PROCEEDINGS OF SCIENCE

Top quark pair reconstruction using an attention-based neural network

Jason Sang Hun Lee,[a] **Inkyu Park,**[a] **Ian James Watson**[a] **and Seungjin Yang**[a,*]

[a] *Department of Physics, University of Seoul,*
Seoul 02504, Republic of Korea
E-mail: jlee@physics.uos.ac.kr, icpark@physics.uos.ac.kr,
ijwatson@physics.uos.ac.kr, seungjin.yang@physics.uos.ac.kr

For many top quark measurements, it is essential to reconstruct the top quark from its decay products. For example, the top quark pair production process in the all-jets final state has six jets initiated from daughter partons and additional jets from initial or final state radiation. Due to the many possible permutations, it is very hard to assign jets to partons. We use a deep neural network with an attention-based architecture together with a new objective function for the jet-parton assignment problem. Our novel deep learning model and the physics-inspired objective function enable jet-parton assignment using jet-wise input variables while the attention mechanism bypasses the combinatorial explosion that usually leads to intractable computational requirements. The model can also be applied as a classifier to reject the overwhelming QCD background, showing increased performance over standard classification methods.

40th International Conference on High Energy physics - ICHEP2020
July 28 - August 6, 2020
Prague, Czech Republic (virtual meeting)

*Speaker

PoS(ICHEP2020)348

1. Introduction

The measurements of top quark properties requires the reconstruction of the top quark from its decay products, leading to the problem of how to assign the reconstructed jets to outgoing partons from top quark decay. Existing jet-parton assignment studies used kinematic fitting or machine learning classifiers to estimate a goodness-of-association between underlying partons and a given jet permutation [1–5]. Since this evaluation is performed for jet combinations, they suffer from combinatorial explosion as the jet multiplicity increases. We introduce SaJa, which is a self-attention based neural network [6] for the jet-parton assignment free from requiring jet permutations. We test SaJa on fully-hadronically decaying $t\bar{t}$ production events. The full article can be found in [7] and the code for SaJa has also been made publicly available [8].

2. The SaJa Network

The jet-parton assignment problem can be translated into a jet-wise multi-class classification task. All jets in fully hadronically decaying $t\bar{t}$ production events can be divided into five categories: a b jet originating from the decay $t \to bW$, two light quark jets from the decay $W \to jj'$, a b jet and two W jets originating from $\bar{t}$ decays, and jets not associated to the top quark decays, which are referred to as other jets. In order to resolve the dificulty of distinguish jets produced from t and jets produced from $\bar{t}$, we introduce arbitrary indices 1 and 2 for the separation of t and $\bar{t}$ and their decay products.

Therefore, the model has the following form:

$$f^{\theta}: \begin{pmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(N)} \end{pmatrix} \to \begin{pmatrix} \hat{y}^{(1)}_{b_1} & \hat{y}^{(1)}_{W_1} & \hat{y}^{(1)}_{b_2} & \hat{y}^{(1)}_{W_2} & \hat{y}^{(2)}_{\text{other}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}^{(N)}_{b_1} & \hat{y}^{(N)}_{W_1} & \hat{y}^{(N)}_{b_2} & \hat{y}^{(N)}_{W_2} & \hat{y}^{(N)}_{\text{other}} \end{pmatrix} \tag{1}$$

where θ denotes the parameters of the model to be optimized, $\mathbf{x}^{(j)}$ indicates the jet in the event with index j and $\hat{y}^{(j)}_{\text{class}}$ indicates the score for jet j to be assigned to the category and the corresponding truth label will be denoted $y^{(j)}_{\text{class}}$ which is 1 if the jet has been truth-matched to the class and 0 otherwise. The assignments can be inconsistent with the topology of fully-hadronic $t\bar{t}$. For example, if there are two jets assigned to b_1 or only single jet assigned to W_2, such events are rejected and then are referred to as topologically invalid.

