Proceedings of the 39th International Conference on Machine Learning, PMLR 162:18281-18292

ParT

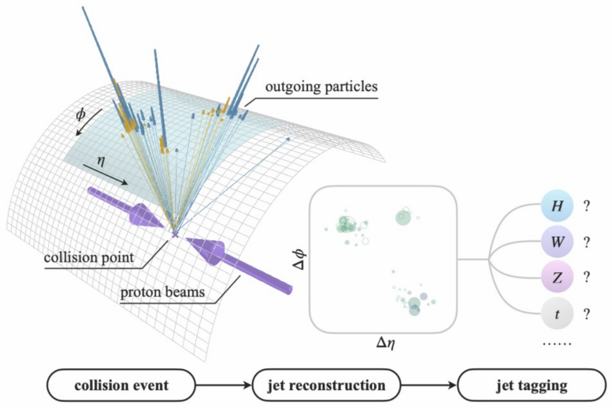
Particle Transformer for Jet Tagging

ArXiv: 2202.03772, Huilin Qu (CERN), Congqiao Li & Sitian Qian (PKU)

The 21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research (***ACAT2022***)
2022/10/24~2022/10/28

Sitian Qian (PKU)

Introduction



- Jet tagging: an ideal bridge linking HEP & ML
  - Important tool for HEP community
  - Well-defined task for ML enthusiasts
- Graph Neural Networks (GNN) + Point Cloud (PC) Representation = State-Of-The-ArT (SOTA)
  - First show up in ParticleNet: previous SOTA for top tagging benchmark
  - Since then, various models are proposed under GNN+PC framework
    - ABCNet
    - ParticleNeXt
    - HEP application of Point Cloud Transformer
  - Can we add another term to the right-hand side?

Attention Mechanism And Transformer

- Attention mechanism has drawn the attention of ML community
  - Success in various ML communities: Natural Language Processing (NLP), Computer Vision (CV)…

BERT (1810.04805):
the very first transformer, overperformed every other model even human in all kinds of NLP task!

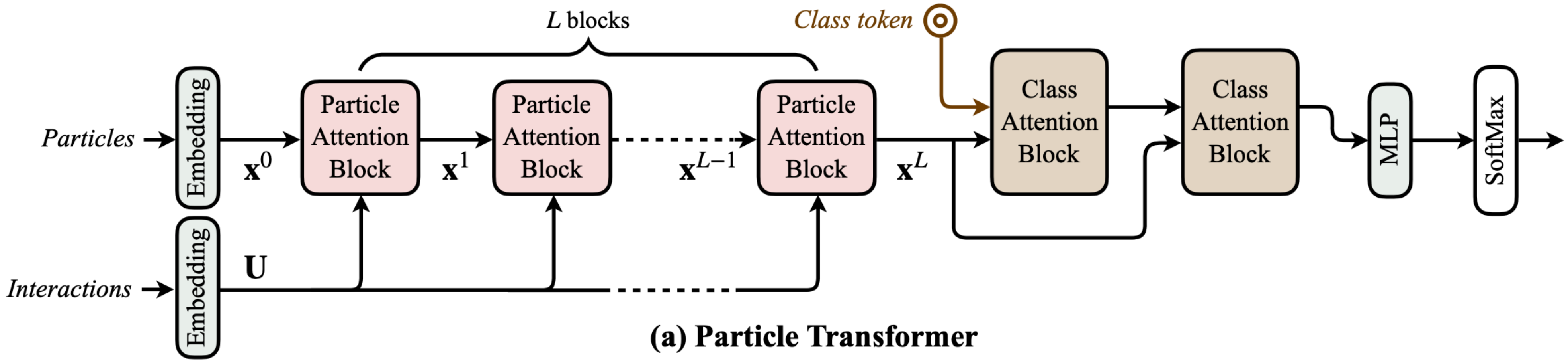
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\mathrm{BERT}_{\mathrm{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\mathrm{BERT}_{\mathrm{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\mathrm{BERT}_{\mathrm{BASE}}$	81.6	-
$\mathrm{BERT}_{\mathrm{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

