Note

Feng-Yang Hsieh

1 CWoLa

The Classification Without Labels (CWoLa) is a weakly supervised learning method. The CWoLa approach trains a model to discriminate the mixed samples, which are mixtures of the original signal and background samples. The optimal classifier in the CWoLa approach is also the optimal classifier in the traditional fully-supervised case where all label information is available. This section utilizes the CWoLa approach to train classifiers on di-Higgs samples.

1.1 Sample

This exercise's signal corresponds to the resonant Higgs boson pairs production in the four-b quarks channel. These Higgs boson pairs are produced via gluon-gluon fusion in the two Higgs doublet model (2HDM). The Higgs boson h ($m_h = 125$ GeV) pair is produced by the heavy CP-even scalar H with mass m_H ranging from 300 GeV to 1200 GeV. The background consists of QCD multi-jet events.

The CWoLa training samples M_1 and M_2 are the mixtures of the signal and background samples. The probability distribution of the mixed sample is a combination of the signal $p_s(x)$ and background $p_B(x)$ distributions:

$$p_{M_1}(x) = f_1 p_S(x) + (1 - f_1)p_B(x)$$
$$p_{M_2}(x) = f_2 p_S(x) + (1 - f_2)p_B(x)$$

$$(1)$$

where f_1, f_2 are the signal fractions, and x represents the observables used for the classification task.

DNN and SPANet network architectures are considered in this exercise. For DNN, the input features are summarised in Table 1, consisting of 16 variables. For SPANet, the input features are a list of final jets, each represented by their 4-momentum $(p_{\mathrm{T}}, \eta, \phi, M)$ and a boolean b-tag.

Table 1: Input variables used to train the dense neural network.

Reconstructed objects	Variables used for training	#
Higgs candidate	$(p_{\mathrm{T}}, \eta, \phi, m)$	8
Subjets	$\Delta R(j_1, j_2)$	2
b-tagging	Boolean for $j_i \in h_{1,2}^{\mathrm{cand}}$	4
Di-Higgs system	$p_{\mathrm{T}}^{hh}, m_{hh}$	2

1.2 Result

The CWoLa training utilizes samples with different signal fractions f_1, f_2 to train the classifiers. The results of CWoLa training are shown in Figure 1 with different signal fractions. When f_1 is far from 0.5, the results tend to approach those of the fully supervised case.
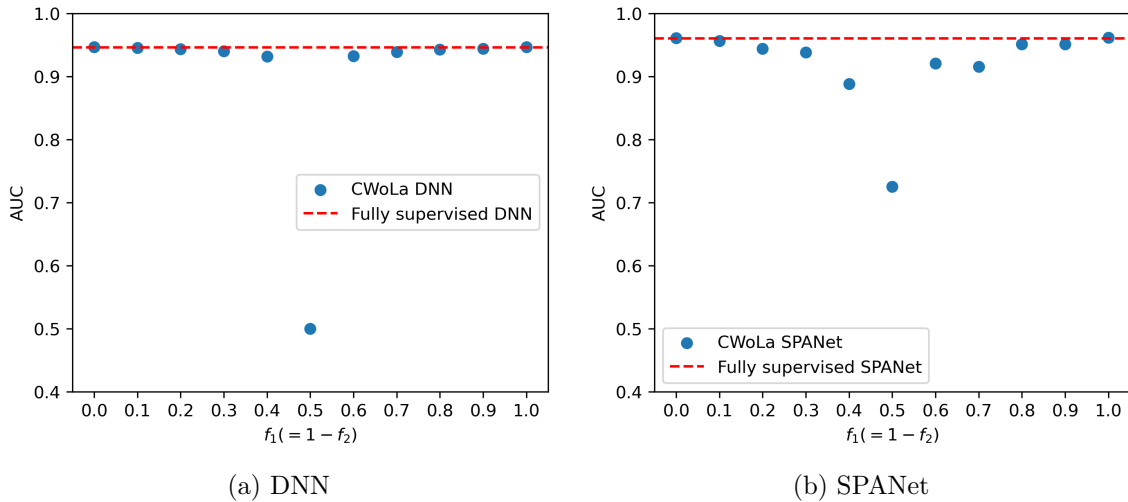

Figure 1: The AUC of CWoLa training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. The horizontal dashed line indicates the fully-supervised AUC.

When $f_1 = 0.5$ the mixed sample M_1 and M_2 have identical distributions, so the classifier can not learn anything in this case. In the case of DNN, the AUC is 0.5, as expected. However, for SPANet, the AUC is more than 0.7.

