Multi-scale Cross-Attention Transformer encoder for event classification

Mihoko Nojiri(IPNS, KEK), with Ahmed Hammad  Stefano Moretti
arXiv 2401.00452
ABOUT MYSELF

• PhD Kyoto (1990) a bit old.


• Collider:

  • 1996: JLC study and Snowmass

  • 2002-2008 LHC BSM study in ATLAS SUSY group. BSM Convener of Les Houches TeV collider workshop twice → Jet substructure study → Deep Learning

• Service: JPS executive board member → member of Science Council of Japan (SCJ) working on Diversity Issues.

• In KEK, we just had DEI workshop last Dec, and trying establish more DEI activities. (https://www2.kek.jp/ipns/en/news/5320/)

“a young mind”, (according to Tilman Plehn) but this makes me cry
The research area "Machine Learning Physics" will begin with the aim of discovering new laws and pioneering new materials.
How machine Leaning help Collider Analysis

Jet clustering

CMS experiment at LHC, CERN
data recorded: Tuesday Sept 27 10:30:59 2016 EDT
run/event/LS: 281707/1308250303/826
Jet classification using ML

QCD jet

W jet

as sets

Particles

Observable

as graphs

permutation invariance

(Energy Flow Network and Particle Flow Network 1810.05165)

sparse data

1902.08570 Particle Net
Dreyer et al LundNet (1807.04758)
Gong et al LorentzNet (2201.08187)

building key and query
2202.03772

CNN Oliverira et al (1511.05190)

from Schwartzman et al

Bogatskiy et al PELICAN (2211.00454)
CONNECTING JET STRUCTURE INFORMATION TO EVENT KINEMATICS

- Non SM Higgs boson (Two Higgs doublet model)
  - \( pp \rightarrow H \) (Heavy Higgs boson) \( \rightarrow hh \) \( \rightarrow 4 \) bjet
  - \( mH=600-2000 \) GeV, \( mh=125.11 \) GeV
- Meta stable vacuum of SM \( \rightarrow \) extension of Higgs sector
- why doing Deep Learning?
  - Sensitivity under S/BG~1 scale by \( 1/\sqrt{N} \) with background rejection \( 1/N \)

Figure 1: Structure of the transformer model used. Here, \( P_{j1}, P_{j2} \) are the number of the leading and second leading jet constituents while the \( P_m \)'s are the reconstructed particles, \( j_1, j_2, \) and \( H \). Also, MHSA stands for multi-heads self-attention layers, and MHCA stands for multi-heads cross-attention layers. Finally, the \( N_i \)'s are the number of the used transformer encoders. The transformer layers are stacked and work sequentially, as pointed out by the black arrow.

Figure 2: Feynman diagram for the signal process.
HOW SIGNAL LOOKS LIKE: KINEMATICAL INPUTS

- Delphes $pp \to H \to hh$ vs $pp \to 4b$, $pp \to tt$
- Delphes Preselection
  - two fatjets (radius $R=1.0$) $p_T$ cut on the fatjet $P_T^{1} > 450\text{GeV}$ $P_T^{2} > 250\text{GeV}$.
  - double b tags for each fatjet (Delphes 80% tagging efficiency) $250\text{GeV} > M(J) > 50\text{GeV}$
  - no pileup (theorist job)

inspired by ATLAS study’’

Kinematical inputs (3, 6)
fatjet 1 = (m_1, \eta_1, \phi_1, p_T^1, E_1), \theta_1
fatjet 2 = (m_2, \eta_2, \phi_2, p_T^2, E_2), \theta_2
H candidate = (m_{12}, \eta_{12}, \phi_{12}, p_T^{12}, E_{12}), \theta_{12} = 0

NOTE: "5 inputs for 4 momentum", H candidate momentum as sum of two fat jets, add \theta,
INPUT TO NETWORK: JET SUBSTRUCTURE INFO AS PARTICLE CLOUD

color singlet

beam direction

jet 1

jet 2

m_{j} \sim 125 \text{GeV}

\eta

\phi

up to 50 constituents:

Regularization speed up the training and reduce the required events.

