Anomalous Jet Identification via Variational Recurrent Neural Network

Authors: Alan Kahn, Julia Gonski, Inês Ochoa, Daniel Williams, Gustaaf Brooijmans
Nevis Laboratories, Columbia University, 136 S Broadway, Irvington NY 10533

0.1 Method

Our method employs a Variational Recurrent Neural Network (VRNN) to model jets as a sequence of constituents. A VRNN is a sequence-modeling architecture which replaces the standard encoder-decoder architecture of a Recurrent Neural Network with a Variational Autoencoder (VAE). This allows the VRNN to perform both sequence modeling as well as *Variational Inference*, which has been shown to be a very powerful tool in the task of Anomaly Detection [Ref]. Being a sequence-modeling architecture, the VRNN is capable of accommodating variable-length inputs, such as the constituent 4-vectors of a jet. Our motivation for choosing a recurrent architecture is to suppress the ability of the model to learn correlations with the number of constituents within a jet. In fixed-length architectures, such as VAEs, the loss function is computed between the input layer and the reconstructed layer, where zero-padded inputs directly affect the value of the loss function, leading to correlations that are difficult to remove. Considering constituents as a sequence rather than as a fixed-length input allows the model to learn a representation in a way which circumvents these drawbacks.

Figure 1 shows a diagram of one VRNN cell. The VAE portion of the architecture is displayed on the top row of layers in the diagram, where an input constitent's four-momentum components are input as a vector $x(t)$, which is encoded into a gaussian distribution in the latent space z, and then decoded to produce a reconstruction of the same input constituent's components $y(t)$. The variable t refers to the *time-step*, which advances as the sequence is processed, and can be interpreted as the constituent number currently being processed by the model.

Inputs to the VRNN consist of sequences of jet 4-vector constituent components p_T, η, and ϕ, where we assume massless constituents. Before training, we apply a pre-processing method which boosts each jet to the same reference mass, energy, and orientation in $\eta - \phi$ space, such that all jets are superficially the same with the only differences being their substructure. In addition, our pre-processing method includes a choice of *sequence ordering*, in which the constituent sequence input into the model is sorted by k_t-distance instead of by constituent p_T. In more detail, the n^{th} constituent in the list, c_n, is determined to be the constituent with the highest k_t-distance relative to the previous constituent, with the first constituent in the list being the highest p_T constituent. This ordering is chosen such that non QCD-like substructure, characterized by two or more separate cores of constituent clusters within in the jet, is more easily characterized by the sequence. When compared to p_T-sorted constituent ordering, our sequence consistently travels between each cluster, making their existence readily apparent and easy to model. As a result, we have observed a significant boost in performance due to this choice.

$$c_n = max(p_{Tn} \Delta R_{n,n-1})$$

The loss function used is very similar to that of an ordinary Variational Autoencoder. It consists of a mean-squared-error (MSE) loss between input constituents and generated output constituents as a reconstruction loss, as well as a weighted KL-Divergence from the latent space prior to the learned approximate posterior distribution. Since softer constituents contribute less to the overall classification of jet substructure, each KL-Divergence term, computed constituent-wise, is weighted by the constituent's p_T-fraction with respect to the jet's total p_T, averaged over all jets in the dataset to avoid correlations with constituent multiplicity. An additional weight coefficient is enforced as a hyperparameter, and has been optimized to be 0.1 in our studies.

$$\mathcal{L}(t) = MSE + 0.1\overline{p_T}(t) D_{KL}$$

The architecture is built with 16 dimensional hidden layers, including the hidden state, with a two-dimensional latent space. All hyperparameters used result from a hyperparameter optimization scan.

Our model is trained on the leading and sub-leading jets of each event, where our input datasets consist of the entirety Background and Black Box events. After training, we evaluate each jet in the dataset and assign it an *Anomaly Score*, defined as follows, where D_{KL} is the KL-Divergence between the encoded posterior distribution and learned prior distribution:

$$\text{Anomaly Score} = 1 - e^{-\overline{D_{KL}}}$$

Since the LHC Olympics challenge entails searching for a signal on the event-level instead of the jet-level, we determine an overall *Event Score* by choosing the most anomalous score between the leading and sub-leading jets. To ensure consistency between training scenarios, Event Scores are subject to a transformation in which the mean of the resulting distribution is set to a value of 0.5, and Event Scores closer to 1 correspond to more anomalous events.