- Already several attempts in HEP context:
  - ABCNet, ParticleNeXt, Point Cloud Transformer

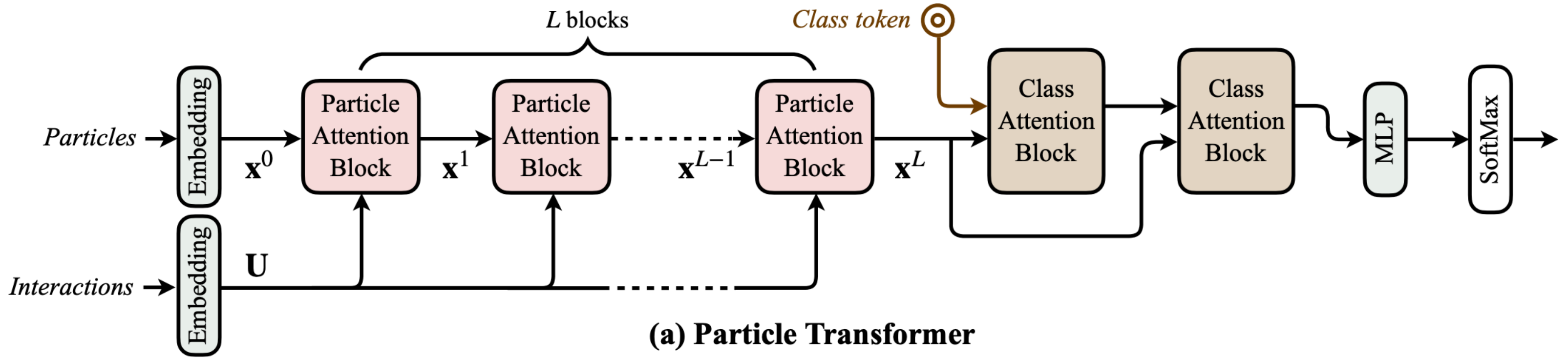
Particle Transformer

- Particle Transformer (ParT): transformer designed for particle physics
  - Input embedding: Not only inject single particle information, but also include pair-wise feature

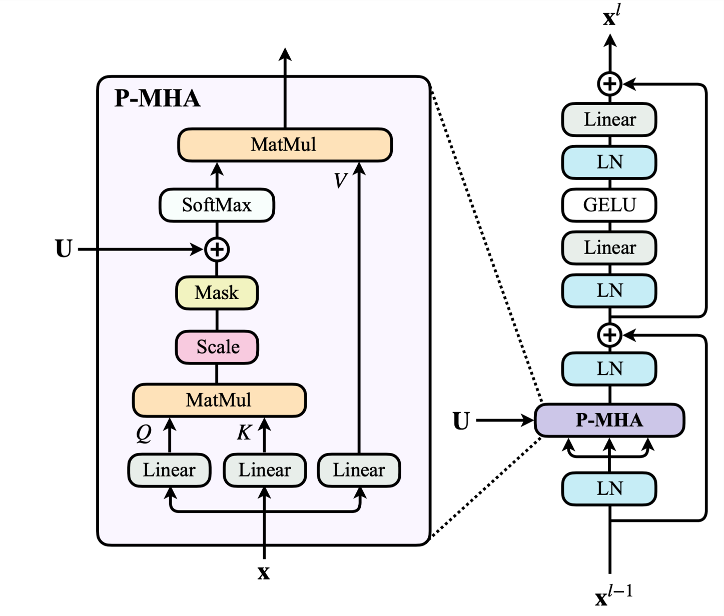


(a) Particle Transformer

Particle Transformer
Particle Attention Block



(a) Particle Transformer



(b) Particle Attention Block

Multi-head Attention (MHA)
powered feature extraction (embedding)

$$\text{P-MHA}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k} + \mathbf{U})V$$

