國立清華大學

物理系

碩士學位論文

基於 SPA-Net 的雙頂夸克全強子衰變事件重建

Event reconstruction of full hadronic Top-quark-pair decays using SPA-Net

系所組別: 物理所物理組

學號姓名:

108022517 何大維 (Ta-Wei Ho)

指導教授 :

張敬民　教授 (Prof. Kingman Cheung)

徐士傑　教授 (Prof. Shih-Chieh Hsu)

中 華 民 國 一 一 ○ 年 五 月

Event reconstruction of full hadronic Top-quark-pair decays using SPA-Net

A Thesis Presented to
the Department of Physics at
National Tsing Hua University
in Partial Fulfillment for the Requirement of
the Master of Science Degree Program

By

Ta-Wei Ho

Advisor

Dr. Kingman Cheung

Dr. Shih-Chieh Hsu

May 2021

Abstract

The top quarks produces by *pp* collision in Large Hadron Collider(LHC), have a very complicated process still can't be well-classified today. In this project, we present a novel approach to the ‘all hadronic decay‘ process of Top quarks base on the neural networks with attention mechanism, we call it ‘Symmetry Preserving Attention Networks‘(SPA-Net). This networks identify the decay products of each quarks unambiguously and without combinatorial explosion. This approach perform a outstanding result compare to the existing state-of-the-are method. Our network can correctly assigning all hadronic decay in 93.0% of 6 jets, 87.8% of 7 jets, and 82.6% of ≥ 8 jets event respectively.

摘要

在大型強子對撞機 (LHC) 實驗中，經由質子對撞所產生的頂夸克對具有非常複雜的過程以及產物，至今仍無法被非常正確的判別以及重建。在本研究中，我們提出了一個利用新穎的機器學習方法來對雙頂夸克全強子衰變過程進行重建。此方法基於 Attention mechanism，我們稱之為 Symmetry Preserving Attention Networks(SPA-Net)。這個模型架構可以在避免組合性爆炸的前提下對所有的衰變產物進行辨識以及重建。此方法對比於傳統的 χ^2 重建方式，表現出了非常巨大的差異。本方法可以在一、存在 6 jets 條件下正確的重建 93% 的事件；二、存在 7 jets 條件下正確的重建 87% 的事件；三、存在大於 8 jets 條件下正確的重建 82.6% 的事件。

Acknowledgements

I am so grateful for the helps from my colleagues, aadvisors, collaborators and friends. They gave me a lots of help when I was struggle in the research and technical issue. Their kindless helps me to come over the advantages and receive the degree.
When I doing this project, my advisor and my collaborator, Prof. Kingman Cheung and Prof Shih-Chieh Hsu gave me so much support that makes me can find a path to accomplish this project. Their suggestions and strong knowledge support are my best backup when I struggled in the mist. My colleague, Yi-Lun Chung, who helps me to resolve the problem of simulation package and theoretical problems, is also a person who deserved my most gratitude.
I appreciate all the collaborator who participate in this project, Mike Fenton, Alexander Shmakov, Daniel Whiteson, and Pierre Baldi. I would not finsihed this project without their help and support. The way that Alexander construct a machine learning model and the way that Mike describing a physics concept really teach me a lot. The guidence of Daniel, Shih-Chieh, and Pierre inspire me to explore more.

我想感謝我的同事、教授、朋友以及合作者。若沒有他們的鼎力相助，我肯定無法順利地完成這個研究。在我遇到各種問題，在困難中掙扎時，他們友善的協助讓我跨越難關，最後終於取得了學位。

張敬民教授以及徐士傑教授作為我的指導教授以及合作者，不論在物理上或是機器學習上，都非常大方地給予我非常多的建議。這些建議讓我能夠以足夠的知識來執行這個研究。他們的建議，每每都在我遇到困難時發揮關鍵的作用，也在我陷入迷霧時給予了我一盞明燈。

鍾沂倫，研究室的博士班學長，也在研究中給予了我相當大的幫助。無論是在各種環境的設定以及工作站的維護，乃至模擬軟體的設定以及物理模型的參數設定，都給了我非常多的幫助。他的協助讓我在維護工作站以及管理上輕鬆了非常多，真的是非常感謝他的幫忙。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

At Large Hadron Collider(LHC), two protons collide with very high energy and produce many kinds of products. A process that pp collision produces a pair of Top quark and result in the 6 jets final state is called **Full Hadronic Top-quark-pair decay**. This process has a very complicated signature due to a large number of combinations. These jets produced by the top quark pair is hard to tag as a specific parton correctly. A traditional method is to reconstruct the event using χ^2 reconstruction, but it takes such a long time to compute and cannot provide enough accuracy to reconstruct an event. The importance of studying Top quark and its full hadronic decay channel is 1. Top quark is the most heaviest fundamental particle in standard model and will decay before hadronization, 2. The brach ratio of full hardonic decay is the biggest part in Top quark decay(46%).