0.2 Results on LHC Olympics

We first present results with the R&D dataset. In this study we want to directly investigate the performance of the Anomaly Score, and therefore refrain from imposing cuts on additional variables. Using the Event Score as a discriminator, we aim to reconstruc the Z' mass peak within a contaminated dataset consisting of 895113 background events and 4498 signal events corresponding to a contamination level of 0.5%,. Figure 2 shows the two-dimensional histogram of dijet mass vs Event Score, where we see the Z' mass at 3500 GeV clearly visible at higher values of the Event Score. Figure 3 shows the one-dimensional dijet mass distributions before and after a cut on the Event Score at a value of 0.65, which is chosen to retain enough statistics in the background to display its smoothly falling behavior. We also plot the local significance of the signal, which we have computed for each bin. We see that a cut on the

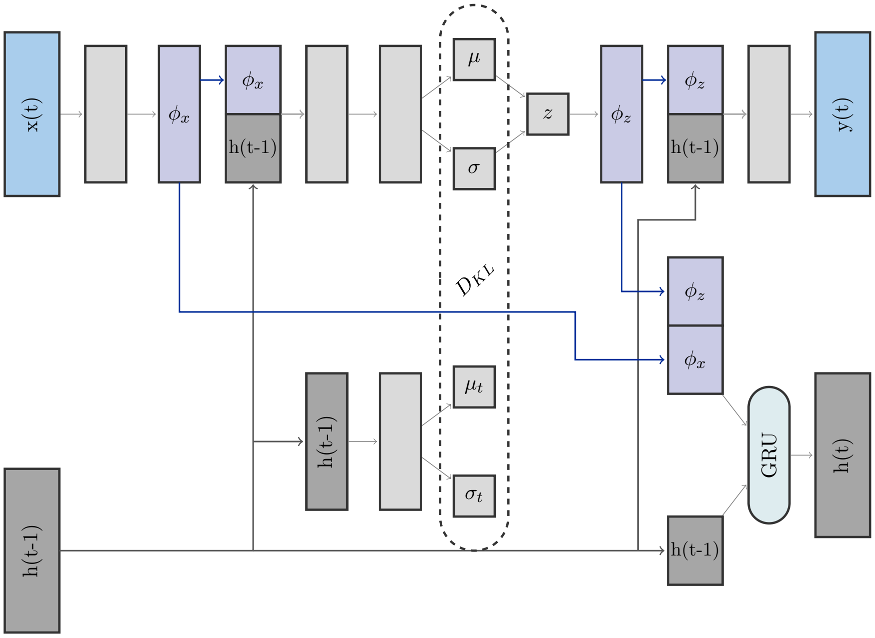

Figure 1. A Variational Recurrent Neural Network cell. The $x(t)$ and $y(t)$ layers represent respectively the input constituent and reconstructed constituents' four-momentum components p_T, η, and ϕ. The phi_x and phi_z layers are *feature-extracting layers* which encode a representation of the features in the input layer $x(t)$ and latent space z respectively. $h(t-1)$ represents the current time-step's hidden state, which is updated each iteration via a transition function between $h(t-1)$, phi_x, and phi_z carried out by a Gated Recurrent Unit (GRU). At each timestep, the prior distribution defined by μ_t and σ_t is determined from the current hidden state

Event Score dramatically increases the significance of the excess. Applying the BinomialExpZ function from RooStats to each bin, and using a relative background uncertainty of 15%, we observe that the peak of the local significance of the signal increases from 0.18σ to 2.2σ at a signal contamination of 0.5% while still retaining the smoothly falling behavior of the background.

Our analysis on the Black Box datasets is performed by applying a cut on the Event Score at a value of 0.75, which is chosen to optimize the signal-to-background ratio, as well as restricting the pseudorapidity of the leading and sub-leading jets to less than 0.75 to ensure that central, high momentum-transfer events are being considered. Applying this method to Black Box 1, we arrive at the mass distributions of Figure 4. Plotted here is the mass of the dijet combination for both the Black Box 1 dataset as well as the Background dataset, displayed as a dashed line. The important feature in this result is the shoulder present at around 4TeV, which corresponds to the present Z' signal at a mass of 3.8TeV.