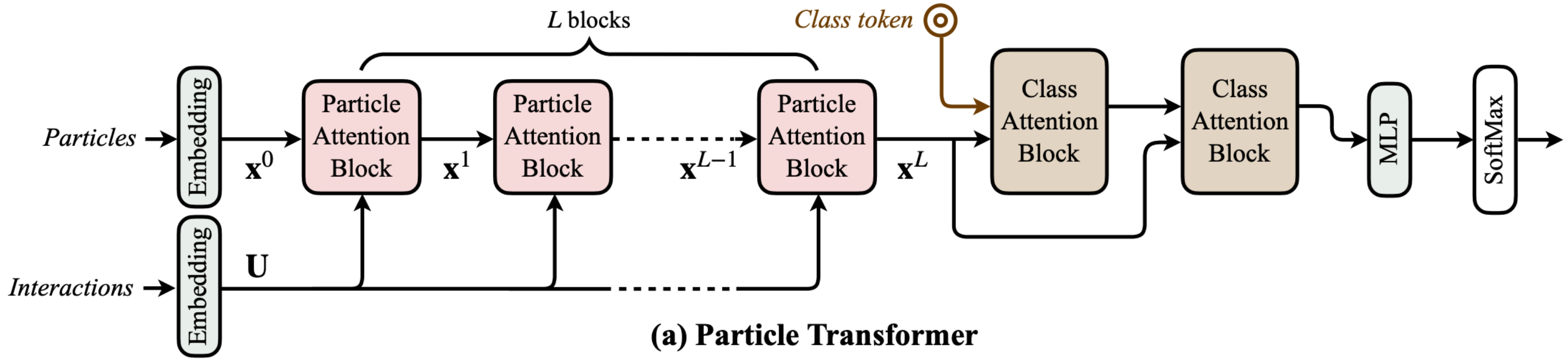
d_k: dimension of K

Kinematic Variables:
based on LundNet

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

Particle interactions come in as pair-wise features

Particle Transformer
Class Attention Block



(a) Particle Transformer

Output of class attention blocks will be imported to MLP + softmax for final classification scores

Multi-head Attention (MHA) powered, class information comes in for classification

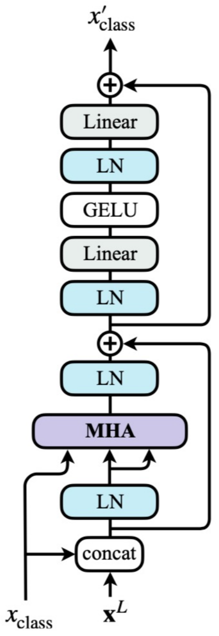
$$\mathrm{MHA_C}(Q_\mathrm{C}, K_\mathrm{C}, V_\mathrm{C}) = \mathrm{SoftMax}(Q_\mathrm{C}K_\mathrm{C}^T/\sqrt{d_{k\mathrm{C}}})V_\mathrm{C}$$

$Q_\mathrm{C} = W_{q\mathrm{C}}x_\mathrm{class} + b_{q\mathrm{C}}$

$$\underbrace{K_\mathrm{C} = W_{k\mathrm{C}}\mathbf{z} + b_{k\mathrm{C}} \qquad V_\mathrm{C} = W_{v\mathrm{C}}\mathbf{z} + b_{v\mathrm{C}}}_{\mathbf{z} = [x_\mathrm{class}, \mathbf{x}^L]}$$