For a problem that contains a large amount of data and highly requires computing resources, machine learning can widely provide powerful support on solving the problem and helps to reduce the time-wasting. The machine learning method helps to discover physics phenomena with very outstanding effort. A remarkable discovery that helps by machine learning is the discovery of Higgs Boson. Both CMS and ATLAS groups apply the machine learning method to promote the searching of Higgs Boson. [1][2]

In this thesis, we perform a novel architechture for parton-jet assignment problem. This method is base on the state-of-the-art machine learning technology, Attention mechanism. We call this novel ML model **Symmetry Preserving Attention NETworks (SPA-NET)**. By applying attention networks, the SPA-NET perform a outstanding performance compare to traditional method while avoiding combinatorial explosion. And thanks to the natural properties of attention network, the network reflect the permutation symmetry naturally and provide a chance to explore in set-based output.

We will discuss the Top physics and the concept of machine learning in chapter 2; and explain our event generation and simulation configuration in chapter 3; then introduce how we analyze the dataset and reconstruct the event using traditional method and ML approach in chapter 4. We will discuss our work in chapter 5 and summerize in chapter 6.

Chapter 2

The Top Physics and Machine Learning

2.1 The Top Physics

Top quark, the most massive fundamental particle in Standard Model(SM), is the only quark that decays semi-weakly. Its large mass leads to a short lifetime and decay before hadronization occurs. Top quark contains so many properties that interest us, such as its mass, couplings, and cross-section, e.t.c. Measure these properties accurately can bring us a worth understanding of fundamental interactions and the key to Beyond Standard Model.[3]

In recent model, Top quark pair produced by pp collision has three decay modes, **all-hadronic channel**, **semi-leptonic channel**, and **dileptonic channel**. The branch ratio of each channel, has shown in the Table 2.1. The decay width of Top quark predict in SM is[4]:

$$\Gamma_t = \frac{G_F m_t^3}{8\pi\sqrt{2}}\left(1-\frac{M_W^2}{m_t^2}\right)^2\left(1+2\frac{M_W^2}{m_t^2}\right)\times\left[1-\frac{2\alpha_s}{3\pi}\left(\frac{2\pi^2}{3}-\frac{5}{2}\right)\right] \tag{2.1}$$

Table 2.1: Top quark pair decay process[3]

Decay Channel	Process	Branch Ratio(%)
All-hadronic	$t\bar{t} \to W^+bW^-\bar{b} \to q\bar{q}'bq''\bar{q}'''\bar{b}$	45.7
Semi-leptonic	$t\bar{t} \to W^+bW^-\bar{b} \to q\bar{q}'b\ell^-\bar{\nu}_\ell\bar{b} + \ell^+\nu_\ell bq''\bar{q}'''\bar{b}$	43.8
Dileptonic	$t\bar{t} \to W^+bW^-\bar{b} \to \ell^+\nu_\ell b\ell'\bar{\nu}_{\ell'}\bar{b}$	10.5

In recent study, the most precise result of Top quark mass is measured in the lepton+jets channel due to its good signal-to-background ratio and the presence of one neutrino final state. Although the all-hadronic channel has the most probability to appears in the Top quark pair decay process, it couldn't provide a precise mass measurement due to its poor signal-to-background ratio. The poor signal-to-background ratio of the all-hadronic channel is due to the difficult QCD background. The CMS and ATLAS group approach a precision of Top mass measurement using the all-hadronic channel with 0.65% and 1.1%.[5][6]

The channel we are interested in this project is the **jet-parton assignment problem in all hadronic decay channel**. The reason that we are interested in this channel is the resolved 6 jets signature and the potential of the machine learning method apply to the ambiguous event reconstruction problem. There exist 6 jets in the final state, 2 b-jets and 4 quark jets, they can be separated into two groups (b, q, q') and $(\bar{b}, q'', q''')$. A schematic of the decay products is shown in 2.1.

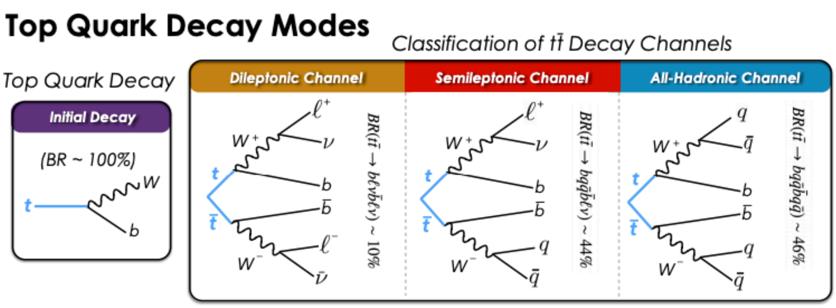

Figure 2.1: The schematic of Top quark decay channels.[7]

2.2 Machine Learning and its application on Particle Physics

Machine Learning has been applied to most of the region in recent age, so dose particle physics. From the search of higgs boson(neural network and BDT) to the b-tagging technology(BDT[8]), physicist already applied several kinds of machine learning method to recent researchs.

In a nutshell, machine learning can break into several cases, it can help to do classification, regression, and clustering problems. It can not only help to accelerate the computation of well-defined problems, and also find a new path to unsolved area. We will use a state-of-the-art machine learning technology, attention mechanism. The attention mechanism is a technology base on the evolution of RNN.[9] The attention mechanism will not only consider the local relationship and the sequence neighbor but also calculate the global relation base on the self-attention calculation shown in Figure 2.2. Using this novel architecture, we will train on the relationship between each jet and try to figure out the correct pair information.