Since we want to find an optimal jet-wise classification model, it is a natural choice to use the average of the jet-wise cross entropy as the objective function. However, arbitrary indices 1 and 2 result in an ambiguity of choosing a permutation of t and $\bar{t}$ between indices 1 and 2. Therefore, we develop the objective function $J(\theta)$ as:

$$J(\theta) = \frac{1}{N}\sum_{j=1}^{N}\left(\min\left(\pi^{(j)}_{12}, \pi^{(j)}_{21}\right) + y^{(j)}_{\text{other}} \log \hat{y}^{(j)}_{\text{other}}\right) \tag{2}$$

where $\pi^{(j)}_{\alpha\beta} = y^{(j)}_{b} \log \hat{y}^{(j)}_{b_\alpha} + y^{(j)}_{\bar{b}} \log \hat{y}^{(j)}_{b_\beta} + y^{(j)}_{W^+} \log \hat{y}^{(j)}_{W_\alpha} + y^{(j)}_{W^-} \log \hat{y}^{(j)}_{W_\beta}$ with $\alpha, \beta \in \{1, 2\}$. As the min function has the permutation invariance property, Eq. 2 is free from the problem of whether

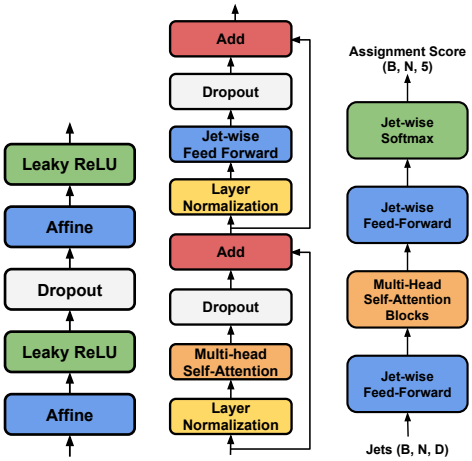

Figure 1: The directed acyclic graph (DAG) of the jet-wise feed-forward network (left), multi-head self-attention block (center), and SaJa (right) are shown. B indicates the batch size. N indicates the maximum jet multiplicity in the batch. D indicates the number of features representing the jet.

to assign t or $\bar{t}$ to the index 1 or the index 2. Note that eq. 1 did not constraint the form of model much, thus armed with eq. 2, any neural network architecture, such as convolutional networks or graph neural networks, can be zero-permutation jet-parton assignment networks at inference time.

Figure 1 shows SaJa and its two building blocks, which are the jet-wise feed-forward network and the multi-head self-attention block. SaJa features the scaled dot-product self-attention, which takes three sets of vectors as input and output a single set of vectors and can capture the underlying patterns of input data. The mechanism of the scaled dot-product self-attention is well documented in the orignal paper [6] and our full article [7].

We will test the predictive entropy, which is a kind of uncertainty for classification models [9]. As wrong predictions tend to have high uncertainty, predictive entropy can be used to veto wrong assignments, which degrade the resolution of reconstructed top quark kinematics. Also, the uncertainty can play a role of out-of-distribution (OOD) test sample detection method. In this study, QCD multijet events are exactly OOD. We found that the predictive entropy enabled SaJa to reject QCD events without additional training on QCD events.

Since SaJa simultaneously predicts on all jets in the event, we use the average of the jet-wise predictive entropy, $\mathbb{H}[\hat{Y}] = \frac{1}{N}\sum_{j=1}^{N}\left(-\sum_{c\in\text{classes}}\hat{y}_c^{(j)}\log\hat{y}_c^{(j)}\right)$. When the predictive entropy is higher than a threshold, the event is not selected.

3. Monte Carlo Samples and Event Selection

We use MG5_aMC@NLO v2.2.2 [10] interfaced to Pythia8.212 [11] to produce fully-hadronic $t\bar{t}$ pair production with up to two additional jets at next to leading order and multijet events at leading order in the final state from proton-proton collisions at $\sqrt{s} = 13$ TeV. For the event generation, the top quark mass is set to 172.5 GeV. We use Delphes v3.4.2 [12] to simulate the response of CMS-like detector. The default Delphes CMS card was used except that we perform anti-k_T jet clustering with the parameter R of 0.4, instead of 0.5, using FastJet v3.3.2 [13].

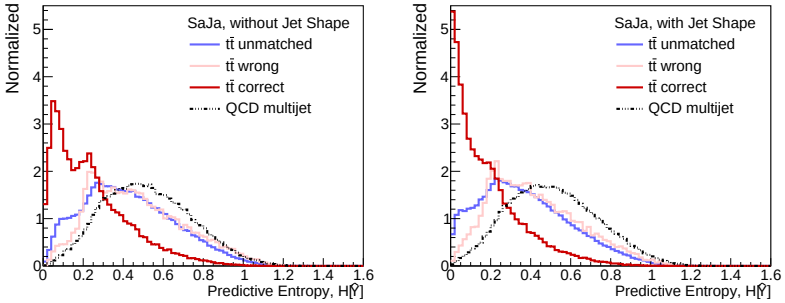

Figure 2: The prediction entropy distribution of SaJa without jet shape (left) and SaJa with jet shape (right).