$d_{k\mathrm{C}}$: dimension of K_C

Concatenate class information and particle embedding



(c) Class Attention Block

Larger Model Calls for Larger Dataset

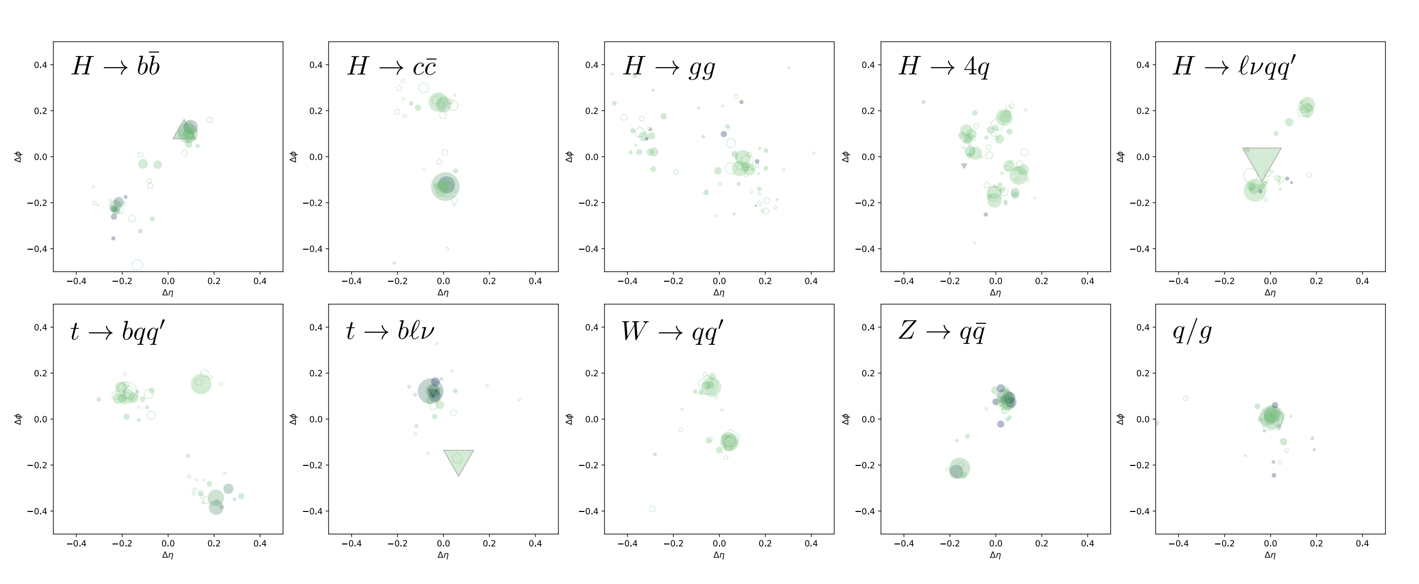
- ParT is one of the largest ML model for jet tagging
  - Require enough statistics to avoid overtraining!
- Existing datasets however are not large
  - Top tagging: 2M jets
  - Quark-gluon tagging: 2M jets
  - JetNet: 500k jets
  - Jedi-Net: 880k jets
  - Higgs boson tagging: 3.9M signal jets, 1.9M background jets
- A large dataset is needed
  - JetClass dataset

Table 2. Number of trainable parameters and FLOPs.

	Accuracy	# params	FLOPs
PFN	0.772	86.1 k	4.62 M
P-CNN	0.809	354 k	15.5 M
ParticleNet	0.844	370 k	540 M
ParT	**0.861**	2.14 M	340 M
ParT (plain)	0.849	2.13 M	260 M

The JetClass Dataset

- JetClass is inclusive:
  - 10 types of jets
  - Kinematics,
  - PID,
  - trajectory displacement



- JetClass is large:
  - 100M jets for training → 10M each class
  - 5M for validation
  - 20M for test → 2M each class

