Study of Identification of b-jets in the $t\bar{t}b\bar{b}$ process using Deep Learning

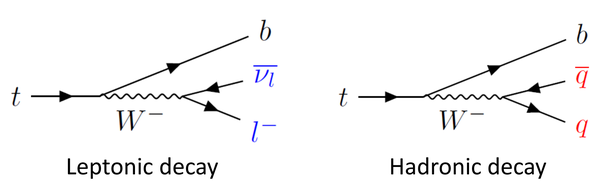
Jongwon Lim, Jieun Choi, Tae Jeong Kim
Hanyang University

Motivation

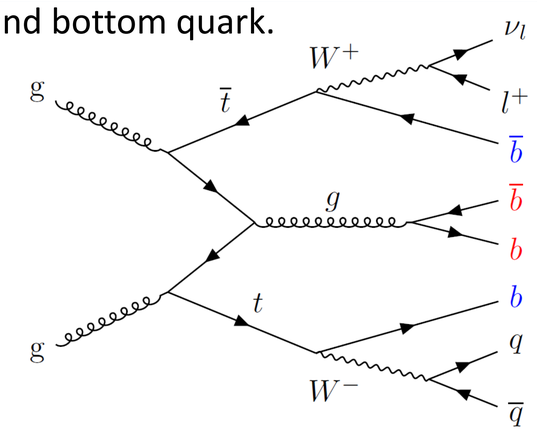
In LHC (Large Hadron Collider), we get much amount of data from particle collision experiment. We are interested in the process which is top quark pair production associated with two bottom quarks ($t\bar{t}b\bar{b}$). Identifying the $t\bar{t}b\bar{b}$ process as a background has crucial roles in separating from $t\bar{t}H(b\bar{b})$ process and searching for FCNC like $tcH(b\bar{b})$. In this study, we will focus on identifying the b-jets of the $t\bar{t}b\bar{b}$ process with particle collision simulation. Traditionally, the b-jets whose ΔR is minimum are selected as additional b-jets from gluon. Also, we will use deep learning method to effectively identify two additional b-jets. Finally, we will compare those two method; minimum ΔR and deep learning method.

Introduction

Top quark ***mostly*** decays into W- boson and bottom quark.
< Two Decay Modes >



Leptonic decay Hadronic decay



Lepton + Jet channel
: both leptonic and hadronic decay
*Jet is a cone produced by hadronization of quarks or gluon.

Additional b-jets from gluon & b-jets from top quark $t\bar{t}b\bar{b}$ process
Where is the origin of the b-jet at a detector level?

Event Simulation

To simulate particle collision experiment like the CMS detector in LHC, we used software products using Monte Carlo simulation.

(1) Matrix Element Calculation ➔ MadGraph & MadSpin
(2) Parton Shower (Hadronization) ➔ PYTHIA
(3) Detector Simulator ➔ Delphes

Collision energy $\sqrt{s}$ = 13 TeV
Total number of Generated Events : 10 million events

Event Selection

[Signal Definition]
Signal : Additional b-jets not from the top quark
Selection 1 (S1) for Generated Additional b-jets : $p_T \geq 20\ \mathrm{GeV},\ |\eta| \leq 2.5$

[Object Selection]
Selection 2 (S2) for a Lepton : $p_T \geq 20\ \mathrm{GeV}, |\eta_e| \leq 2.5, |\eta_\mu| \leq 2.4$
Selection 3 (S3) for Jets : $p_T \geq 20\ \mathrm{GeV},\ |\eta| \leq 2.5$