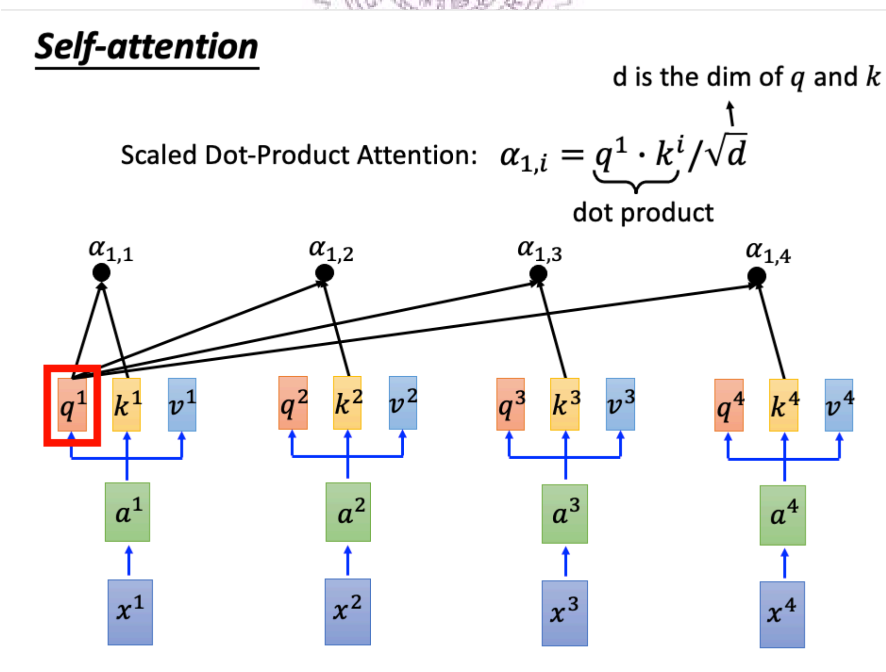

Figure 2.2: How self-attention works.[10]

Chapter 3

Event Generation

3.1 MC samples

For data preparation, we generate our dataset using a custom docker image with MadGraph_aMC@NLO(v2.7.2), Pythia8(v.8.2), and Delphes(v3.4.2) for showering, hadronization, and detector simulation. We apply the ATLAS parametrization during detector simulation. The data are generated at Leading order, including quantum chromodynamics(QCD). The top mass is configured as $m_{top} = 173\ GeV$. We force the W quark decay hadronically into a (q, q') pair. Following is our configuration:

```
generate p p > t t~ QED=0, (t > W+ b, W+ > j j), (t~ > w- b~, w- > j j )
output <file_path>
launch <file_path>
shower=Pythia8
detector=Delphes
analysis=OFF
done
set nevents = 10000
set iseed = 1
Delphes/cards/delphes_card_ATLAS.tcl
done
exit
```

Listing 3.1: Configuration for generating samples

To get a more general performance, we scan the iseed value from 1 to 30000, each value has around 100 files with 10 thousand events before event selection. The reason for scanning iseed value is that the iseed value is the key to the random generation. Originally, the program will choose the iseed value randomly and generate different samples. By scanning the issed value, we can check whether the network can work well on different iseed number or not.

Chapter 4

Data analysis and Event reconstruction

In this chapter, we will first explain the way we analysis the dataset, then how we apply the truth matching, traditional reconstruct method, and the machine learning approach.

4.1 Data analysis

4.1.1 Event selection

The top all hadronic decay channel has 2 b-jets and 4 quark jets, all of them in our configuration are not in the boosted region. Follow the event selection used in the reference[7], we apply a event selection that an event should at least exists **2 b-jets** and **4 quark jets** satisfied p_T larger than **25 GeV** and $|\eta|$ less than **2.5**. A cutflow table and figure can help us to understand how many events are killed by the selections. We may apply 5 cuts and see the evolution of survived event numbers. The rule of cuts is shown in Table 4.1, and the cutflow is shown in Figure 4.1. As the result, we found around 1820% of events will survive after the event selection.

Table 4.1: Rule of cuts. All the cuts require a kinematic limitation that $p_T > 25$ GeV and $|\eta| < 2.5$.

#Cut	Number of b-jets	Number of quark jets
C1	0	4
C2	0	5
C3	0	6
C4	1	6
C5	2	6

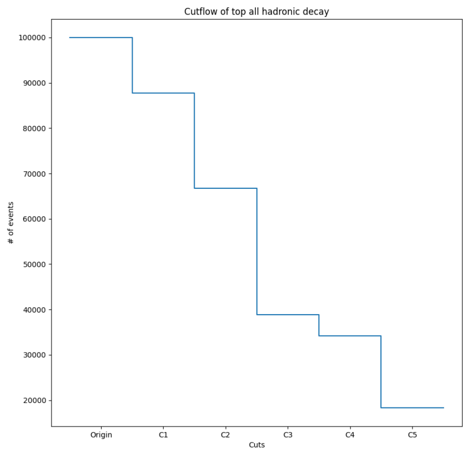

Figure 4.1: Cutflow of all hadronic top decay.