1. shift coordinate to (0,0)
2. rotate jet based on covariant matrix
3. flip \eta so that \( E(\bar{\eta} > 0) > E(\bar{\eta} < 0) \)
4. particles are ordered by pT and we take up to 50

\[ p_i = (\bar{\eta}_i, \bar{\phi}_i, p_{T_i}, \log p_{T_i}) \rightarrow (50, 4) \text{ data} \]
HOW TO COMBINE JET STRUCTURE AND EVENT KINEMATICS

Naive approach "simple concatenation"

![Diagram of neural network structures](image)

**FIG. 2.** The schematic plots for neural network structures: (a) conventionally used one in previous studies only with concatenation and (b) our proposed one with a regularized attention mechanism.

(a) [Jet momentum (parton momentum)] + [jet concatenation] does not work. Because of imbalance of "importance" of two information → the minor one can be ignored in the training.

Pre-training and freeze substructure analysis? We would loose the correlation to global kinematics.
**OUR CROSS ATTENTION MODEL**

**Physics Example**

Leading Jet

Sub-leading Jet

Data Pre-processing:

- **Centering:** Jet contents are shifted such that the jet axis is in the center.
- **Rotation:** Rotate the jet contents in the eta-phi plane such that the jet axis is vertical.
- **Flipping:** Jet contents are reflected over the vertical axis such that the right side contains the hard radiations.

After preprocessing, the Higgs jet exhibits a two prong structure while QCD multi-jets process shows a wide range of radiation. The background consists of 90% of QCD and 10% of $t\bar{t}$.

**Multihead Cross Attention Layers**

**Multihead Self Attention Layers**

**Step 1:** Self attention

$$[\text{structure}] \times [\text{structure}] \times [\text{jet kin}]$$

**Step 2:** Cross attention

transform jet kin by cross Att. $[\text{structure}] \times [\text{jet kin}]$
TAKEAWAYS

- use "cross attention" when you combine the “high scale information” to the “low energy scale”, because cross attention layer gives extra emphasis to the information linked to the high energy kinematics.

- skip connection and Interpretation: Skip connection helps to maintain some connection to the inputs

- More Physics: Heavy particles decay into colored particles (discovery, spin, color structure?) Cross attention network probably more useful to resolve correlation of jet structures.
STEP 1 SELF ATTENTION LAYERS

output size = input size

Attention matrix mix all features. Higher attention elements indicates important correlations

transformation $V \rightarrow V'$ does not change the dimension. Structure of $V$ retained for the next transformation.

We adopt 50x50 self attention for jet and 3x3 self attention for kinematics, with $n_{head} = 5$

$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

ATTENTION Matrix mix all features. Higher attention elements indicates important correlations.

transformation $V \rightarrow V'$ does not change the dimension. Structure of $V$ retained for the next transformation.

We adopt 50x50 self attention for jet and 3x3 self attention for kinematics, with $n_{head} = 5$.
STEP 2 CROSS ATTENTION LAYERS

- Choose cross attention (jet kin) x (jet str.)
- Jet momentum: hard physics of partons $Q, V$
- Jet substructure: parton shower, hadronization $K$
- Substructure output $K$ and Jet kinematics output $Q$ make attention matrix. The pairs update $V$ (jet Kin)
- High scale feature relevant for classification gives extra weight to the corresponding jets through backward propagation
Comparision with other approach

Naive approach "simple concatenation"


(a) Conventionally used one in previous studies only with concatenation and (b) our posed one with a regularized attention mechanism.