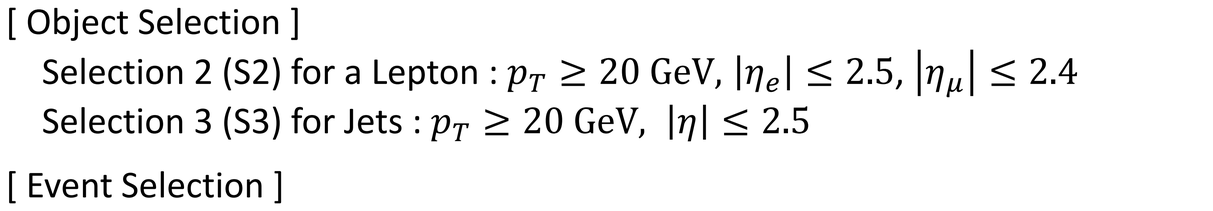
[Event Selection]
Number of jets ≥ 6, Number of b-jets ≥ 3, Number of leptons = 1

[Acceptance]
Acceptance = the number of Selected Events / Simulated Events (10M)
Number of Events after each selection (Acceptance (%))
S1 : 1887355 (18.87%) S2 : 563519 (5.64%) S3 : 55649 (0.556%)
The number of events after the event selection is 55649.

Minimum ΔR Method

[Matching the reco b-jet and gen additional b-jet]
Additional b-jets from gluon have narrow ΔR unlike other b-jets.
By using this property, we can find out the $t\bar{t}b\bar{b}$ process

Reconstructed (Reco) jet : a jet which is detected by a detector
Generated (Gen) jet : a jet which is about to enter the detector



Matching Efficiency : 27.86%
Matching efficiency = the number of Matched Events / Selected Events
Matched Event : 15503 Selected Event : 55649

Deep - Learning Method

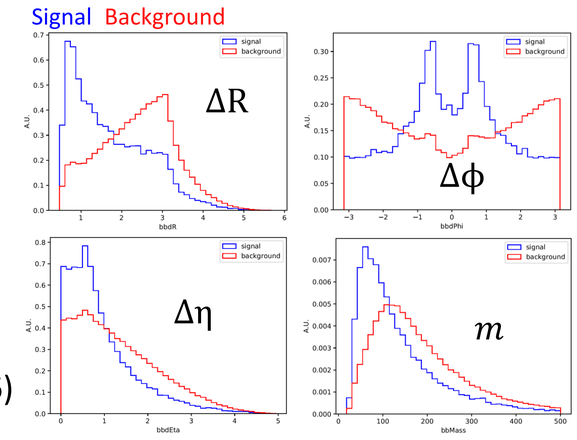
Deep learning (Deep Neural Network algorithm, DNN) is a computational method for complex calculation.

[Deep Learning Environment]
- Using Keras backend Tensorflow
- Activation Function : ReLU
- Loss : Binary Cross Entropy
- Optimizer : Adam
- Batch size : 1024
- Number of Input Variables : 78

✓ p_T, η, energy of first, or second b-jet (6)
✓ $\Delta R, \Delta\eta, \Delta\phi, p_T, \eta$, m, m_T, H_T of
two b-jets (8)
Lepton + two b-jets (8)
MET + two b-jets (8)
first, or second b-jets + Lepton (16)
first, or second b-jets + MET (16)
first, or second b-jets + W (16)

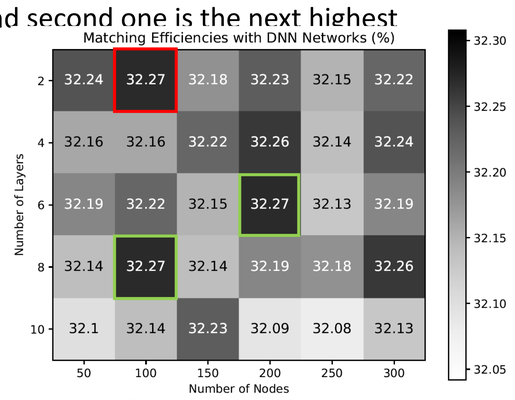
*W : a pair of two jets (not b-jets) that has the nearest mass with W-boson
*First b-jet has the highest p_T among b-jets and second one is the next highest

< Distributions of Some Input Variables (2 b-jets) >
Signal Background



[Hyperparameter Optimization]
Hyperparameter : Node, Layer, Epoch

We choose 100 nodes with 2 layers having the highest matching efficiency for the best nodes and layers.