This is because SPANet is trained on both pairing and classification tasks simultaneously. The pairing part introduces some asymmetries between signal and background samples, leading to the AUC that deviates from 0.5.

To investigate the effect of the pairing task on SPANet's performance, the weight of the

pairing component is set to zero, meaning that SPANet focuses solely on the classification task. Figure 2 shows the SPANet training results without pairing task. In this case, as expected, the AUC is close to 0.5 when $f_1 = 0.5$.
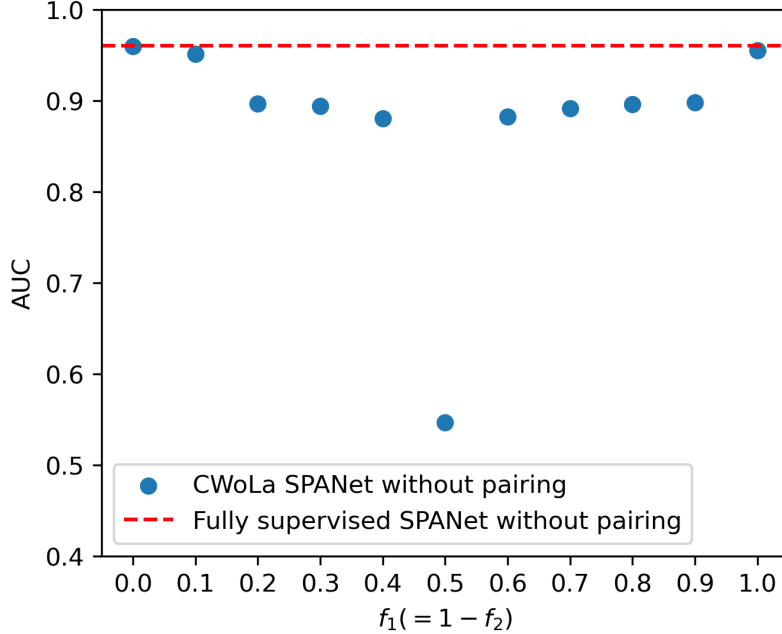

Figure 2: The AUC of CWoLa SPANet training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. Here, SPANet is trained on the classification task only.

2 CWoLa hunting

2.1 Sample

The signal is the resonant Higgs boson pairs production in the four-b quarks channel. In this section, the Higgs boson pair is produced by the heavy CP-even scalar H with mass $m_H = 1000$ GeV. The background consists of QCD multi-jet events. The basic requirement is the "four-tag cut," which requires at least four b-tagged $R = 0.4$ anti-k_t jets with $p_T > 40$ GeV and $|\eta| < 2.5$. Only the events passing the four-tag cut are used in the following analysis.

The CWoLa hunting approach utilizes the signal region and sideband region to create the mixed training sample. The total invariant mass m_{hh} of the di-Higgs system is utilized to determine the signal and sideband region. This quantity is computed from the four b-jets

with the highest transverse momentum. Figure 3 presents the m_{hh} distribution of signal and background samples. The signal region is $[800, 1050]$ GeV and the sideband region is $[700, 800] \cup [1050, 1100]$ GeV. This signal and sideband region are chosen such that the corresponding cross-sections are closed.
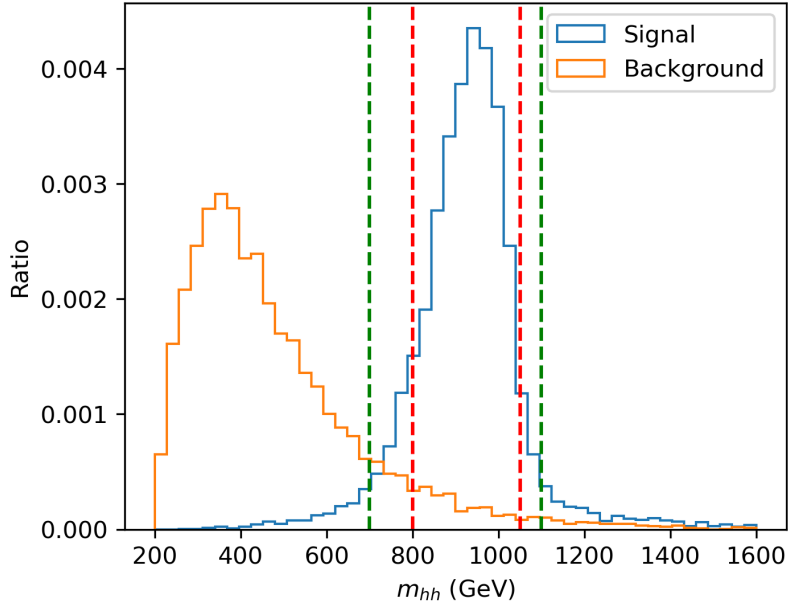

Figure 3: The totoal invriant mass m_{hh} distribution of signal and background samples. The signal region is $[800, 1050]$ GeV between the red dashed lines. The sideband region is $[700, 800] \cup [1050, 1100]$ GeV between the green dashed lines and exclude the signal region.