(b) self attention matrix of combined information

\[
A \ V = \begin{pmatrix}
Q(\text{Sub}) \times K(\text{Sub}) & Q(\text{Kin}) \times K(\text{Sub}) \\
Q(\text{Sub}) \times K(\text{Kin}) & Q(\text{Kin}) \times K(\text{Kin})
\end{pmatrix}
\]

our network kill this term and keep off diagonal part only

\[
V = Q(\text{kin}) \ K(\text{kin}) \ V(\text{kin}) + \ldots
\]
PHYSICS

• a jet:

\[ P(\text{hadrons in jets} \mid \text{parton or jet}) = P(\{x_i\} \mid y) \]

• a fatjet or a jet with substructure

\[ P(\{x_i\} \mid \{y_\alpha\}) \]

• two fatjets in an event

\[ P(\{x_i\}, \{x'_j\}, \{y_\alpha\}, \{y'_\beta\}) \sim P(\{x_i\} \mid \{y_\alpha\}) P(\{x'_j\} \mid \{y'_\beta\}) P(\{y_\alpha\}, \{y'_\beta\}) \]

\[ P(\{x_i\}, \{x'_j\}, \{y_\alpha, y'_\beta\}) \sim P(\{x_i\} \mid \{y_\alpha, y'_\beta\}) P(\{x'_j\} \mid \{y_\alpha, y'_\beta\}) P(\{y_\alpha, y'_\beta\}) \]

cross attention  jet kinematics
fluenced by the random partitioning of the training and test data sets, and the network limited training and testing samples. For example, the network performance can be in-

factor 5 improvement at the same acceptance.

AUC (Jets only (self-attention))= 84.4%
AUC (Kinematics only (self-attention))= 91.6%
AUC (Jets+Kinematics (self-attention))= 95.0%
AUC (Jets+Kinematics (cross-attention))= 98.8%

Cross attention improve the rejection efficiency significantly

Simple estimation of the upper limits

95\% C.L. limit $\sigma(gg \rightarrow H \rightarrow h_h)$, $\mathcal{L} = 3000$ fb$^{-1}$
INPUT TO DL: EVENT KINEMATICS

Kinematical inputs (3, 6)

fatjet 1 = \( (m_1, \eta_1, \phi_1, p_T_1, E_1), \theta_1 \)
fatjet 2 = \( (m_2, \eta_2, \phi_2, p_T_2, E_2), \theta_2 \)

H candidate = \( (m_{12}, \eta_{12}, \phi_{12}, p_{T12}, E_{12}), \theta_{12} = 0 \)

NOTE: "5 inputs for 4 momentum", H candidate momentum as sum of two fat jets, add \( \theta \).
Conversely, when the information of the jet constituents is included using the cross-attention layer, the attention output distributions for background events are broader, and the signal distributions are narrower. The fact that background jets lack a multi-prong structure with broader soft radiations influences the attention output for background events, increasing the output variations in the feature space.

Finally, we include, alongside the described kinematical information, also the rotation angle \( \theta \) aligning the fat jet axis to the "direction after shifting the jet and" to the center of the "plane. This information allows the network to reconstruct the full events and access the correlation of the jet shape to the other fat jet and the beam axis. In Fig. 8, we show the ROC curve of the network trained without the \( \theta \) inputs (red) compared to the ROC curve of our cross-attention model (blue). The improvement on the background rejection is a factor of four for a signal efficiency of 80%. Therefore, including \( \theta \) results in a drastically increased performance. The model with \( \theta \) has higher efficiency at \( m_J \geq m_h \) and \( p_T \geq m_H/2 \). In short, the model can focus more on the momentum with \( \theta \) inputs.

We also looked for simple correlations among \( \theta \) and the other kinematical variables, such as \( \div J \), but did not find any apparent ones contributing to the selection improvement. (The correlations within the internal structures of the jet will be investigated in future publications.)

<table>
<thead>
<tr>
<th></th>
<th>Kinematics</th>
<th>Kin + ( \theta )</th>
<th>jet str.+kin</th>
<th>jet str +Kin + ( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>91.01%</td>
<td>91.6</td>
<td>97.23-98.16</td>
<td>98.68-99.28</td>
</tr>
</tbody>
</table>

adding rotation angle \( \theta \) improve classification when both jet str. and kinematical information available.

We are working on to identify the origin. (color connection? momentum resolution? )
For QCD and top event, fatjets are likely color connected to the other activities of the event.

