Quark/Gluon Discrimination and Top Tagging with Dual Attention Transformer

Daohan Wang
daohan.wang@oeaw.ac.at

HEPHY INSTITUTE OF HIGH ENERGY PHYSICS
ÖAW AUSTRIAN ACADEMY OF SCIENCES

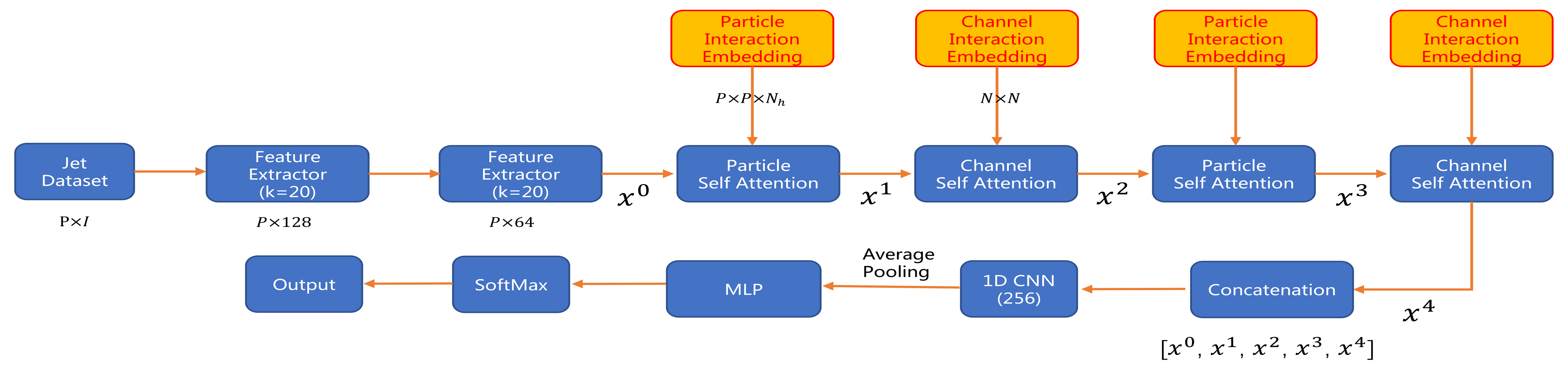
Introduction

Jet tagging is a crucial task in HEP experiments. Over the past decade, deep learning approaches have been extensively adopted to enhance the jet tagging performance. Various jet representations and the corresponding architectures have been proposed, including image, sequence, tree, graph and point cloud. Notably, the Transformer architecture[3] has been adapted for HEP with point cloud representation through models like the Point Cloud Transformer (PCT)[1] and the Particle Transformer[2]. In this study, we introduce the Particle Dual Attention Transformer (P-DAT) for jet tagging. This novel transformer architecture stands out by concurrently capturing both global and local information, while maintaining computational efficiency.

Model Implementation

- PYTORCH framework with Binary Cross Entropy loss
- AdamW optimizer with lr = 0.0005 on a mini-batch of 64 samples
- 100 epochs with learning rate decreasing by a factor of 2 every 10 epochs to a minimal of 10^{-6}
- Chunk Loading strategy: Within a loop, input data chunks are dynamically loaded for training, validation, and test. Each chunk contains 1280 events. During each iteration, once the chunk is processed for training/validation/test, the loaded data is removed to free up memory resources and the next chunk of data is loaded for next iteration.

Particle - Dual Attention Transformer



2 Feature Extractor (1 EdgeConv + 3 Conv2D + 1 AvgPool) + 2 Particle Attention modules + 2 Channel Attention modules + 1D CNN + MLP.

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.
- Channel interactions:
  Ratios of $\{E,\ p_T,\ \sum p_{Tf},\ \sum E_f,\ \overline{\Delta\eta},\ \overline{\Delta\phi},\ \overline{\Delta R},\ \text{PID}\}$
- Particle interactions:
  ΔR, m^2, $\min(p_{T,a}, p_{T,b})\Delta R$, $\min(p_{T,a}, p_{T,b})/(p_{T,a} + p_{T,b})$, Δp_T, $\delta_{i,j}$ of each pair of particles.
- Particle Attention Module computes the attention weights between each pair of particles, with particle interaction matrix U_1 as a bias term.

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{N_h})$$
$$\text{where head}_i = \text{softmax}\left[\frac{\mathbf{Q}_i(\mathbf{K}_i)^T}{\sqrt{C_h}} + \mathbf{U_1}\right]\mathbf{V}_i \quad (1)$$

- Channel Attention Module computes the attention weights between each pair of particle features, with channel interaction matrix U_2 as a bias term.

$$\mathcal{A}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = \text{softmax}\left[\frac{\mathbf{Q}_i^T\mathbf{K}_i}{\sqrt{C}} + \mathbf{U_2}\right]\mathbf{V}_i^T \quad (2)$$

- Combination:
  The two particle attention modules ($P \times P$ attention maps) and two channel attention modules ($C \times C$ attention maps) are stacked while maintaining a consistent feature dimension of $N = 64$. By alternatively applying these two types of modules, the local and global information can be captured simultaneously and complement each other.

Quark/gluon Discrimination

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$	Parameters	FLOPs
ResNeXt-50	0.821	0.9060	30.9	80.8	1.46M	-
P-CNN	0.827	0.9002	34.7	91.0	354k	15.5M
PFN	-	0.9005	34.7±0.4	-	86.1k	4.62M
ParticleNet-Lite	0.835	0.9079	37.1	94.5	26k	-
ParticleNet	0.840	0.9116	39.8±0.2	98.6±1.3	370k	540M
ABCNet	0.840	0.9126	42.6±0.4	118.4±1.5	230k	-
SPCT	0.815	0.8910	31.6±0.3	93.0±1.2	7k	2.4M
PCT	0.841	0.9140	43.2±0.7	118.0±2.2	193.3k	266M
LorentzNet	0.844	0.9156	42.4±0.4	110.2±1.3	224k	-
ParT	0.849	0.9203	47.9±0.5	129.5±0.9	2.13M	260M
P-DAT	0.839	0.9092	39.2±0.6	95.1±1.3	498k	144M

Top Tagging

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$	Parameters	FLOPs
ResNeXt-50	0.936	0.9837	302±5	1147±58	1.46M	-
P-CNN	0.930	0.9803	201±4	759±24	354k	15.5M
PFN	-	0.9819	247±3	888±17	86.1k	4.62M
ParticleNet-Lite	0.937	0.9844	325±5	1262±49	26k	-
ParticleNet	0.940	0.9858	397±7	1615±93	370k	540M
JEDI-net	0.9263	0.9786	-	590.4	-	-
SPCT	0.928	0.9799	201±9	725±54	7k	2.4M
PCT	0.940	0.9855	392±7	1533±101	193.3k	266M
LorentzNet	0.942	0.9868	498±18	2195±173	224k	-
ParT	0.940	0.9858	413±16	1602±81	2.13M	260M
P-DAT	0.932	0.9768	228±8	876±39	498k	144M

References

[1] Vinicius Mikuni and Florencia Canelli. Point cloud transformers applied to collider physics. *Mach. Learn. Sci. Tech.*, 2(3):035027, 2021.

[2] Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for Jet Tagging. 2 2022.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.