

# Top quark pair reconstruction using an attention-based neural network

Zero-Permutation Jet-Parton Assignment

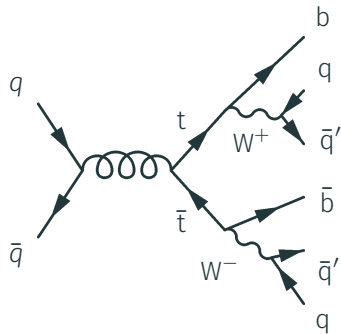
---

Jason Sang Hun Lee, Inkyu Park, Ian James Watson, and Seungjin Yang  
University of Seoul

ICHEP 2020,  
July 31, 2020

# Introduction

- The reconstruction of events with intermediate states decaying to jets requires a technique to assign jets to partons.
- The number of jets can be greater than the number of partons because of additional QCD radiation. It makes the assignment harder.
- We introduce **the self-attention for jet assignment (SAJA) network without requiring jet permutations**. We apply SAJA to find jet-parton assignments of fully-hadronic  $t\bar{t}$  events to test the performance.



## ■ $\chi^2$ method [CMS, Eur. Phys. J. C 79 (2019) 313]

$$\chi^2 = \sum_{j \in \text{jets}} \left[ \frac{(p_{T,j}^{\text{reco}} - p_{T,j}^{\text{fit}})^2}{\sigma_{p_{T,j}}^2} - \frac{(\eta_j^{\text{reco}} - \eta_j^{\text{fit}})^2}{\sigma_{\eta_j}^2} - \frac{(\phi_j^{\text{reco}} - \phi_j^{\text{fit}})^2}{\sigma_{\eta_j}^2} \right]$$

## ■ Kinematic likelihood fitting [J. Erdmann, Nucl.Instrum.Meth.A 748 (2014) 18-25]

$$\mathcal{L} = B(m_{q_1 q_2 q_3} | m_t, \Gamma_t) \cdot B(m_{q_1 q_2} | m_W, \Gamma_W) \cdot B(m_{q_4 q_5 q_6} | m_t, \Gamma_t) \cdot B(m_{q_4 q_5} | m_W, \Gamma_W) \cdot \prod_{i=1}^6 W_{\text{jet}}(E_{\text{jet},i}^{\text{meas}} | E_{\text{jet},i})$$

## ■ Machine Learning [M. Erdmann, JINST 12 (2017) P08020], [J. Erdmann, JINST 14 (2019) P11015]

$$\text{correct or wrong} = \text{model}(\text{assignment})$$

All of the above methods follow the same steps.

- 1 Compute all (promising) jet permutations.
- 2 Evaluate how well each permutation agrees with the underlying event topology. ( $\chi^2$ ,  $\mathcal{L}$ , the sigmoid output of NN, BDT score)
- 3 Choose the best permutation.

⇒ Combinatorial explosion  $O(n!)$

Jet-parton assignment as jet-wise classification.

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

- $f^\theta$  is DL model with weights  $\theta$ .
- $x^{(i)}$  denotes the reconstructed jet variables and there is no order between jets.
- Since it is hard to distinguish jets originating from  $t$  and jets originating from  $\bar{t}$ , arbitrary indices 1 and 2 are introduced.
- 'other' indicates additional jets produced in the event.
- When the assignment doesn't agree with the underlying topology (e.g. no  $b_1$  or 3  $W_2$ ), the assignment is said to be topologically invalid and is not selected.

# Objective function

Since the indices 1 and 2 are arbitrary, the general objective function cannot be used here.