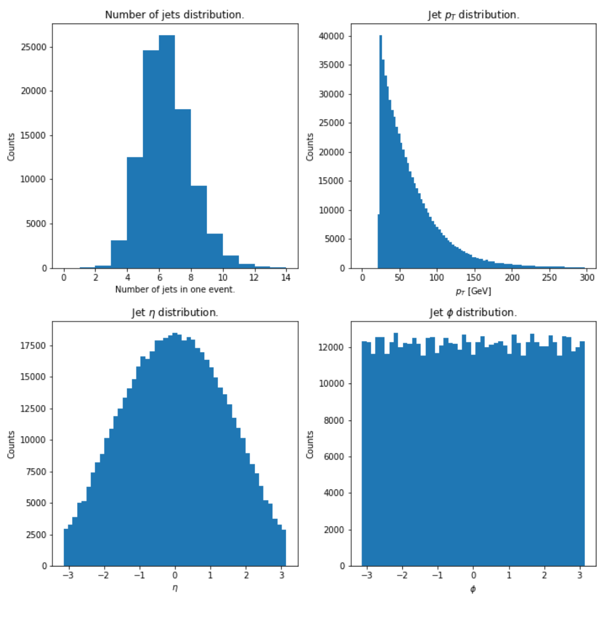

Figure 4.2: Demonstration of distributions.

4.1.2 Truth matching

The **truth matching**, which is also called **ΔR matching**, is to match the detector simulation(i.e. jet information generate by Delphes) data to truth record(i.e. Parton level information). To calculate the ΔR value, we will find the daughters of top quarks, W boson, and b quark. After the daughters of top quarks are found, we will find the daughters of W bosons. Finally, we will get six partons that come from the decay of top quark pairs. These six partons can match the jets identically by considering their distances. The formula of calculating ΔR is:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{4.1}$$

By using the kinematic properties provide in parton level and detector simulation information, we can calculate the ΔR value between each parton and jets. Using the result of the calculation, we may assign each parton to a specific jet.

4.1.3 Custom barcode system

To specify the relation between each parton, and the relation between mothers and daughters, we design a barcode system that helps us to declare the relationship.

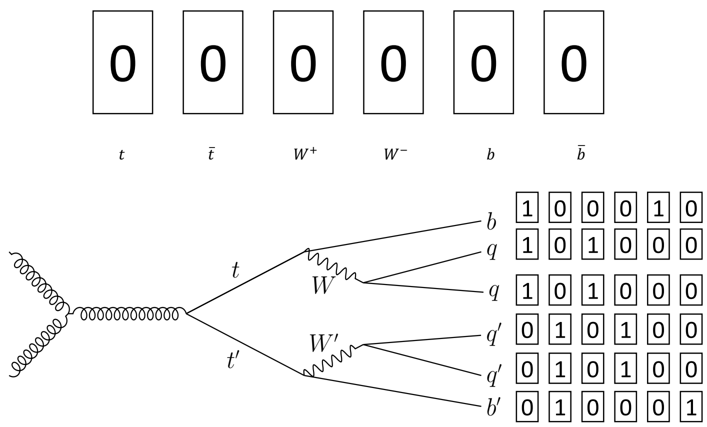

Figure 4.3: Design of barcode.

In Figure 4.3, we define a six-digit barcode, the first two digits are to show which top quark is the mother of this parton, the last four digits of the sequence is to declare which daughter of the top quark is the mother of parton. In case, we can use this barcode system to break six parton(jet) candidates into two subsets which contains 3 elements. The benefit of using this barcode system is not only can specify the relation without losing the information, but also provide a permutation relationship to our network. We will discuss this in the following section.

4.2 Event reconstruction

4.2.1 χ^2 minimization method

The χ^2 minimization method is a traditional method to reconstruct an event. For an event that exists 6 jets, it has about $6!/(2 \times 2 \times 2) = 90$ possible combinations. The number of probable combinations is proportional to the number of existing jets in an event. The χ^2 minimization method will calculate all the candidates and try to find the candidate which has the smallest χ^2 value. This method base on the masses of the W boson and top quark(or you may consider the difference of the top masses reconstructed by two subsets.). The equation of χ^2 minimization in this model is:

$$\chi^2 = \frac{(m_{bqq'} - m_{\bar{b}q''q'''})^2}{\sigma^2_{\Delta_{m_{bqq'}}}} + \frac{(m_{qq'} - m_W)^2}{\sigma^2_W} + \frac{(m_{q''q'''} - m_W)^2}{\sigma^2_W} \tag{4.2}$$

4.2.2 Machine Learning Approach

For a machine learning model, equivariance and invariance are the important properties that may affect the performance of the model. Such as the computer vision problem, the object should be invariant under translation to prevent affect the prediction. Convolutional Neural Network(CNN) can produce object recognition outcomes that are invariant under translations. The properties of invariance can be generalized to another geometry structure, e.g. manifolds and groups. In all hadronic top decay, we have two subsets with the same elements (b, q, q') and $(\bar{b}, q'', q''')$. These subsets should remain invariant

under permutations of the input jets order.
The attention mechanism is an architecture that allows the network to propagate the information selectively by using a "mask". By the implementation of the mask, the neural network can learn from partial information and update the parameters with selected information, and allow the network to infer the relationships between different elements in a sequence.
By the invariant feature of attention architecture, rearranging the elements in a sequence leaves the attention weight unchanged. We may use this permutation symmetry present the attention-based model can handle the reconstruction of all hadronic top decay process efficiently. In case, the network output of the all hadronic top decay process should identify two distinct interchangeable subsets, each contains interchangeable qq‘ pair. This invariant property on the output is the unique feature of our dataset and the model should take into account.

