1], **Benjamin Lieberman**[1], **Abhaya Swain**[1], **Bruce Mellado**[12]

[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: finn.david.stevenson@cern.ch

Abstract. This proceedings depicts the optimisation and evaluation of a Variational Auto-encoder plus Discriminator model used for $Z\gamma$ background final state event generation. This work has been completed to evaluate the use of deep learning models instead of traditional more computationally expensive physics event production and classification mechanisms.

1 Introduction

This proceedings documents the secondary stages, evaluation and optimisation, of the development of a deep neural network generative model for production of $Z\gamma$ final state data for physics analysis. Following the discovery of the Higgs boson in 2012 [1, 2], which can be said to have completed the Standard Model (SM) of particle physics, there are still a number of discrepancies to the SM in the form of unexplained phenomena and anomalies in the data that prompt searches for new bosons. The work described in this paper fits into the bigger picture of searching for new bosons as this work can specifically effects the efficiency and accuracy of the search within the $Z\gamma$ final state [3, 4]. A main limiting factor related to completing the aforementioned search is the requirement of a large quantity of $Z\gamma$ final state background events. Traditionally in similar searches, the requirement of copious amounts of data is satisfied through the use of computationally expensive Monte Carlo (MC) production mechanisms. Using pre-trained deep learning models to produce data instead can have alleviate some of the finite CPU hours used in the analysis as well as speed up the time taken. Over and above the use of deep learning models in data generation, such models can also be used in the search for signal classification purposes. Variational Auto-encoders (VAEs) are a type of deep learning model that can somewhat uniquely be used as both a data generation and signal classification model, using the same trained model. This bi-functionality again adds to the efficiency and time improvement. The aim of this work is to evaluate the use of deep generative models, specifically variational auto-encoders and derivatives for both data generation and signal classification tasks

in the search for new bosons. This proceedings specifically concentrates on the evaluation and optimsation of the VAE based generative capacity of the model, further work will be completed to evaluate the classification capability.

1.1 MC $Z\gamma$ Final State Data

In this work, simulated $Z\gamma$ background data has been used, which 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 $Z\gamma$ sample MC events have been generated using `Madgraph5` [5] with the `NNPDF3.0` parton distribution functions [6]. Here, the Standard Model (SM) of particle physics has been utilized for which the UFO model files required by the Madgraph are from `FeynRules` [7]. The parton level generation is followed by the parton showering and hadronization by `Pythia` [8] and then the detector level simulation is performed using `Delphes`(v3) [9]. The jets at this level has been constructed using Fastjet [10] which involves the anti-K_T jet algorithm with $P_T > 20$ GeV and radius $R = 0.5$. While generating the sample we decayed the $Z\gamma$ boson to leptons. Some baseline cuts have also been applied on the leptons and photons at the Madgraph level to enhance the statistics.

1.2 Centre of Mass and Kinematic Features

The analysis focuses around the centre of mass of 150GeV (132GeV$< m_{\ell\ell\gamma} <$ 168GeV). The kinematic features used in the study are $Z\gamma$ invariant mass, $m_{\ell\ell\gamma}$; the transverse momentum, azimuthal angle, pseudo-rapidity and energy of the leading lepton, sub-leading lepton and photon respectively, $Pt_{\ell_1\ell_2\gamma}$, $\Phi_{\ell_1\ell_2\gamma}$, $\eta_{\ell_1\ell_2\gamma}$ and $E_{\ell_1\ell_2\gamma}$; missing transverse energy E_T^{miss} and it's azimuthal angle, $\Phi_{E_T^{miss}}$; the number of jets, N_j, the number of central jets, N_{cj}; and $\Delta R_{\ell\ell}$ ($\Delta R \equiv \sqrt{(\Delta\eta_{ll})^2 + (\Delta\phi_{ll})^2}$), $Pt_{\ell\ell}/m_{\ell\ell\gamma}$, $\Delta\Phi_{\ell\ell}$ and $\Delta\Phi(E_T^{miss}, Z\gamma)$.

2 Hypothesis

A well trained Variational Auto-encoder can aid in the search for new bosons in the $Z\gamma$ final state for both data generation and signal classification purposes. The work presented in this proceedings ha a more specifc hypothesis as follows: The addition of a discriminator network to the overall VAE model as well as a notion of adversarial training similar to that of a Generative Adversarial Network can aid in the training of the model and produce better generated events.

3 Methodology

Previous proceedings works have described the initial development of the base VAE model and therefore the base model will only be briefly described here. Concentration will be centred around the addition of the discriminator network to the VAE overall model that aids in the training of the overall model to produce better generated events in terms of a number of selected metrics.