The JetClass Dataset

- Simulation details:
  - MadGraph5_aMC@NLO: ME level, production & decay of top, W/Z & Higgs boson
  - Pythia8: Parton showering and hadronization
  - Delphes: fast simulation of detector response, CMS configuration
  - Jets: anti-k_T algorithm, R=0.8 on Delphes E-Flow objects,
    - $|\eta|<2$, 500 GeV < p_T < 1000 GeV
    - "high quality" jets only: jets fully containing decay products.
- Proposals of evaluation metrics for classification with JetClass:
  - Common metrics: Accuracy (Acc.) and Area Under (ROC) Curve (AUC)
  - HEP interest: Background rejection at given signal efficiency

$$\boxed{\mathrm{Rej}_{X\%} \equiv 1/\mathrm{FPR} \text{ at } \mathrm{TPR} = X\%}$$

Numerical Experiments With JetClass

	All classes		$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to 4q$	$H \to \ell\nu qq'$	$t \to bqq'$	$t \to b\ell\nu$	$W \to qq'$	$Z \to q\bar{q}$
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
PFN	0.772	0.9714	2924	841	75	198	265	797	721	189	159
P-CNN	0.809	0.9789	4890	1276	88	474	947	2907	2304	241	204
ParticleNet	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT	**0.861**	**0.9877**	**10638**	**4149**	**123**	**1864**	**5479**	**32787**	**15873**	**543**	**402**
ParT (plain)	0.849	0.9859	9569	2911	112	1185	3868	17699	12987	384	311

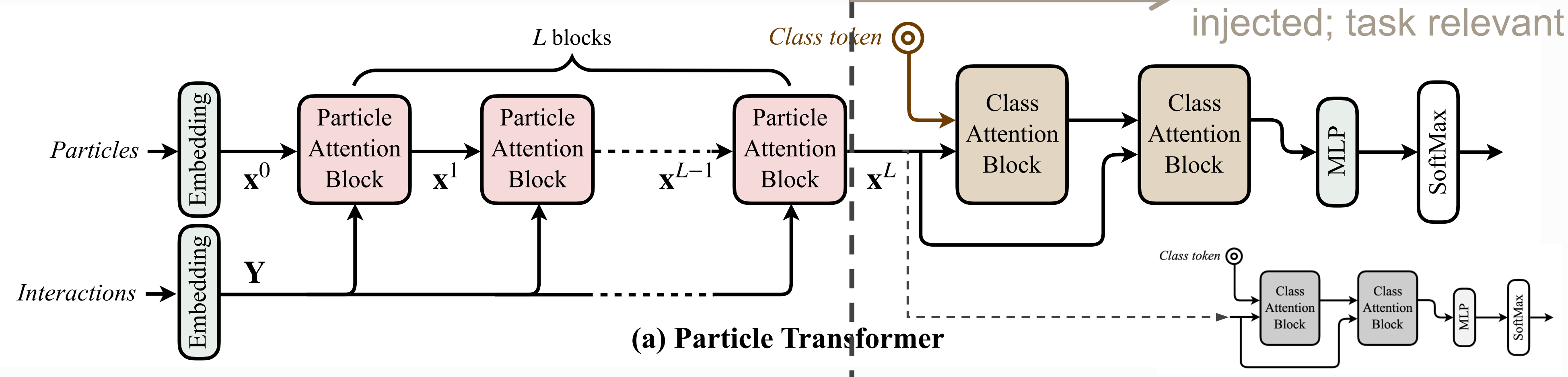
ParT (plain): no pair-wise feature mask applied

- ✓ ParT achieves SOTA performance in every classification task!
- ✓ Adding pair-wise feature enhances performance even further!
- ✓ Take H → ccbar as an example: doubled background rejection → ~ 1.4x significance!
  - ✓ Same significance reach with half data!

Pre-training: The Lesson From JetClass

- Pre-train + Fine-tune becomes the trend in ML community
  - Self-attention from transformer → <u>task irrelevant</u> embedding
  - Large dataset → embedding captures <u>generic</u> information

Class information injected; task relevant



(a) Particle Transformer

No class information; task-irrelevant

Task-irrelevant information can be used for different tasks: Top tagging, quark/gluon discrimination

Fine-tuning
With different datasets

Top Tagging

Table 3. Comparison between ParT and existing models on the top quark tagging dataset. ParT-f.t. denotes the model pre-trained on JetClass and fine-tuned on this dataset. ParT refers to the model trained from scratch on this dataset. Results for other models are quoted from their published results: P-CNN and ParticleNet (Qu & Gouskos, 2020), PFN (Komiske et al., 2019b), JEDI-net (Moreno et al., 2020), PCT (Mikuni & Canelli, 2021), LGN (Bogatskiy et al., 2020), and rPCN (Shimmin, 2021).