In Black Box 2, our model shows no significant excesses. Furthermore, the behavior of

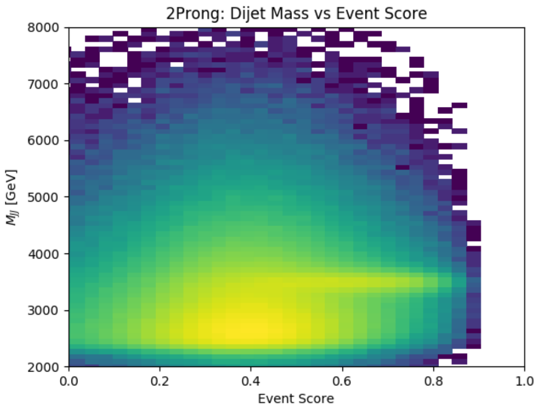

Figure 2. Description of the figure. Reproduced from Ref. [**?**].

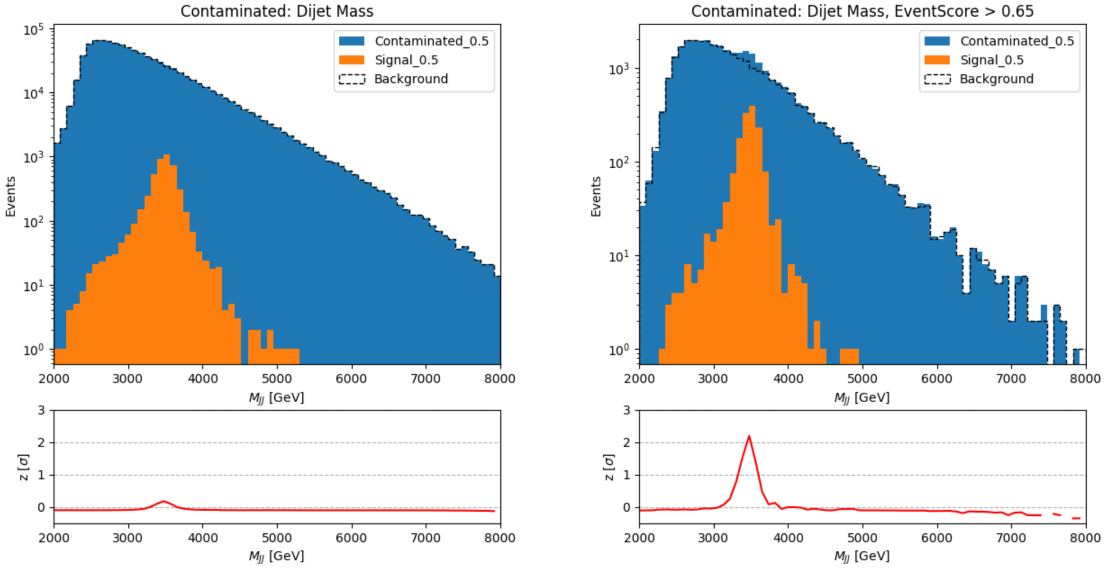

Figure 3. Dijet mass distributions before (left) and after (right) a cut on the Event Score, at a signal contamination of 0.5%

the Event Score consistently affects the distributions of both Black Box 2 and the Background dataset. It is important to note that the model was trained independently on each dataset, and the resulting Event Scores are from entirely unique sets of network weights.

Figure 6 shows our results for Black Box 3. Our model is specifically sensitive to boosted final states, and as a result, we are insensitive to the signal present in this Black Box.

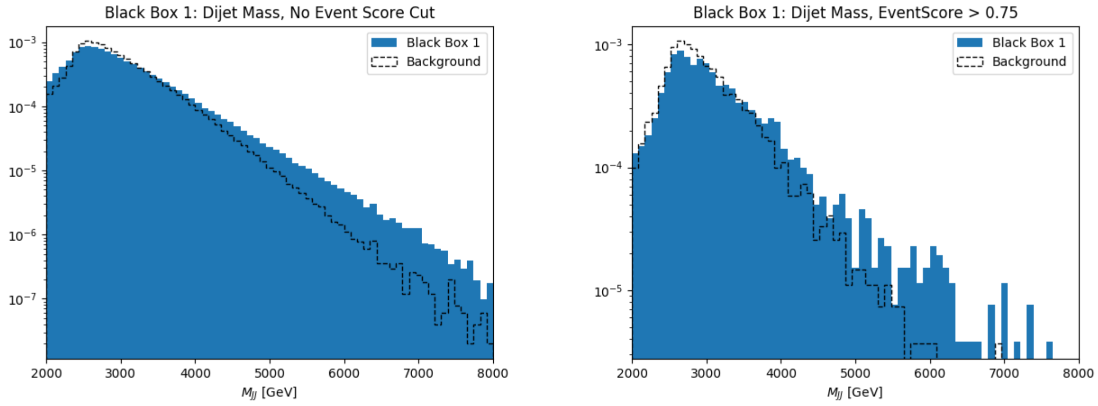

Figure 4. 2-Prong dijet mass distributions before (left) and after (right) a cut on the Event Score, at a signal contamination of 0.5%