Therefore, a new cross-entropy based objective function is introduced.

$$J(\theta) = \frac{1}{N} \sum_{j=1}^N [\min(\pi_{12}^{(j)}, \pi_{21}^{(j)}) + y_{\text{other}}^{(j)} \log \hat{y}_{\text{other}}^{(j)}],$$

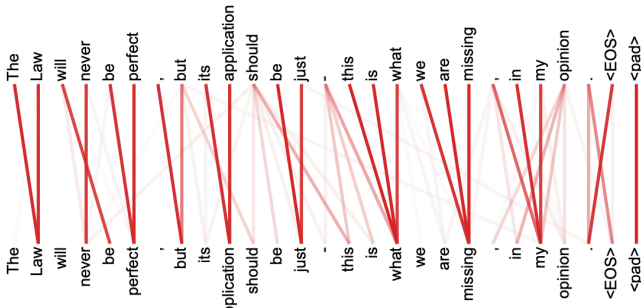
where

$$\pi_{\alpha\beta}^{(j)} = y_b^{(j)} \log \hat{y}_{b\alpha}^{(j)} + y_{\bar{b}}^{(j)} \log \hat{y}_{\bar{b}\beta}^{(j)} + y_{W^+}^{(j)} \log \hat{y}_{W^+\alpha}^{(j)} + y_{W^-}^{(j)} \log \hat{y}_{W^-\beta}^{(j)}.$$

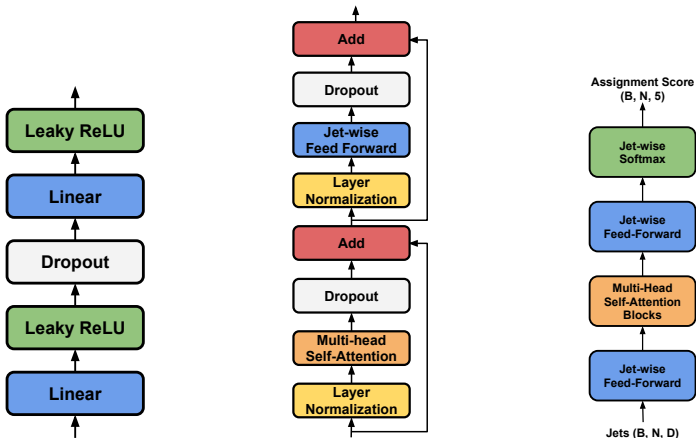
⇒ A deep learning model of any architecture can become a zero-permutation jet-parton assignment model if it is fit by minimizing the above objective function.

# Self-Attention based Model Architecture

- The model implementation is based on TRANSFORMER, which is the neural machine translation model and features self-attention. [A. Vaswani, arXiv:1706.03762]
- Self-attention is a weight sum of the elements of the input set, where the weight matrix is also computed from the elements.
- Self-attention based model can learn the dependency between elements.



# Self-Attention based Model Architecture (SAJA)



(a) Jet-wise feed-forward network

(b) Self-attention block

(c) Jet-parton assignment model

Figure 1:  $B$  denotes the batch size.  $N$  indicates the maximum number of jets in the batch.  $D$  indicates the number of features representing the jet.

## ■ Generation

- pp collision with  $\sqrt{s} = 13$  TeV using MADGRAPH5\_AMC@NLO and PYTHIA8
- Fully hadronic  $t\bar{t}$  with up to two additional jets at NLO,  $m_t = 172.5$  GeV
- QCD multijet at LO.

## ■ Detector response

- DELPHES3
- CMS-like detector

## ■ Jet Finding

- FASTJET3
- anti- $k_T$  algorithm with  $R=0.4$



Selection follows the trigger selection used in the CMS  $t\bar{t}$  all-jets analysis. [CMS, Eur. Phys. J. C 79 (2019) 313].

## ■ Jet

- $p_T > 30$  GeV
- $|\eta| < 2.4$
- + b-tag: true/fake based on the efficiency for b quark and misidentification rates for gluon, light quark jets and c-jets.