Higgs bosons are color isolated.
Conversely, when the information of the jet constituents is included using the cross-attention layer, the attention output distributions for background events are broader, and the signal distributions are narrower. The fact that background jets lack a multi-prong structure with broader soft radiations influences the attention output for background events, increasing the output variations in the feature space. Finally, we include, alongside the described kinematical information, also the rotation angle \( \theta \) aligning the fat jet axis to the "direction after shifting the jet and to the center of the \( \phi \) plane. This information allows the network to reconstruct the full events and access the correlation of the jet shape to the other fat jet and the beam axis.

In Fig. 8, we show the ROC curve of the network trained without the \( \theta \) inputs (red) compared to the ROC curve of our cross-attention model (blue). The improvement on the background rejection is a factor of four for a signal efficiency of 80%. Therefore, including \( \theta \) results in a drastically increased performance. The model with \( \theta \) has higher efficiency at \( m_{J1} \geq m_h \) and \( p_T \geq m_{H2} \). In short, the model can focus more on the \( H \) kinematics with \( \theta \) inputs.

We also looked for simple correlations among \( \theta \) and the other kinematical variables, such as \( \phi \), but did not find any apparent ones contributing to the selection improvement. (The correlations within the internal structures of the jet will be investigated in future publications.)

**Figure 8:** Left) The ROC curve and error band of the full model using \( \theta \) input (red) and the model without \( \theta \) input (blue). The ROC is obtained by using 20,000 signal and background testing events. The error is estimated as in Fig. 6. The middle (right) plot shows the signal efficiency as varying \( m_{J1} \) (\( p_T \)). The ratio is calculated at 80% of the signal efficiency for 20,000 signal samples. The efficiency (without) using \( \theta \) is shown by blue (red) bars indicating statistical errors. The acceptance of the full model is higher than the one without \( \theta \) input at \( m_{J1} \geq m_h \) and \( p_T \geq m_{H2} \).

**Interpretation of the transformer encoder results**

In the following section, we discuss additional methods to interpret and analyze the results of the transformer encoder with cross-attention, which performs best in Fig. 6. The interpretation methods are generic and can be further applied to other networks to interpret their results. As attention-based transformer models excel in capturing intricate spatial relationships and global context within data, their interpretability becomes paramount. Interpretation methods for attention-based transformers aim to elucidate the visual cues, features, and regions that contribute significantly to the model's predictions. Common Interpretation Methods are:

- **Attention Maps:** Attention maps visualize the focus of the model by highlighting the particles in the cloud that receive higher attention. These maps provide a direct better selection of Higgs mass rejecting high PT events.
Deep Learning suffers low interpretability and it is always annoying.

- skip connection of attention blocks helps connecting input data to extracted feature (transformed quantity) in some level.

Figure 1: The Transformer - model architecture.

3.1 Encoder and Decoder Stacks
Encoder:
The encoder is composed of a stack of $N=6$ identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection around each of the two sub-layers, followed by layer normalization. That is, the output of each sub-layer is $\text{LayerNorm}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension $d_{\text{model}} = 512$.

Decoder:
The decoder is also composed of a stack of $N=6$ identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization. We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with the fact that the output embeddings are offset by one position, ensures that the predictions for position $i$ can depend only on the known outputs at positions less than $i$.

3.2 Attention
An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum.

1706.03762 Vaswani et al. “Attention is all you need”
First few "particle" token express Higgs nature efficiently

Cross attention map: Particle in the jet (50) and parent particle (3)

Signal

sum of fatjet momenta capture signal

maybe number of averaged particles are different
GRAD-CAM (1610.02391)

- Output of last attention layers (some correlation with original inputs)

5000 signal events

Y : class score

F : output from last attention layer

\[ \alpha_k(\tilde{\eta}, \tilde{\phi}) = \frac{1}{Z} \sum \frac{\partial Y_c}{\partial F_k(\tilde{\eta}, \tilde{\phi}, \vec{p}_T)} \]

Grad-CAM(\tilde{\eta}, \tilde{\phi}) = \frac{1}{k} \sum_k \alpha_k(\tilde{\eta}, \tilde{\phi}) F_k(\tilde{\eta}, \tilde{\phi}, \vec{p}_T)

Still see some connection between particle location and transformed coordinates.