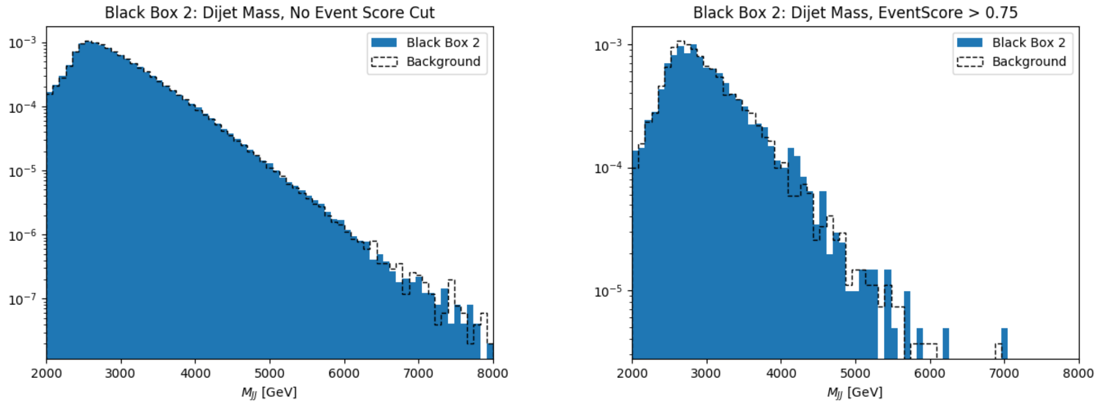

Figure 5. 2-Prong dijet mass distributions before (left) and after (right) a cut on the Event Score, at a signal contamination of 0.5%

0.3 Lessons Learned

This challenge presented a highly accessible avenue for development of our model. We are particularly surprised by the effect of our pre-processing method on the performance of the model, specifically by our choice of k_t-ordered sequencing, and find it to be a very interesting result which may have a number of applications in a wide range of analysis contexts. Overall, we view the VRNN as a model capable of identifying objects within a contaminated dataset as long as they can be characterized by sequential data. While we limited our scope in this study to be entirely unsupervised with absolutely no external information helping the model understand the elements of data that it is modeling, the accessibility of the VRNN's architecture, being both a RNN and a VAE, opens the possibility of accommodating more

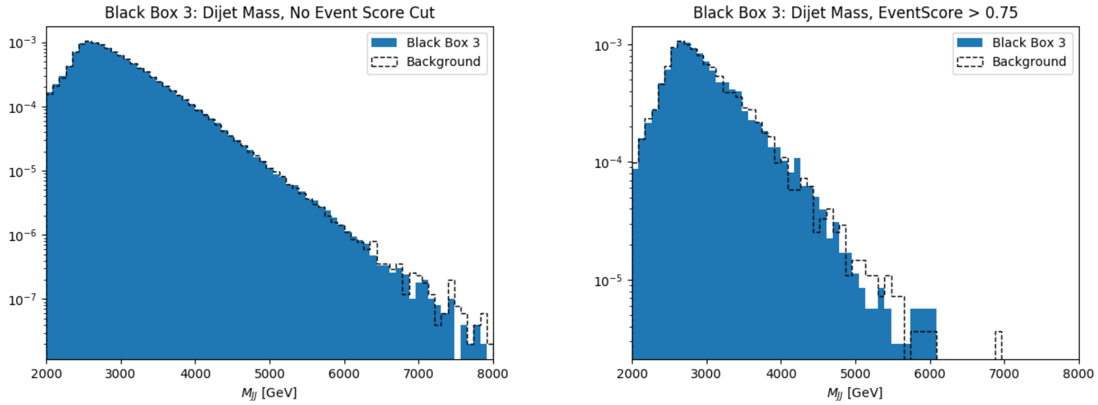

Figure 6. 2-Prong dijet mass distributions before (left) and after (right) a cut on the Event Score, at a signal contamination of 0.5%

supervised training scenarios. We also restricted some of our development tracks in the interest of simplicity, but feel that a number of adjustments to the architecture, such as a dedicated adversarial mass de-correlation network, or an additional input layer representing high-level features, are worthwhile avenues of exploration.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Science under contract DE-AC02-05CH11231.

*For the references, please use names from Ref. [**?**]. If your paper is not there or is not updated, please submit a MR!*