## ■ Event

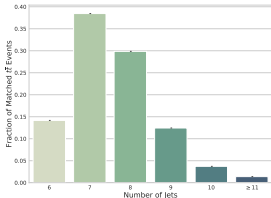
- $N_{\text{jet}} \geq 6$
- At least one b-tagged jet of the six most energetic jets
- $p_T(\text{jet}_6) > 40$  GeV
- $H_T \equiv \sum_{\text{jet}} p_T > 450$  GeV

# Jet-Parton Matching

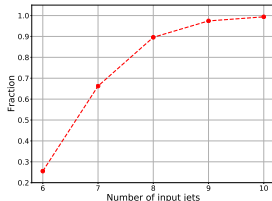
After the event selection, only about 20% of  $t\bar{t}$  events satisfy the following jet-parton matching condition without ambiguity and are called **matched**.

$$\Delta R(\text{jet}, \text{parton}) = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$$

Only the matched  $t\bar{t}$  events are used to train the DL model.



**Figure 2:** The distribution of the number of jets in the matched  $t\bar{t}$  events



**Figure 3:** The fraction of matched  $t\bar{t}$  events, where all partons can be matched with the most energetic  $N$  jets.

## ■ Event

All jets in the event are used as input to the model.

For simplicity, the jet is high level reconstructed variables.

$$(p_T, \eta, \phi, \frac{p_T}{H_T}, \text{b-tag})$$

## ■ Jet Shape

Gluon-initiated jets should always be assigned to 'other'. So the effect of the following jet shape variables is studied. [CMS, arXiv:1409.3072]

- $p_{TD} = \frac{\sum_i p_{T,i}^2}{\sum_i p_{T,i}}$
- Major and minor axes of the jet axis in  $\eta - \phi$  space
- Charged hadron, neutral hadron, electrons, muon, and photon multiplicities

## ■ Pre-processing

All features except b-tag is scaled to [0, 1] through min-max scaling  $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

# Predictive Entropy

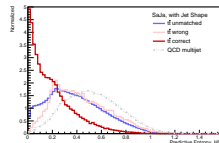
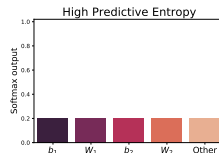
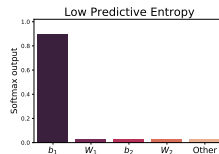
NB

In this slides, the uncertainty is ML-side terminology.

- To suppress poor jet-parton assignments, the DL model uncertainty is studied.
- The **predictive entropy** quantifies the uncertainty in the prediction of the classifier.

$$H[\hat{Y}] = \frac{1}{N} \sum_{j=1}^N \left[ - \sum_{C \in \text{partons}} \hat{y}_C^{(j)} \log \hat{y}_C^{(j)} \right]$$

- When the jet-parton assignment with the predictive entropy higher than the threshold, the event is not selected.
- + The uncertainty is also used to detect out-of-distribution test data (corresponding to QCD in this study).



# Benchmark: Kinematic Likelihood Fitting

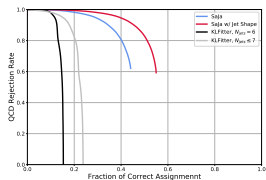
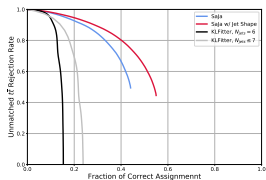
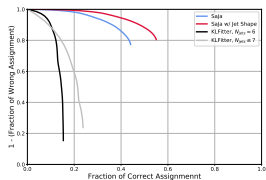
- Kinematic likelihood fitting is performed as a baseline study, the `KLITTER` library is used. [J. Erdmann, Nucl.Instrum.Meth.A 748 (2014) 18-25]

$$\mathcal{L} = B(m_{q_1q_2q_3}|m_t, \Gamma_t) \cdot B(m_{q_1q_2}|m_W, \Gamma_W) \cdot B(m_{q_4q_5q_6}|m_t, \Gamma_t) \cdot B(m_{q_4q_5}|m_W, \Gamma_W) \cdot \prod_{i=1}^6 W_{\text{jet}}(E_{\text{jet},i}^{\text{meas}}|E_{\text{jet},i}),$$

where  $B$  indicates the Breit-Wigner distribution and  $W$  does the the transfer functions

- As more jets are used, the number of permutations increases explosively. Therefore, two case are studied.
  - Most energetic 6 jets,  $N_{\text{jets}} = 6 \rightarrow$  Avg. 18 permutations
  - Up to 7 most energetic jets  $N_{\text{jets}} \leq 7 \rightarrow$  Avg. 126 permutations
- When the best permutation has a likelihood lower than the threshold, the event is not selected.