	Accuracy	AUC	$\mathrm{Rej}_{50\%}$	$\mathrm{Rej}_{30\%}$
P-CNN	0.930	0.9803	201 ± 4	759 ± 24
PFN	—	0.9819	247 ± 3	888 ± 17
ParticleNet	0.940	0.9858	397 ± 7	1615 ± 93
JEDI-net (w/ $\sum O$)	0.930	0.9807	—	774.6
PCT	0.940	0.9855	392 ± 7	1533 ± 101
LGN	0.929	0.964	—	435 ± 95
rPCN	—	0.9845	364 ± 9	1642 ± 93
ParT	0.940	0.9858	413 ± 16	1602 ± 81
ParT-f.t.	**0.944**	**0.9877**	$\mathbf{691 \pm 15}$	$\mathbf{2766 \pm 130}$

q/g discrim.

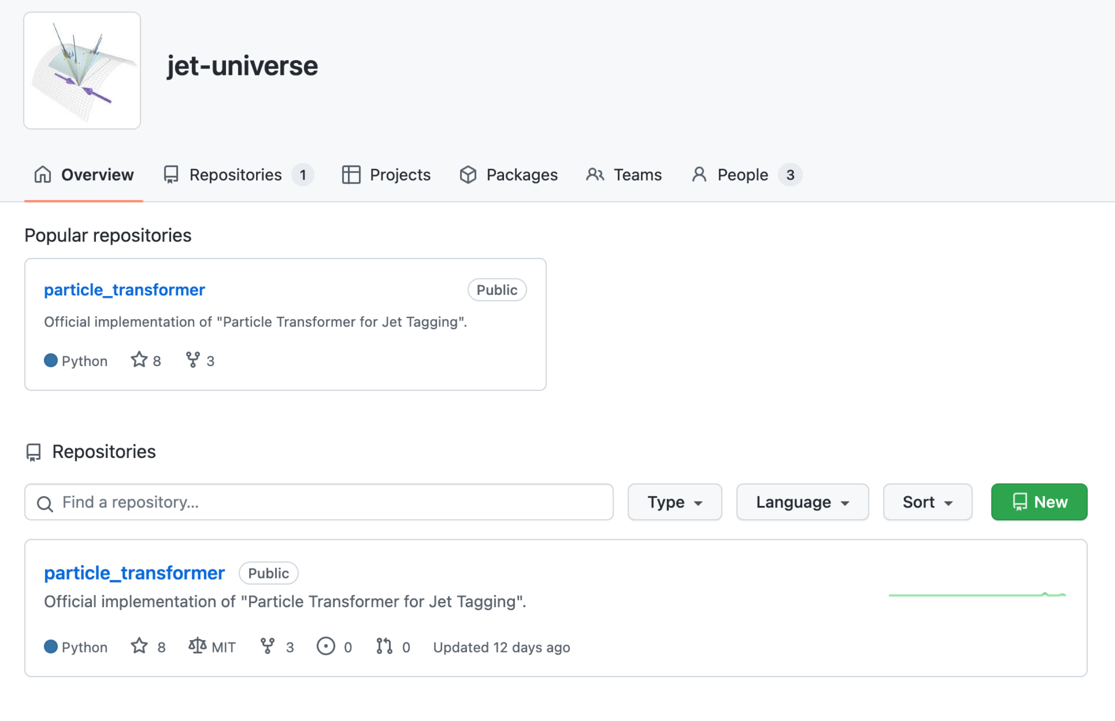
Table 4. Comparison between ParT and existing models on the quark-gluon tagging dataset. ParT-f.t. denotes the model pre-trained on JetClass and fine-tuned on this dataset. ParT refers to the model trained from scratch on this dataset. Results for other models are quoted from their published results: P-CNN and ParticleNet (Qu & Gouskos, 2020), PFN (Komiske et al., 2019b), ABCNet (Mikuni & Canelli, 2020), PCT (Mikuni & Canelli, 2021), and rPCN (Shimmin, 2021). The subscript "exp" and "full" distinguish models using partial or full particle identification information.