Inference vs Attention depth
TAKEAWAYS

- **use "cross attention"** when you combine the “high scale information” to the “low energy scale”, because cross attention layer gives extra emphasis to the information linked to the high energy kinematics.

- **skip connection and Interpretation**: Skip connection helps to maintain some connection to the inputs.

- **More Physics**: Heavy particles decay into colored particles (discovery, spin, color structure?) Cross attention network probably more useful to resolve **correlation of jet structures**.

- **Result looks very good to me and I am still worrying about bugs...**
NEED TO BE IMPROVED

• Current GPU requirement: 2 x NVIDIA RTX A6000 (48GB) with 80% and 30% utilization in tensor flow mirror strategy. 96% consumption /card 20min/training.

• We definitely have to change “jet substructure part” to simpler one, keeping cross attention structure(this part is generic)

• Ex: “Modulated Network of HL variables”

• QCD vs top, Amon Furuichi(Nagoya), Sung Hak Lim(Rutgers) and M. Nojiri arXiv 2312.11760[hep-ph] work as good as Particle Transformer.

• ....... but are they robust for color connection?
BACK UP SLIDES
Top jets often have two or three substructures as they decay into a bottom quark and two light quarks. Therefore, structure in two-point and three-point energy correlation is essential to discriminate top jets from QCD jets. We use IRC safe two-point correlation spectrum, which is defined as follows:

\[ S_{2,ab}(R) \equiv \sum_{i \in a} \sum_{j \in b} p_{T,i} p_{T,j} \delta(R - R_{ij}). \]

Here, \( a \) and \( b \) are labels for subsets of jet constituents, \( i \) and \( j \) are labels for their constituents and \( R_{ij} \) is defined as:

\[ R_{ij} = \sqrt{\|\vec{p}_{T,i}\|^2 + \|\vec{p}_{T,j}\|^2}. \]

Notably, all EFP information is included in \( S_{2,ab} \). The structure indexes \( a, b \) are \( \{J_{\text{trim}}, J_{c\text{trim}}\} \) or \( \{J_{\text{lead}}, J_{c\text{lead}}\} \) and we call corresponding \( S_{2,ab} \) inputs as \( x_{\text{trim}} \) and \( x_{\text{lead}} \). These are collectively referred to as \( x_{S_2} \). The \( S_2 \) is binned by \( R = 0.1, 0.2, 0.3 \).

The module of the networks that further compresses the \( S_2 \) information is a simple MLP with two hidden layers. The inputs are combined with \( x_{\text{kin}} \) as shown in Fig. 3a. Two sets of outputs of dimension five each goes into the final convolution layers:

\[ z_{\text{trim}} = \text{trim}(x_{\text{trim}}, x_{\text{kin}}) \]
\[ z_{\text{lead}} = \text{lead}(x_{\text{lead}}, x_{\text{kin}}) \]

See [9] for the detailed setup of the network.
NETWORK USING HL INPUTS (ANALYSIS MODEL=AM)

(b) A schematic diagram of subjet recursive module.

- input: subjet with multiple cone size (R=0.1, 0.2, 0.3) = information of clustering
- Shared MLP for 2nd to 5th subjet to reduce parameters
AM demonstrated much smaller training uncertainty than ParT. Importantly, the smaller behavior in AM through bootstrap methods applied to our jet tagging problems, and our particularly in understanding the di-jets. Identifying the HLFs exclusive to

Figure 7:

with 35% GPU utilization. need lots of preprocessing

AM model: 1GB GPU memory on GeForce 1080Ti GPU (11.3TFLOPS) with 35% GPU utilization, need lots of preprocessing

ParT: 14GB GPU memory RTX A6000 (38.7TFLOPS) GPU utilization 95%