- The assignment performance is visualized as ROC-like curves drawn by varying the threshold value for the predictive entropy of SAJA or the likelihood of KLFITTER.
- SAJA shows more powerful performance than KLFITTER.
- Predictive entropy not only reduces poor jet-parton assignments but also helps reduce unmatched  $t\bar{t}$  and QCD multijet events without additional training process.
- Jet shape increases the fraction of correct assignment.



# Reconstructed W Boson and Top Quark Mass Distributions

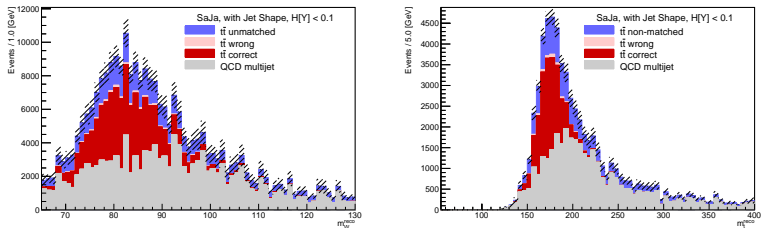
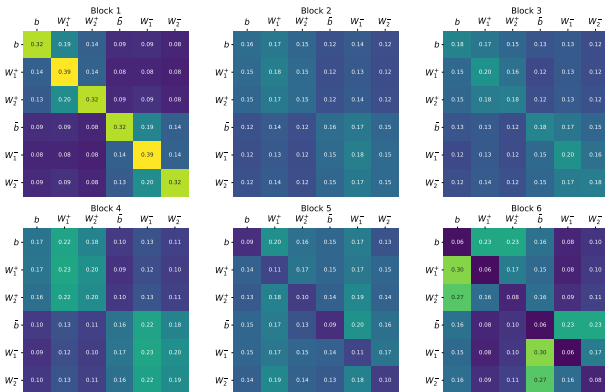


Figure 4: (Left) Reconstructed W boson mass distribution and (Right) reconstructed top quark mass distribution.

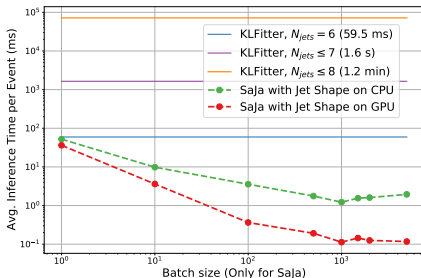
# Model Interpretability



- The average of attention matrices for the correct assignments in the matched  $t\bar{t}$  events.
- Since SAJA is invariant under the permutation of the jets, it is possible to sort the jets using the jet-parton matching information w.l.o.g, for display purposes.
- $W_1^+$  indicates higher  $p_T$  one of the two quarks, the decay products of  $W^+$  boson.



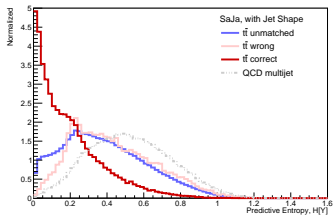
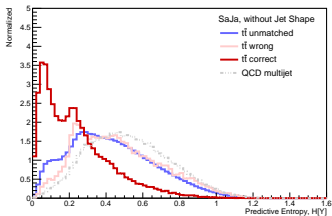
# Inference Latency



- The graph shows that even for a batch size of one, the zero-permutation SAJA network is outperforming KLFITTER, while a two-order-of-magnitude speed up in inference time is possible by fitting in large batches.