3.1 Variational Auto-encoder

A VAE is an autoencoder (AE) with architectural changes and an additional component added to the loss function that facilitates regularised training and imroves the generative capability of the model by ensuring appropriate latent space properties. As shown in Figure 1, the VAE architecture is composed of two main composite networks, the encoder, and the decoder. The VAE is trained to minimise the loss between the input data (kinematic variable event) and the encoded-decoded output (reconstructed event). However, in the case of the VAE, instead of encoding an input event as a single vector, the input is encoded as a distribution over the latent space of the VAE. This allows for some regularisation of the latent space.

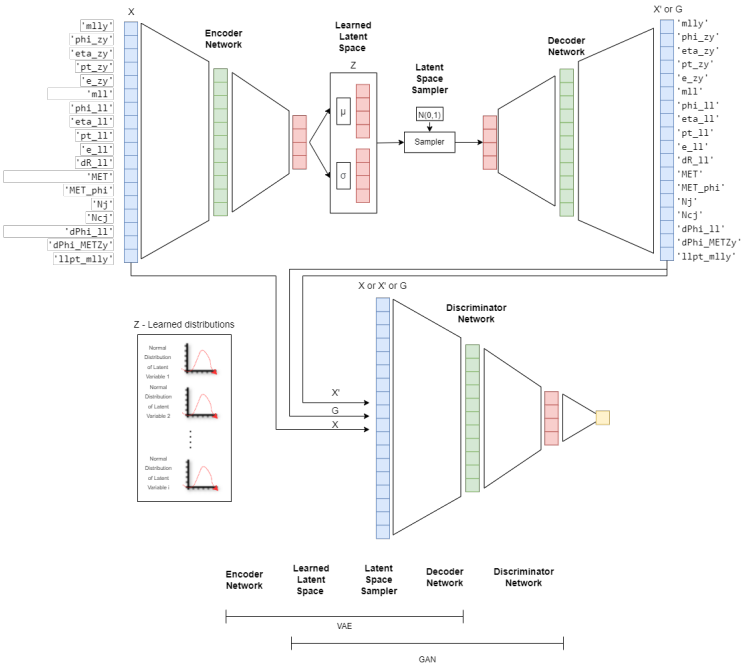

Figure 1. Diagram of VAE Base and greater VAE+D model Architecture, showing encoder network, decoder network, learned latent space and discriminator network.

These latent space distributions are forced to be normal Gaussian so that the encoder can be configured to return latent space vectors that represent the mean μ and the covariance σ of the normal Gaussian distributions because the Kullback-Leibler divergence between two the Gaussian distributions has a form that can be directly expressed in terms of the means and the covariance matrices of the two distributions. As a result of encoding an input event to a distribution rather than a single vector, it is possible to regularise the latent space. The loss function that is minimised when training a VAE is composed of a reconstruction loss component that is responsible for forcing the output of the decoder to be as close to the input, and secondly, a regularisation loss component, that serves to regularise the organisation of the latent space by making the distributions returned by the encoder close to a standard normal distribution. The loss function also contains a coefficient for the KL-Divergence loss term, β_V. This can be used during optimisation to weigh the importance of the KL-Divergence loss term against that of the reconstruction loss term.

$$L_{VAE} = L_R + \beta_V * L_{KL} \tag{1}$$

$$L_R = \overline{((X' - X)^2)} \tag{2}$$

Where X is the input event and X' is the reconstructed event.

$$L_{KL} = \sum (\sigma^2 + \mu^2 - \log \sigma - 1) \tag{3}$$

3.2 Variational Auto-encoder + Discriminator (VAE+D)

The addition of a discriminator and the notion of adverserial training network helps in the training of the VAE encoder-decoder network. Figure 1 shows the architecture diagram of the VAE+D model. The VAE+D loss functions see below are slightly different to the VAE, whilst

still including the main loss components of the original VAE. Unlike the VAE, the VAE+D has a loss function for each individual network, the encoder, decoder and discriminator and each network's weights are updated individually at different times during a forward pass of the overall VAE+D. Similar to a GAN, the discriminator and the VAE are trained simultaneously, with the discriminator learning to differentiate fake events from real events and the VAE learning to reproduce real events accurately. The VAE+D loss functions and components are as follows:

$$BCE_{real} = criterion_{BCE}(X) vs. 1s \tag{4}$$

$$BCE_{recon} = criterion_{BCE}(X') vs. 0s \tag{5}$$

$$BCE_{gen} = criterion_{BCE}(G) vs. 0s \tag{6}$$

$$L_{disc} = BCE_{real} + BCE_{recon} + BCE_{gen} \tag{7}$$

$$L_{dec} = \gamma * Loss_R - Loss_{disc} \tag{8}$$

$$L_{enc} = Loss_{KL} + \gamma * Loss_R \tag{9}$$

Where $criterion_{BCE}$ is the binary cross entropy between either the actual data against actual data, reconstructed data or generated data. The results is then compared to a 1s label (for the real data) or a 0s label (for reconstructed and generated) to obtain the final discriminator loss function, L_{disc}. γ is a similar variable to the variational beta in the standard VAE, however, instead is a coefficient of the reconstruction loss instead of the KL divergence like in the VAE.