We propose an attention-based network, called **Symmetry Preserving Attention NETwork(SPA-NET)** in this project. Its structure is shown in Figure 4.4. The input of SPA-NET is a list of unsort jet information, with their 4-momentum (p_T, η, ϕ, M) as well as the b-tag information provided by Delphes. We take the logarithm to M and p_T and normalize them to have zero mean and unit variance. The input jets will be sent to the network and be embedded into a D-dimensional latent space representation. After embedding, the output will be sent to a stack of transformer encoder layers, this layer will learn the relationship between each element in the input sequence. While the encoding is finished, the encoded output will be forwarded into an important architecture in this network, a two branches structure that computes the output individually. There is a transformer encoder in each branch, this encoder layer will extract the information of top quark and the tensor attention layer will produce the top quark distribution.
The most important part in this network is the **Symmetry Preserving Tensor Attention**. Consider a set of weight $\theta \in \mathbb{R}^{D\times D\times D}$, This θ is not inherently symmetric at all. To make the θ become an invariant attention weighting, we apply the following transformation(eq. 4.3). This transformation will transform the θ into an auxiliary weights tensor $S^{ijk} \in R^{D\times D\times D}$. Using S^{ijk} and the embedded

jets tensor $X \in \mathbb{R}^{N\times D}$(N is the number of jets), we can calculate the dot-product attention. The dot-product attention working in flat Euclidean space and produce the output tensor O^{ijk}. The summation product tensor S^{ijk} guarantees the interchangeable of the first two dimensions of S will be symmetric and ensuring that $O^{ijk} = O^{jik}$. These properties enforcing the qq' invariance.

$$\begin{aligned} S^{ijk} &= \frac{1}{2}\left(\theta^{ijk} + \theta^{jik}\right) \\ O^{ijk} &= X_n^i X_m^j X_l^k S^{nml} \end{aligned} \tag{4.3}$$

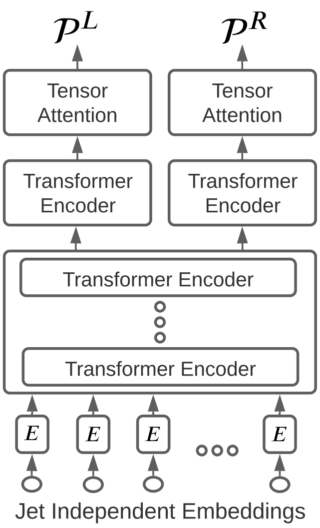

Figure 4.4: High-level structure of SPA-NET.

To obtain the probability distribution P^L and P^R, we apply a 3-dimensional softmax on O^{ijk} to generate the joint triplet probability distribution.

$$[h]P(i,j,k) = \frac{expO^{ijk}}{\sum_{ijk} expO^{ijk}} \tag{4.4}$$

We use equation 4.4 produce the individual probability distribution of two top quarks and produce the single triplet from each by selecting the peak of these distributions.

During the training, a suitable loss function is indeed to deal with the double output probability distributions. We design the loss function base on the cross-entropy between the output probability and truth distribution on the all hadronic top decay. The loss function must ensure the symmetry the top quark pairs are invariant concerning the permutation $tt' \leftrightarrow t't$. We create a symmetry loss function $\mathcal{L}$ by the following function:

$$\mathcal{L} = min(\mathcal{L}_1(P^L, T_1, P^R, T_2), \mathcal{L}_1(P^L, T_2, P^R, T_1)) \tag{4.5}$$

Where $\mathcal{L}_1(P_1, T_1, P_2, L_2) = \mathcal{H}(T_1, P_1) + \mathcal{H}(T_2, P_2)$, and $\mathcal{H}$ is the cross-entropy that $\mathcal{H} = \sum_{(x,y)\in(X,Y)} -xlog(y)$. It is possible that both two branch produce the same output pairs. To make sure the network produce unique predictions, we will select the higher probability one and re-evaluate the other one, then compute the loss function. The Figure 4.5 is a example of the output produce by SPA-NET.

