Mass Decorrelated Neural Network Classifier

Benno Kaech, Alessandro Calandri, Christina Reissel, Christoph Grab, Krunal Gedia, Pirmin Berger

ETH Zurich

June 4, 2020
Look at the 0 lepton channel from year 2017: Zvv2017

The signal region we consider is SR_{medhigh}_Zvv

\[ \text{MET}_{pt} > 150 \text{ GeV} \& \text{MET}_{pt} < 250 \text{ GeV} \]

Want a mass decorrelated network, i.e. a network where the classification is independent on the mass variable

In all trainings the mass is therefore excluded from the features used in the training

If it is not mass decorrelated, we see sculpting:

In the highest percentile of the score, the mass distribution of the highest DNN score quantile of the background looks like the mass distribution of the signal
**Vocabulary**

- **score**: Output of the last layer of the classification network. Is a number between 0 and 1. Describes, when using softmax in the final layer, the probability of an event corresponding to the signal class.

- **mass regression network**: Neural net that tries to estimate the mass, using only the score as an input.

- **χ²**: χ² distance between two histograms,

  \[ \chi^2 = \sum \frac{(n_i - m_i)^2}{\sigma_n^2 + \sigma_m^2} \]

  In the following \( n_i \) are the entries in the normalized histogram of the mass of all backgrounds. This is what I call the uncut background. Similarly \( m_i \) stands for the normalized mass histogram in a certain DNN score quantile (mostly the highest one).

- **significance**: bin-by-bin asimov significance applying the classification.

- **hyperparameter**: parameter used to control the learning process of the neural net. It is set before the training and remains constant during the whole training.
Since the hyperparameters have a big impact on the training, it is important to optimize them.

Grid searches make it easy to see which region of the hyperparameter configurations gives acceptable results. In a grid search all possible configuration of a set of hyperparameters is evaluated.

Random searches have generally a faster convergence.

Therefore some grid searches have been conducted at first to get a broad overview. After that random searches in the chosen region lead to best configuration of hyperparameters.

Example for a heatmap of the significance
Quantization of sculpting with $\chi^2$

- The $\chi^2$ only is a shortcut to not have to check all mass distribution plots, for a given network configuration.
- It only marks plots where we see clear sculpting, this is for a $\chi^2 > 10$.
- However it is not a quantity that is used anywhere else, it just makes grid/random searches of hyperparameters quicker to evaluate.

$\chi^2 < 10$, but we still need to check the mass distribution in order to check mass correlation.

High $\chi^2 \rightarrow$ we clearly see sculpting.
Add DisCo correlation term times a hyperparameter $a$ (the importance) to the classifier loss function. This term regularizes the amount of decorrelation, higher $\lambda$ means more decorrelation. However this at the cost of classification accuracy.

$$L_{clf} \rightarrow L_{clf} + a\rho_{disco}(m_{jj}, \text{score})$$

DisCo accounts for non linear correlations between two distributions
DisCo correlation is zero, if and only if the two distributions are independent
DisCo correlation is always calculated on the score of the previous epoch and $m_{jj}$ and then added to the loss function for the current epoch
Minimizing the loss function therefore leads to minimizing the DisCo, i.e. the correlation between the score and $m_{jj}$
Is robust w.r.t the hyperparameter $a$, but has a hard time to account for small correlations.
Quantify the correlation implicitly. If score is independent from $m_{jj}$ it is not possible to estimate $m_{jj}$ from the score.

Adds neural network with 4 layers which tries to estimate the mass using just the score.

Subtract the loss of this net from the loss function of the classifier:

$$L_{clf} \rightarrow L_{clf} - bL_{Reg}$$

$b$ is used a knob to adjust decorrelation, for a higher value of $b$ we expect more decorrelation.

Minimization of this loss implies a maximization of the regression loss.

The regression network has a order of magnitudes bigger learnrate than the classification network, this is so the regression network can account for changes in the classifier quicker.

Again note that the mass regression always uses the score of the previous epoch.
First the network is trained with the DisCo term in the loss function for \( m \) epochs. After \( m \) epochs the DisCo term in the loss function is replaced by the loss of the mass regression for the remaining \( n_{tot} - m \) epochs.

\[
\text{FOR N EPOCHS: } L = L_{\text{crossentropy}} + a\rho_{\text{DisCo}}
\]

\[
\text{AFTER N EPOCHS: } L = L_{\text{crossentropy}} - bL_{\text{reg}}
\]

Note that \( a, b \) differ in few orders of magnitude.

The idea behind this is to use the advantage of both methods. First get a fair amount of decorrelation using the robust DisCo method, and then account for smaller correlations using the adversarial decorrelation technique.
After few grid searches only scanning the importance $\lambda$, a random search with 500 hyperparameter configurations in the number epochs $n$, $a$ and the learnrates at different epochs was conducted.

For the best run the a significance of 1.89 and the following mass distribution is obtained.
For the adversarial multiple grid searches were conducted scanning $\lambda$ and the total epochs $n$

For the best run a significance of 1.70 and the following mass distribution is obtained
For the DisCo adversarial a grid search was conducted scanning $a_{DisCo}$, $\lambda_{reg}$ and $\frac{m}{n_{tot}}$

Until now there has no random search been conducted, in order to move forward

For the best run a significance of 1.82 is obtained with the following mass distribution.
The visibility of the diboson peak around 90 GeV was used as an important criteria for mass decorrelation. This is because, if we do not use the invariant dijet mass as a feature, the diboson peak has exactly the same signature as the signal class. Therefore we would expect a lot of background events around the diboson peak. Due to normalization this is seen as an excess of events over the uncut background. Right now we investigate whether an exclusion of the diboson sample from the MC samples results in the disappearance of the peak around 90 GeV from the mass distribution in the highest score quantile.