- We introduce SAJA network for jet-parton assignment without jet permutations.
- SAJA shows better jet-parton assignment and background rejection performance compared to KLFITTER.
- Predictive entropy makes it possible to improve performance without additional training.
- SAJA is fast even without GPU acceleration.

# Predictive entropy distribution



# Effect of Jet Shape

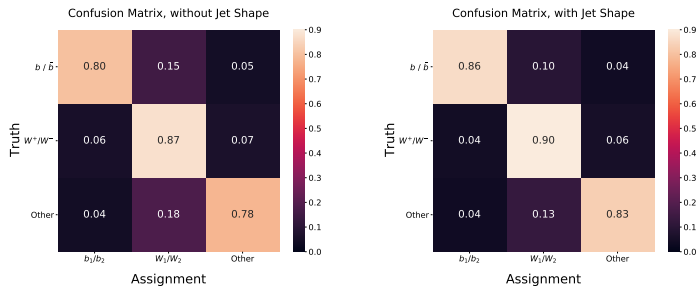
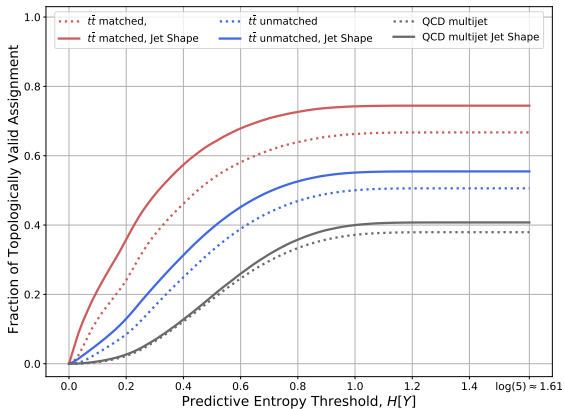


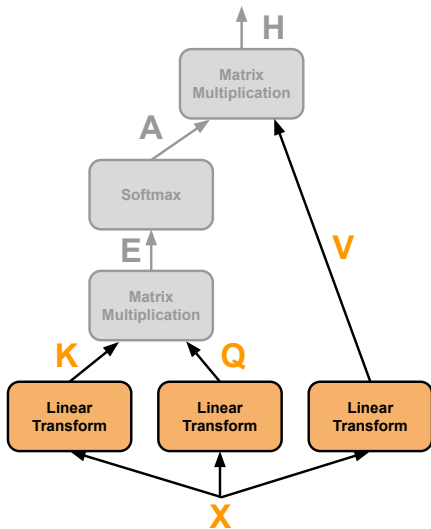
Figure 5: Confusion matrices at the jet-level with an without jet shape.

Jet shape reduces the case where b-initiated jets and additional jets are assigned to W boson.