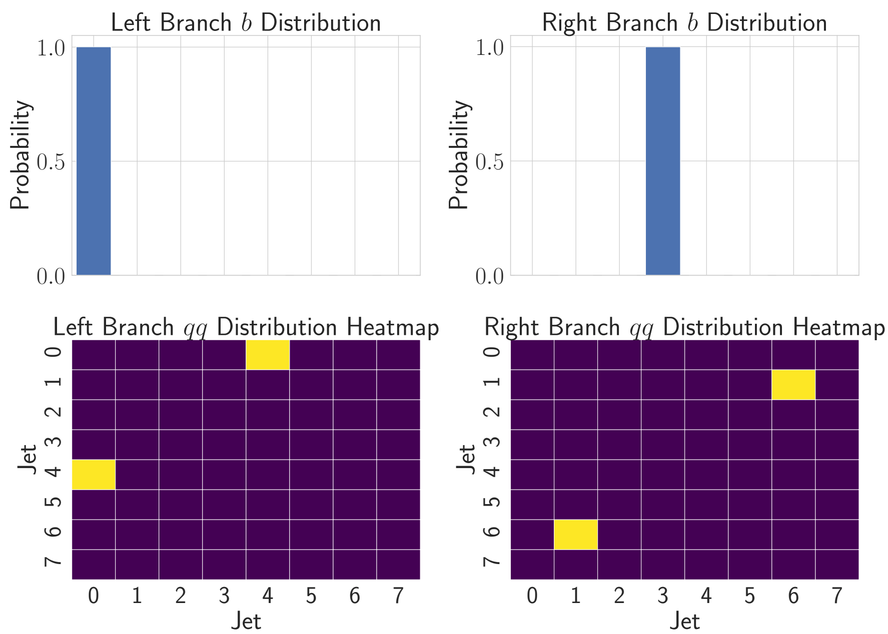

Figure 4.5: A visualization of the example single event output produce by SPA-NET. The top two plots is the projected b quark distribution, and the bottom two plots is the qq distribution respectly.

Chapter 5

Result and Discussion

In this chapter, we will discuss our result and compare the performance between the traditional method and machine learning approach.

5.1 Invariant mass and reconstruct efficiency

Before discussing the result, we should define the category for the reconstruction efficiency. In case, treating the truth matching result as the target and the prediction of χ^2 or SPA-NET may produce three kinds of results:

- Correct matched: An event that both top quarks are correctly predicted by the reconstruction method.
- Incorrect matched: An event that one or both top quarks is incorrectly predicted by the reconstruction method.
- Unmatched: An event that one of the truth match results contains an element that does not match to any jets.

Base on the category above, we can compute the reconstruction efficiency. For the reconstruction efficiency, we separate it into two parts: event-based efficiency and quark-based efficiency. For event-based efficiency, we will calculate the efficiency base on how many **events** are matched correctly, incorrectly, and unmatched. Meanwhile, the quark-based efficiency calculates the efficiency base on how much **quarks** are assigned correctly. The efficiency is shown in the table below:

Table 5.1: Performance of the χ^2 and SPA-NET assignments assessed by per-event efficiency ϵ^{event} and per-top efficiencies ϵ^{top} inclusively and by jet multiplicity N_{jets}.

	χ^2 Method			SPA-NET		
N_{jets}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}
6	61.8%	65.0%	24.2%	80.7%	84.1%	56.7%
7	40.8%	50.4%	24.6%	66.8%	75.7%	56.2%
$\geq$8	23.2%	35.5%	20.2%	52.3%	66.2%	52.9%
Inclusive	**37.7%**	**47.0%**	**23.0%**	**63.7%**	**73.5%**	**55.2%**

The χ^2 has performed a 37.7% efficiency on overall events, while SPA-NET archives an efficiency 63.7%. The χ^2 method has the best performance on the 6 jets category and has the worst effort on the category that an event contains more than 8 jets. The SPA-NET perform much better the χ^2 in all category. For the event that contains two identifiable, the χ^2 method achieves an efficiency ϵ_2^{top} 65.0%, and SPA-NET archives an ϵ_2^{top} 84.1%. Due to we train the SPA-NET we the events that contain two top quarks, so it is reasonable that the ϵ_1^{top} of SPA-NET is lower than ϵ_2^{top}. Also, since our definition of equation 4.2 is base on the difference between reconstructing invariant mass of 2 two quark candidates, so an event that only contains one identifiable top quark, χ^2 method is hard to assign the jet properly. Note that in our evaluation dataset, 8.1% of events in which both tops are identifiable have at least one b-quark matched to non-b-tagged jets. These quarks, which are impossible for the χ^2 to correctly reconstruct, is reconstructed by SPA-NET with an efficiency of 29.4%.

5.1.1 Reconstructed invariant mass

Using χ^2 minimization, we may obtain the best assignment under the constraint of parameters. In this project, the parameters are configured as: $m_W = 81.3GeV$, $\sigma_W = 12.3GeV$, and $\sigma_{m_{bjj}} = 26.3GeV$, all of the parameters are found by fitting the mass distribution of W boson and top quark. While computing $\chi^2 value$, we use the b-tagging information to assign the b quark candidates. This may help to reduce the number of permutations but prevent the event that contains a jet mistagged be assigned correctly.
Following is the reconstructed mass distribution of W boson and top quark using χ^2 minimization method and SPA-NET.

