Quark/Gluon Discrimination and Top Tagging with Dual Attention Transformer

— EuCAIFCon24 —

Daohan Wang

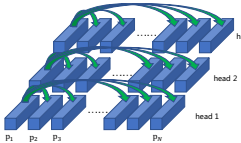
Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences (OeAW)

April 30, 2024

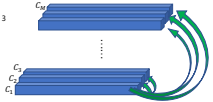
Dual Attention Mechanism

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_{N_h})$$

$$\text{where } \mathrm{head}_i = \mathrm{softmax}\left[\frac{\mathbf{Q}_i(\mathbf{K}_i)^{\mathrm{T}}}{\sqrt{C_h}} + \mathbf{U_1}\right]\mathbf{V}_i$$



Particle Attention Map $P \times P$

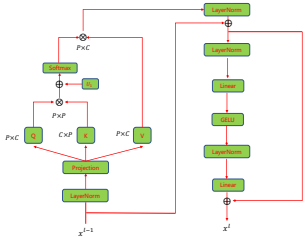


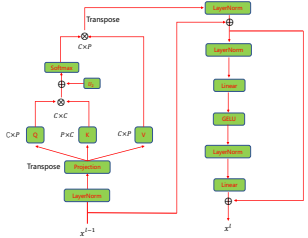
Particle Feature Attention Map $C \times C$

$$\mathcal{A}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = \mathrm{softmax}\left[\frac{\mathbf{Q}_i^{\mathrm{T}}\mathbf{K}_i}{\sqrt{C}} + \mathbf{U_2}\right]\mathbf{V}_i^{T}$$

Particle interaction matrix $\mathbf{U_1}$:

$$\Delta R = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2},$$
$$k_{\mathrm{T}} = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})\Delta,$$
$$z = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})/(p_{\mathrm{T},a} + p_{\mathrm{T},b}),$$
$$m^2 = (E_a + E_b)^2 - \|\mathbf{p}a + \mathbf{p}_b\|^2,$$
$$\Delta p_{\mathrm{T}} = |p_{\mathrm{T},a} - p_{\mathrm{T},b}|$$





Channel interaction matrix $\mathbf{U_2}$:

Straightforward ratios of
$\{\mathrm{E_J}, p_{\mathrm{T_J}}, \sum p_{Tf}, \sum E_f, \overline{\Delta\eta}, \overline{\Delta\phi}, \overline{\Delta R}, \mathrm{PID}\}$

where $\overline{\Delta\eta}$, $\overline{\Delta\phi}$ and $\overline{\Delta R}$ correspond to the transverse momentum weighted sum of the $\Delta\eta$, $\Delta\phi$, ΔR of all the constituent particles inside the input jet, respectively. Here $\Delta\eta$, $\Delta\phi$ and ΔR refer to the distances in the $\eta - \phi$ space between each constituent particle and the input jet.

Model Architecture

- Input features: $\log E$, $\log p_T$, $\frac{p_T}{p_{TJ}}$, $\frac{E}{E_J}$, $\Delta\eta$ $\Delta\phi$, ΔR, PID of leading 100 particles.
- The particle attention module ($P \times P$ attention map) and the channel attention module ($C \times C$ attention map) are stacked while maintaining a consistent feature dimension of $N = 64$ and they can complement each other.
- Particle - Dual Attention Transformer: 2 Feature Extractor (1 EdgeConv + 3 Conv2D + 1 AvgPool) + 2 Particle Attention modules + 2 Channel Attention modules + 1D CNN + MLP.