4 Model Optimisation

VAEs have many hyper-parameters that can be optimised in order to achieve the best model. This hyper-parameter optimisation can be done using a variety of methodologies and available libraries, however in this work a manual optimisation loop was created. The optimisation loop was created to loop through each of the parameters shown in Table 1, on each iteration building, training and evaluating a model with the loop iteration parameters. There have also been some other considerations taken into account in the code involving looping through the architectural based parameters because of the fact some architectural parameters are constrained by others ones to achieve a plausible architecture.

Table 1. Table showing VAE Selected Hyper-parameters and value options

Hyper-parameter Optimisation			
Hyper-parameter	Model	Value Options	Brief Description
Learning Rate	VAE, VAE+D	[0.01, 0.001, 0.0001]	Standard machine learning hyper-parameter that determines how drastically the model changes it's weights each iteration in attempt to minimise the loss function and achieve convergence.
Batch Size	VAE, VAE+D	[1, 16, 64, 256, 512]	The batch size refers to the number of training examples used in one training iteration of the model.
Latent Dimension Size	VAE, VAE+D	[8, 16, 32, 64]	Number of latent dimensions variables.
Number of Hidden Layers	VAE, VAE+D	[1, 2, 3]	Number of hidden layers in between input and latent layers.
Number of Nodes in the Hidden Layers	VAE, VAE+D	[16, 32, 64, 128, 256, 512]	Number of nodes in the hidden layers
Variational Beta	VAE, VAE+D	[1, 10, 100, 500, 1000, 5000]	Coefficient of the KL-Divergence loss term in loss function.

After running of the aforementioned hyper-parameter optimisation loop, the best parameters were found for both the VAE and the VAE+D models.

5 Results

The Results of the addition of the discriminator network to the VAE can be seen in the figures below. It can be seen that the optimised VAE+D model is able to generate more realistic events.

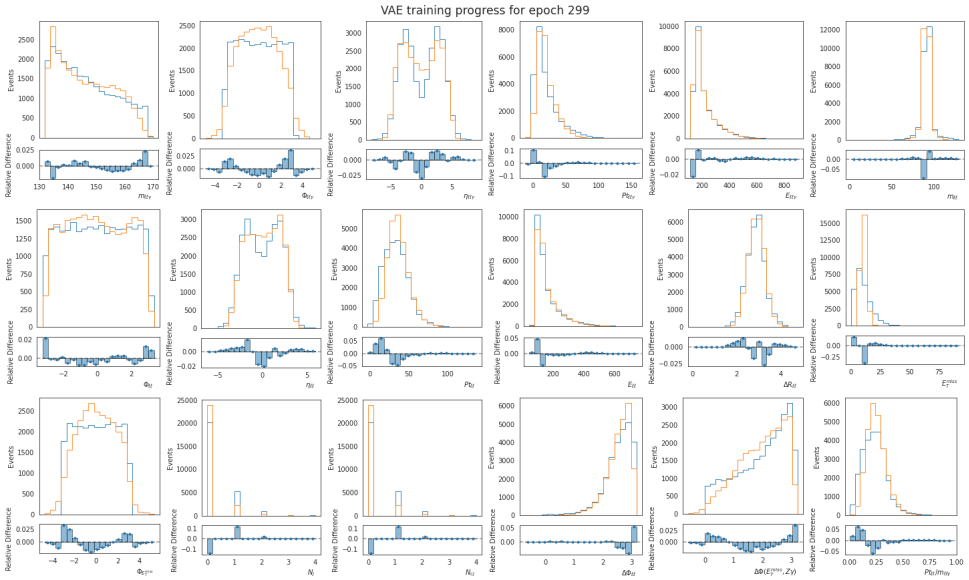

Figure 2. Distribution Graph of Generated Event Features vs. MC Data for VAE Model.