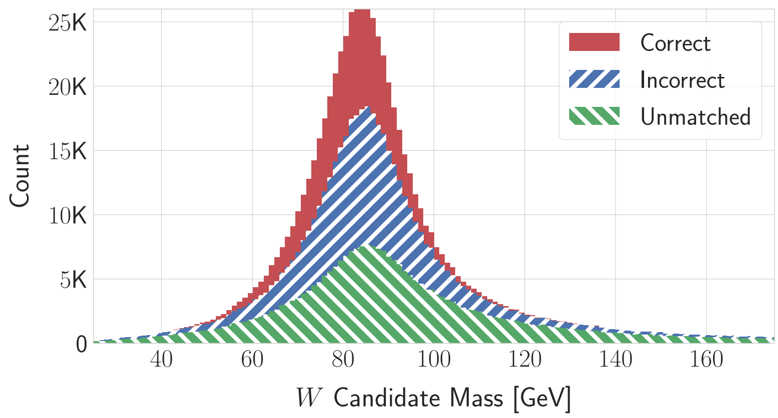

Figure 5.1: W boson mass reconstructed by χ^2 minimization method

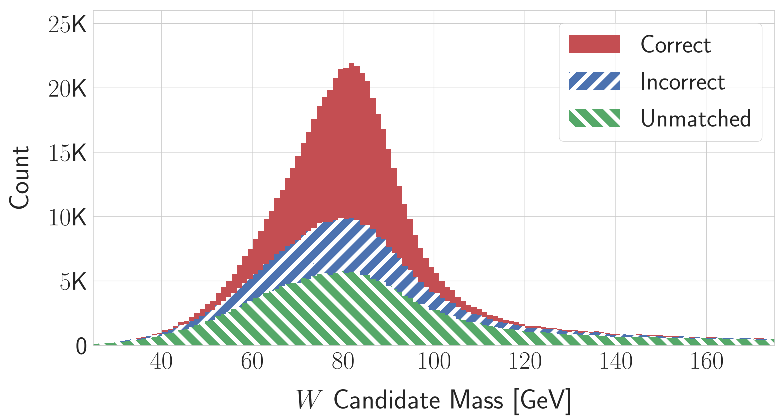

Figure 5.2: W boson mass reconstructed by SPA-NET

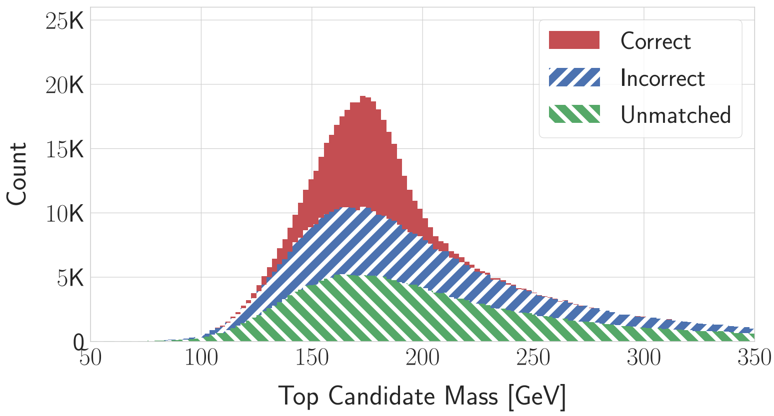

Figure 5.3: Top quark mass reconstructed by χ^2 minimization method

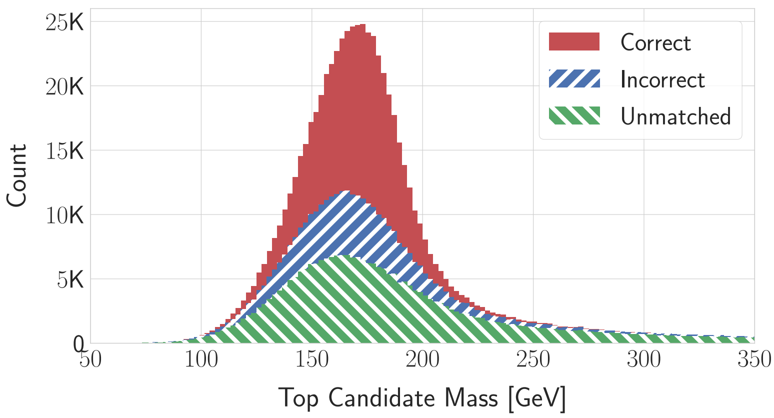

Figure 5.4: Top quark mass reconstructed by SPA-NET

Comparing the result shown in Figure 5.1 and Figure 5.2, we found the χ^2 method has a narrow peak around W boson mass than the SPA-NET. This shape comes from the incorrect and unmatched events and can be explained by the presence of m_W in equation 4.2. Another point is the Figure 5.4 and Figure 5.3 shows that the SPA-NET has the more peaked distribution compare to χ^2 method with comparable incorrect/unmatched events.

5.1.2 ROC curve

The Receiver operating characteristic (ROC) curve is a good target to estimate the performance of a machine learning model. The ROC curve of SPA-NET apply on events with one and two reconstructable top quarks is shown in Figure 5.5 and Figure 5.6. Note that the targets are defined as 1 if the prediction is correct, otherwise the 0 represents the incorrect prediction.

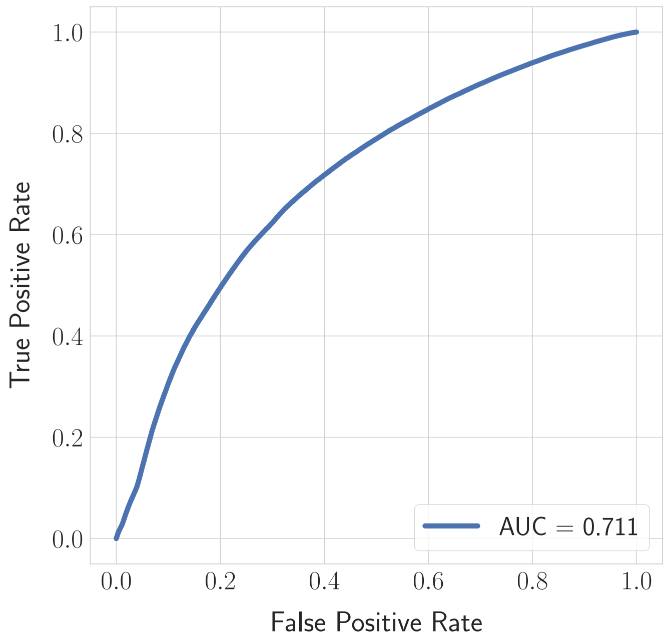

Figure 5.5: ROC curve of SPA-NET apply on events with one reconstructable top.