Table 2 is the cutflow table of the selection cuts. The number of events used in mixed training samples could be computed from these cross-sections. The training sample size is presented in Table 3.

Table 2: The cross sections for the di-Higgs signal and background processes at different selection cuts.

	Cross section (fb)			$\mathcal{L} = 139$ fb^{-1}
	Signal	Background	S/B	$S/\sqrt{B}$
Four tag	0.081	6.03e+03	1.34e-05	0.0123
Signal region	0.063	3.32e+02	1.90e-04	0.0408
Sideband region	0.010	3.19e+02	3.03e-05	0.0064

Consider the DNN CWoLa classifier. The Higgs candidates are reconstructed by the min-ΔR pairing method. The input features are similar to the previous case (Table 1),

Table 3: The training sample size for the mixed sample. The luminosity is $\mathcal{L} = 78 \ \text{fb}^{-1}$ because the generated samples are not enough for now.

	True label	
Mixed sample	Signal	Background
M_1	5	26k
M_2	1	25k

but the b-tagging information and total invariant mass of the di-Higgs system are excluded. For min-ΔR pairing, it only uses the b-tagged jets. Total invariant mass is already used to determine the signal and sideband region.

2.2 Training results

Table 4 presents the DNN classification training results. These numbers are evaluated from the pure samples. The training datasets with and without signal events have similar results. This suggests that the DNN fails to distinguish the signal and background samples but learns the difference between the signal and sideband region. Moreover, the results also imply the input features may correlate to the total invariant mass of the di-Higgs system.

Table 4: The CWoLa DNN training results. The average and standard deviation of 10 training are presented.

	ACC	AUC
With signal	0.868 ± 0.024	0.925 ± 0.023
No signal	0.850 ± 0.033	0.909 ± 0.026

Figure 4 shows the signal score distributions. Even though the signal scores are very different for signal and background distributions, the difference probably stems from the m_{hh} distribution.

There are two issues:

- The input features might correlated to the observables used to determine the signal and sideband region. We need to construct other independent input variables.

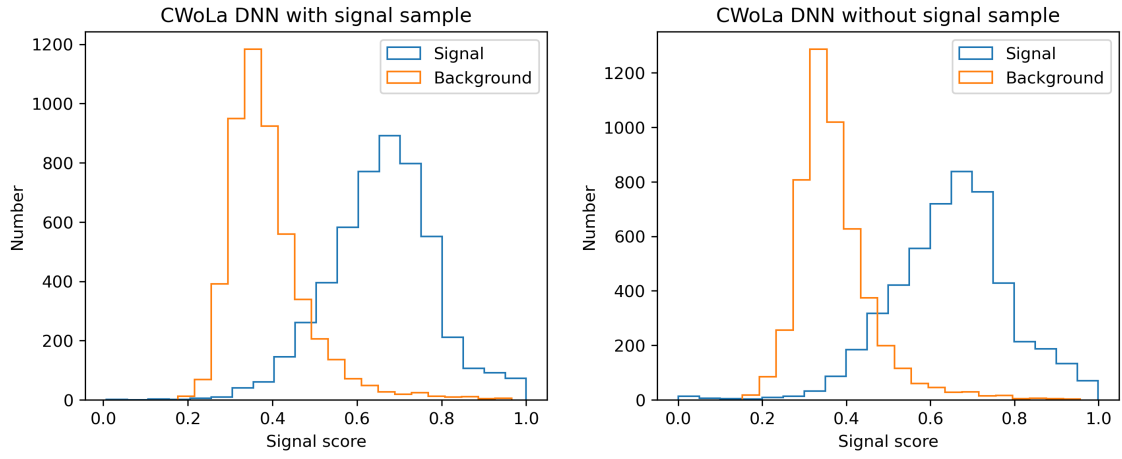- The signal fraction is too low. It is hard to learn something about signal events.

Figure 4: The signal score distributions. We apply the CWoLa DNN on pure samples to obtain the signal score distributions.