# Fraction of Topologically Valid Assignments



# Dot-Product Attention (1)

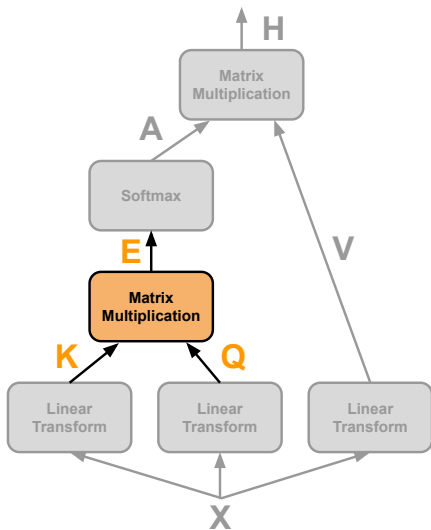


$$\vec{q}_i = W_Q \vec{x}_i \rightarrow Q = \{\vec{q}_1, \dots, \vec{q}_N\},$$

$$\vec{k}_i = W_K \vec{x}_i \rightarrow K = \{\vec{k}_1, \dots, \vec{k}_N\},$$

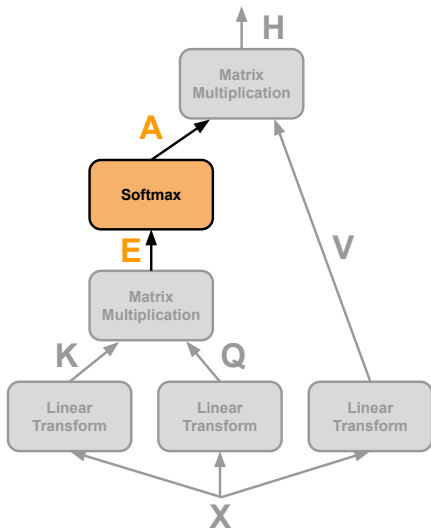
$$\vec{v}_i = W_V \vec{x}_i \rightarrow V = \{\vec{v}_1, \dots, \vec{v}_N\}.$$

## Dot-Product Attention (2)



$$\begin{aligned} E &= QK^T \\ &= \begin{bmatrix} \vec{q}_1 \\ \vec{q}_2 \\ \vec{q}_3 \end{bmatrix} \begin{bmatrix} \vec{k}_1 & \vec{k}_2 & \vec{k}_3 \end{bmatrix} \\ &= \begin{bmatrix} \vec{q}_1 \cdot \vec{k}_1 & \vec{q}_1 \cdot \vec{k}_2 & \vec{q}_1 \cdot \vec{k}_3 \\ \vec{q}_2 \cdot \vec{k}_1 & \vec{q}_2 \cdot \vec{k}_2 & \vec{q}_2 \cdot \vec{k}_3 \\ \vec{q}_3 \cdot \vec{k}_1 & \vec{q}_3 \cdot \vec{k}_2 & \vec{q}_3 \cdot \vec{k}_3 \end{bmatrix} \end{aligned}$$

## Dot-Product Attention (3)

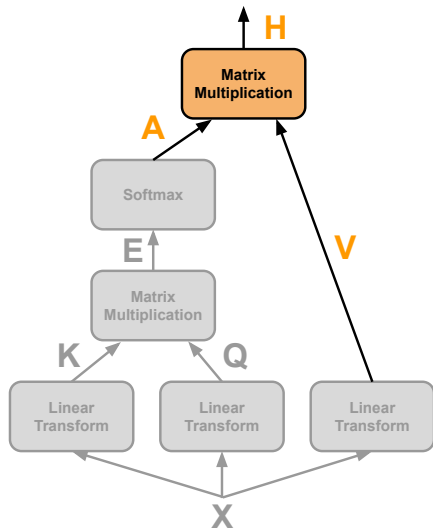


$$A = \begin{bmatrix} \frac{e^{E_{11}}}{Z_1} & \frac{e^{E_{12}}}{Z_1} & \frac{e^{E_{13}}}{Z_1} \\ \frac{e^{E_{21}}}{Z_2} & \frac{e^{E_{22}}}{Z_2} & \frac{e^{E_{23}}}{Z_2} \\ \frac{e^{E_{31}}}{Z_3} & \frac{e^{E_{32}}}{Z_3} & \frac{e^{E_{33}}}{Z_3} \end{bmatrix},$$

where  $Z_j = \sum_i e^{E_{ij}}$ .



## Dot-Product Attention (4)



$$H = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \vec{v}_1 \\ \vec{v}_2 \\ \vec{v}_3 \end{bmatrix}$$
$$= \begin{bmatrix} \sum_i A_{1i} \vec{v}_i \\ \sum_i A_{2i} \vec{v}_i \\ \sum_i A_{3i} \vec{v}_i \end{bmatrix} \cdot$$

# Training in detail

- Training, validation, test set: 300k, 80k, 100k events
- Adam Optimization algorithm
- Learning rate schedule, which reduces the learning rate when the metric on the validation set has stopped improving.
- PYTORCH v1.3