We use the trigger selection used in the CMS fully-hadronic $t\bar{t}$ analysis for the event selection [1]. Jets are required to have $p_T > 30$ GeV and $|\eta| < 2.4$. We select events with at least 6 jets with $p_T > 40$ GeV, at least one of which is b-tagged and then require $H_T = \sum_{\text{jets}} p_T > 450$ GeV.

After the event selection, we perform geometric matching between jets and the six partons from fully hadronic $t\bar{t}$. Top pair events, where all partons are matched is called *matched* events. The fraction of matched events is about 20%. The *unmatched* $t\bar{t}$ events are considered as background to study the performance of SaJa

We use all jets in the event as input to SaJa, where the jet is represented using reconstructed variables: p_T, η, ϕ, $\frac{p_T}{H_T}$, and whether the jet is b-tagged. Also, an additional eight jet shape variables are tested based on the idea that gluon jets should be assigned to the *other* class in the fully hadronic $t\bar{t}$ topology. We apply Min-Max normalization to scale all features into the range in $[0, 1]$ in order to make the training converge faster.

We trained SaJa by minimizing the objective function in eq. 2 using the Adam optimization algorithm with an initial learning rate of 0.001 while decaying the learning rate by 2 when the validation loss stopped decreasing over 10 epochs. A batch size of 512 is used during the training.

We also used KLFitter [3] for the kinematic likelihood fitting in order to compare the performance of SaJa. We only studied two cases because of combinatorial explosion. The first is the most energetic 6 jets, resulting in 18 permutations on average. The second is up to 7 most energetic jets, giving 126 permutations on average. We chose the permutation with the highest likelihood as the jet-parton assignment and reject the event if the likelihood is lower than a veto threshold.

4. Results

The predictive entropy distributions of SaJa with and without additional jet shape variables are shown in figure 2. Correctly assigned $t\bar{t}$ events have lower uncertainty compared to wrongly assigned $t\bar{t}$, unmatched $t\bar{t}$, or multijet events. We observed that the multijet distribution shows more bell-shaped and has a peak at a larger entropy value. These two facts mean that the predictive entropy of SaJa is well calibrated and so is effective in rejecting wrongly assigned signal events and background events. The ability to detect an out-of-distribution events could be used for a model-independent new physics search, by training the model on the expected Standard Model processes.

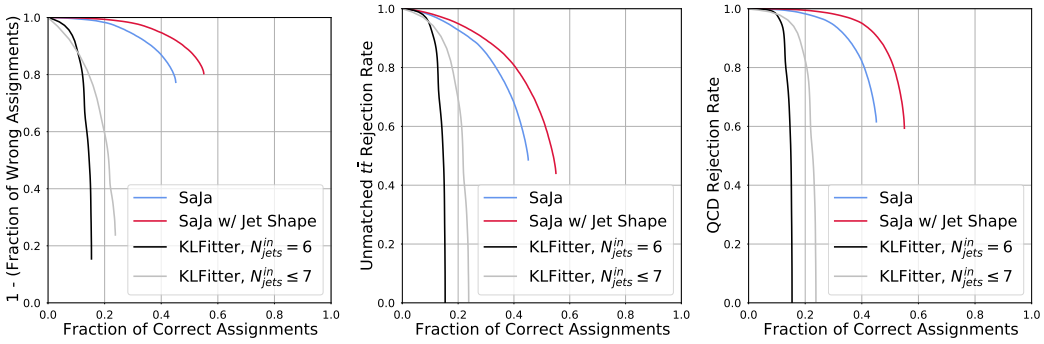

Figure 3: The performance measurement curve. 1 - the fraction of wrong assignments for matched $t\bar{t}$ (left), the unmatched $t\bar{t}$ rejection rate (middle) and multijet rejection rate (right) as a function of the fraction of correct assignments for matched $t\bar{t}$. The curves of SaJa without jet shape (blue) and with jet shape (red) are cut short due to the rejection of topologically invalid events.