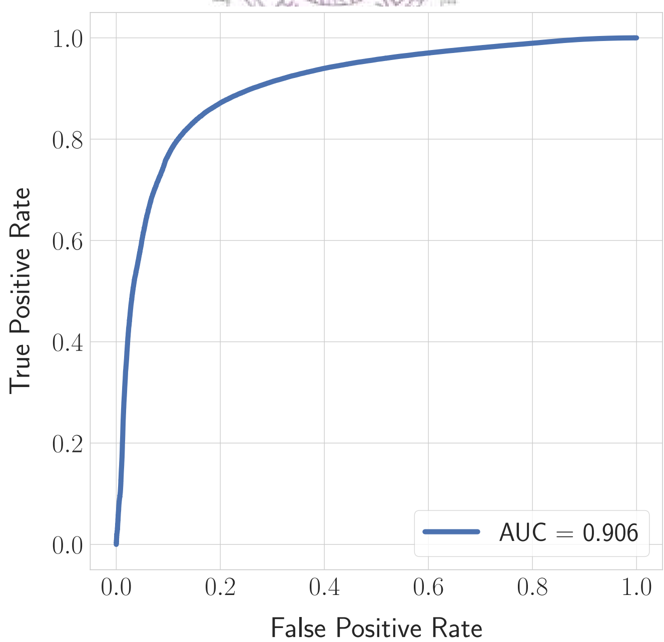

Figure 5.6: ROC curve of SPA-NET apply on events with two reconstructable top.

5.2 Outlook

Base on the design of SPA-NET, the input is a sequence of jets information and their relationships. The network considering the permutation relation between each subset and possible permutations. In case, a physics model which has a permutation relationship might be a good target to explore with SPA-NET. For example, ttH all hadronic decay process or all hadronic four top decay process is a potential target to studying with SPA-NET. Further more, not only the parton-jet assignment problem but also the clustering problem, graph matching problem, and so does other problem which contains a permutaion symmetry, is a potential item to study with SPA-NET. We plan to explore the SPA-NET with some interesting BSM problem, such as the tri-Higgs production in 2HDM model.

Chapter 6

Conclusion

We perform a novel approach of parton-jet assignment using a symmetry preserving attention mechanism. This network could learn the permutation symmetry that appears in a physics process and archives a remarkable performance. Using this architecture can lead to a great improvement of performance and reduce the cost of computing time. And we shown that a well-trained deep learning network can replace the permutaion based algorithms and avoid the combinatorial explosion by estimating symmetry-aware pair-wise similarities.

Also, this novel architecture can not only provide an improvement to all hadronic top decay parton-jet assignment problem but also can be generalized to the other model which also contains the permutation symmetry.

Bibliography

[1] G. Aad *et al.* [ATLAS], "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Phys. Lett. B **716**, 1-29 (2012) doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[2] S. Chatrchyan *et al.* [CMS], "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC," Phys. Lett. B **716**, 30-61 (2012) doi:10.1016/j.physletb.2012.08.021 [arXiv:1207.7235 [hep-ex]].

[3] P. A. Zyla *et al.* [Particle Data Group], PTEP **2020**, no.8, 083C01 (2020) doi:10.1093/ptep/ptaa104

[4] A. Quade, "Top quark physics at hadron colliders," doi:10.1007/978-3-540-71060-8

[5] A. M. Sirunyan *et al.* [CMS], "Measurement of the top quark mass in the all-jets final state at $\sqrt{s} = 13$ TeV and combination with the lepton+jets channel," Eur. Phys. J. C **79**, no.4, 313 (2019) doi:10.1140/epjc/s10052-019-6788-2 [arXiv:1812.10534 [hep-ex]].

[6] M. Aaboud *et al.* [ATLAS], "Top-quark mass measurement in the all-hadronic $t\bar{t}$ decay channel at $\sqrt{s} = 8$ TeV with the ATLAS detector," JHEP **09**, 118 (2017) doi:10.1007/JHEP09(2017)118 [arXiv:1702.07546 [hep-ex]].

[7] T. Mccarthy, "Measurement of the Top Quark Mass in the All-Hadronic Top-Antitop Decay Channel Using Proton-Proton Collision Data from the ATLAS Experiment at a Centre-of-Mass Energy of 8 TeV," CERN-THESIS-2015-275.

[8] M. Paganini [ATLAS], J. Phys. Conf. Ser. **1085**, no.4, 042031 (2018) doi:10.1088/1742-6596/1085/4/042031 [arXiv:1711.08811 [hep-ex]].

[9] A. Vaswani *et al.*, "Attention is all you need," Advances in Neural Information Processing Systems, NIPS (2017)

[10] Hung-Yi, Lee, "NTU course lecture note-Transformer"