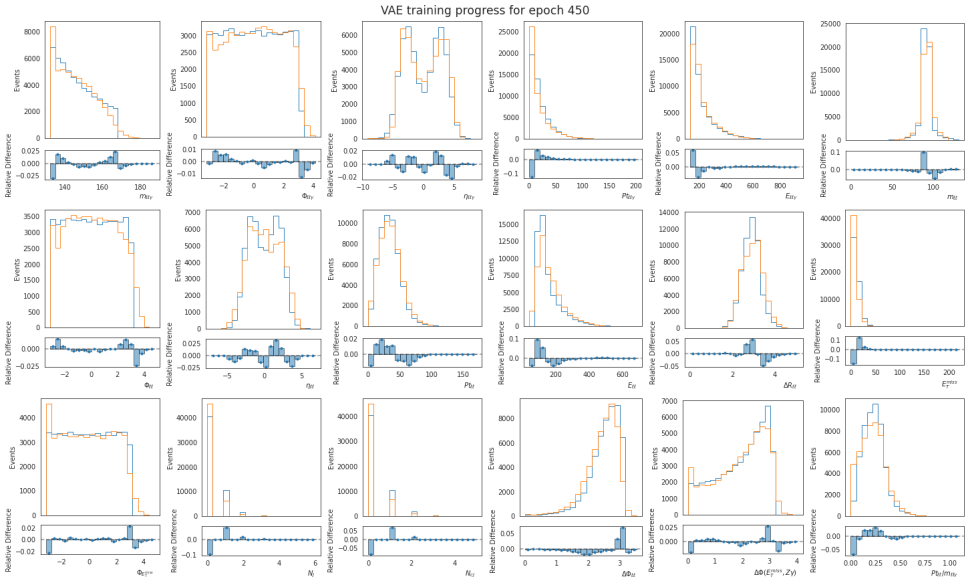

Figure 3. Distribution Graph of Generated Event Features vs. MC Data for VAE+D Model.

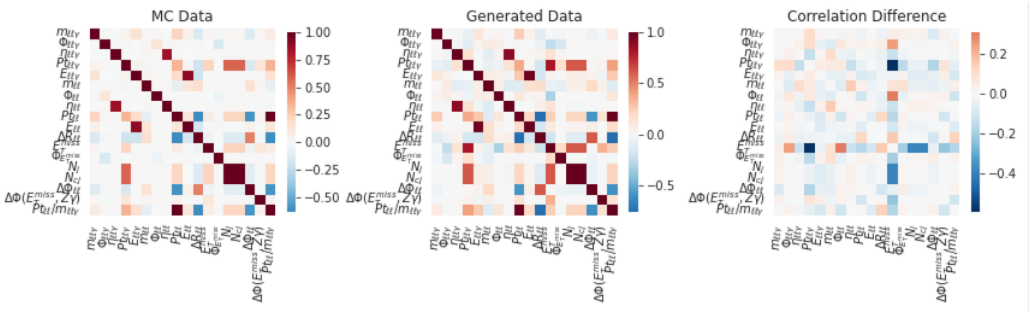

Figure 4. Correlation Comparison of Generated Event Features vs. MC Data for VAE Model.

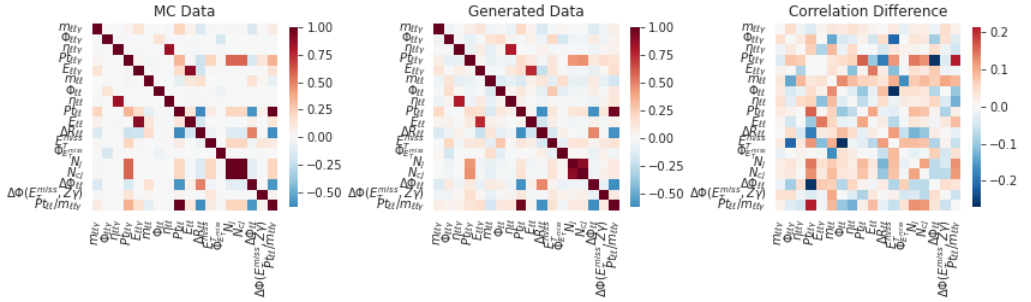

Figure 5. Correlation Comparison of Generated Event Features vs. MC Data for VAE+D Model.

6 Conclusion and Further Work

The addition of the discriminator network to the VAE model has improved the generation results with limited initial optimisation. With further adjustments and hyper-parameter optimisation, the results could achieve better convergence than shown in the results section. A further adjustment that may yield significant improvements is the addition of normalising flows to the VAE model. The addition of normalising flows to the model will allow for more complicated probability distributions to be retained in the latent space, instead of simply forcing the latent space distributions to be normal Gaussian.

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] 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)
[4] 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)
[5] Alwall J, Frederix R, Frixione S, Hirschi V, Maltoni F, Mattelaer O, Shao H S, Stelzer T, Torrielli P and Zaro M 2014 *JHEP* **07** 079 (*Preprint* 1405.0301)
[6] Ball R D *et al.* (NNPDF) 2015 *JHEP* **04** 040 (*Preprint* 1410.8849)
[7] Alloul A, Christensen N D, Degrande C, Duhr C and Fuks B 2014 *Comput. Phys. Commun.* **185** 2250–2300 (*Preprint* 1310.1921)
[8] Sjostrand T, Mrenna S and Skands P Z 2006 *JHEP* **05** 026 (*Preprint* hep-ph/0603175)
[9] 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)
[10] Cacciari M, Salam G P and Soyez G 2012 *Eur. Phys. J.* **C72** 1896 (*Preprint* 1111.6097)