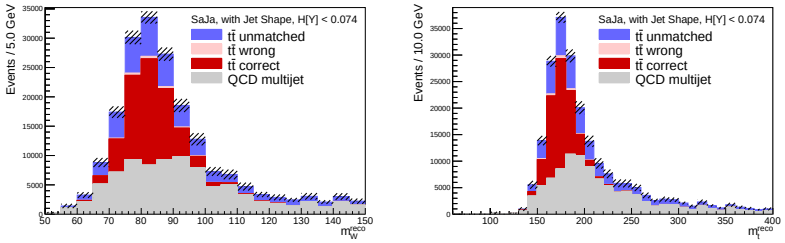

Figure 4: The reconstructed mass distribution of W boson (left) and top quark (right). unmatched $t\bar{t}$ (blue), wrongly assigned $t\bar{t}$ (pink), correctly assigned $t\bar{t}$, and multijet (gray).

Figure 3 shows the performance of jet-parton assignment methods in the manner of the receiver operating characteristic (ROC) curve, where the fraction of wrongly assigned matched $t\bar{t}$ and the rejection rate for background events are shown as a function of the fraction of correctly assigned matched $t\bar{t}$. The curves are drawn by varying the threshold value of the predictive entropy for SaJa or the negative log-likelihood for KLFitter. The higher the curve toward the upper right, the more powerful jet-parton assignment performance and figures shows that SaJa exceeds KLFitter. The curves of SaJa are cut short because the topologically invalid assignments are rejected.

Figure 4 shows the reconstructed W and top mass distribution, which are obtained using SaJa with jet shape information and a predictive entropy threshold of 0.074. The total integrated luminosity of 35.91 fb^{-1} is used for the normalization. Clear peaks are observed in the W and top mass range.

5. Conclusion

In these proceedings, we introduced the SaJa network, which uses the self-attention to solve the jet-parton assignment problem without requiring jet permutations. We also introduced a new objective function to train the SaJa network on jet-parton assignment task for fully-hadronic top pair events. SaJa achieved better assignment performance and faster inference speed compared to the traditional kinematic likelihood fitting method, as implemented in KLFitter. As the SaJa network

is easily extended to more complex topologies where previous methods were computationally infeasible, such as ttH, it has great potential for future use.

References

[1] CMS Collaboration, *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** (2019) 313, [arXiv:1812.10534].

[2] ATLAS Collaboration, *Measurement of the top-quark mass in the fully hadronic decay channel from ATLAS data at* $\sqrt{s}$ *= 7 TeV*, *Eur. Phys. J.* **C 75**, 158 (2015), [arXiv:1409.0832]

[3] J. Erdmann, S. Guindon, K. Kroeninger, B. Lemmer, O. Nackenhorst, A. Quadt and P. Stolte, *A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework*, *Nucl.Instrum.Meth.* **748** *(2014) 18* , [arXiv:1312.5595].

[4] M. Erdmann, B. Fischer and M. Rieger, *Jet-parton assignment in ttH events using deep learning*, *JINST* **12** (2017) P08020, [arXiv:1706.01117].

[5] J. Erdmann, T. Kallage, K. Kröninger and O. Nackenhorst, *From the Bottom to the Top-Reconstruction of* $t\bar{t}$ *Events with Deep Learning*, *JINST* **14** (2019) P11015, [arXiv:1907.11181]

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, in the proceedings of *Neural Information Processing Systems* (NIPS 2017), December 4-9, Long Beach, U.S.A. (2017), [arXiv:1706.03762].

[7] , J. S. H. Lee, I. Park., I. J. Watson, and S. Yang, *Zero-Permutation Jet-Parton Assignment using a Self-Attention Network*, [arXiv:2012.03542]

[8] S. Yang, I. J. Watson, J. S. H. Lee, and I. Park. *(2020, December 8). CPLUOS/SaJa: v1.* Zenodo. doi:10.5281/zenodo.4311381

[9] Y. Gal, *Uncertainty in deep learning*, University of Cambridge (2016).

[10] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, *MadGraph 5 : Going Beyond*, *JHEP* **06** (2011) 128, [arXiv:1106.0522].

[11] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012]

[12] The DELPHES 3 collaboration, *DELPHES 3: a modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346].

[13] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J.* **C 72** (2012) 1896, [arXiv:1111.6097].

[14] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization* [arXiv:1412.6980].