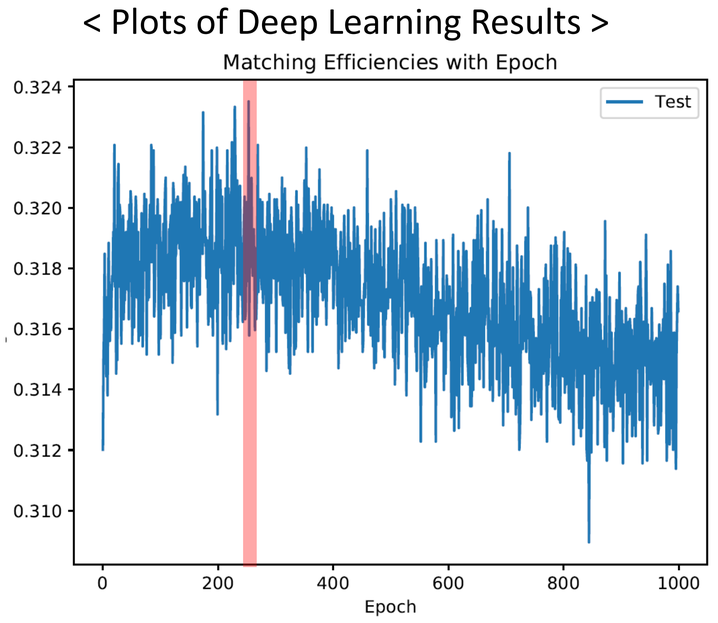


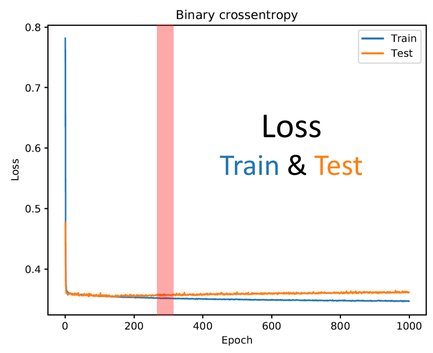
Epoch is set to 100 for test.

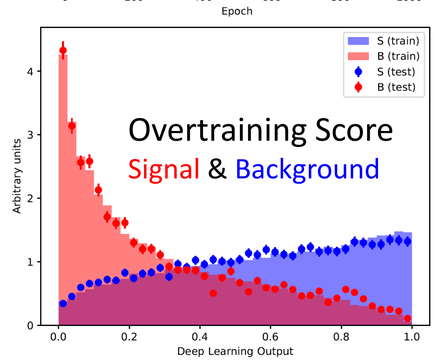
[DNN Result with the Best Network]
Matching Efficiency : 32.35%
We used 80% of selected events for training.

Train Events : 44518 Test Events : 11131 (Matched: 3583)
Training Time : 4184.2 seconds (1 hour 10 minutes)
Epoch for the Best Model : **254**

< Plots of Deep Learning Results >







- After 254th epoch, the matching efficiency decreases.
- Overtraining was not occurred for the best model.

Conclusion

- Deep learning method has a higher matching efficiency than the traditional method, minimum ΔR method. 27.86% ⟶ 32.35%
- By using deep learning, we can better identify the $t\bar{t}b\bar{b}$ process and as a major background, separate from other important signal processes.

References

[1] Jo, Y. K., Choi, S. Y., Roh, Y. J., & Kim, T. J. (2015). Study of the top-quark pair production in association with a bottom-quark pair from fast simulations at the LHC. *Journal of the Korean Physical Society*, *67*(5), 807-812.
[2] Hoffmann, F. (2014). ttbbProduction at the CMS Experiment: Monte Carlo Simulation and Cross Section Measurement. *Karlsruhe Institute of Technology (KIT)*.