	Accuracy	AUC	$\mathrm{Rej}_{50\%}$	$\mathrm{Rej}_{30\%}$
$\mathrm{P\text{-}CNN}_{\mathrm{exp}}$	0.827	0.9002	34.7	91.0
$\mathrm{PFN}_{\mathrm{exp}}$	—	0.9005	34.7 ± 0.4	—
$\mathrm{ParticleNet}_{\mathrm{exp}}$	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3
$\mathrm{rPCN}_{\mathrm{exp}}$	—	0.9081	38.6 ± 0.5	—
$\mathrm{ParT}_{\mathrm{exp}}$	0.840	0.9121	41.3 ± 0.3	101.2 ± 1.1
$\mathbf{ParT\text{-}f.t.}_{\mathbf{exp}}$	**0.843**	**0.9151**	$\mathbf{42.4 \pm 0.2}$	$\mathbf{107.9 \pm 0.5}$
$\mathrm{PFN}_{\mathrm{full}}$	—	0.9052	37.4 ± 0.7	—
$\mathrm{ABCNet}_{\mathrm{full}}$	0.840	0.9126	42.6 ± 0.4	118.4 ± 1.5
$\mathrm{PCT}_{\mathrm{full}}$	0.841	0.9140	43.2 ± 0.7	118.0 ± 2.2
$\mathrm{ParT}_{\mathrm{full}}$	0.849	0.9203	47.9 ± 0.5	129.5 ± 0.9
$\mathbf{ParT\text{-}f.t.}_{\mathbf{full}}$	**0.852**	**0.9230**	$\mathbf{50.6 \pm 0.2}$	$\mathbf{138.7 \pm 1.3}$

Take home:
Pre-train with JetClass helps ParT to reach SOTA performance!

Welcome to Jet-Universe!

- We are more than glad to share our work to the whole community:
  - Both source code of ParT and JetClass dataset are now public at Jet Universe
  - Hope to see more enthusiasts onboard!



Summary

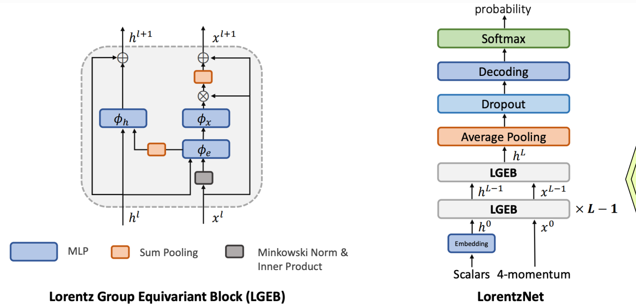
- Particle Transformer (ParT):
  - Dedicated transformer architecture for jet tagging
  - SOTA in various benchmarks
  - Particle interactions help ParT to perform better
- JetClass dataset:
  - Large and inclusive: order of magnitude higher in statistics and classes of jets
  - Pretrain with JetClass enhance ParT's performance on other datasets
- ParT and JetClass are publicly available on Github!
  - Welcome to Jet Universe

End of the Journey?

- There is still room for further improvement:
  - "Physics" should be a term added to the "SOTA equation"
    - Pair-wise features from particle interactions enhance ParT's performance
    - Improvements has been seen in physics inspired models:
      - LundNet, LorentzNet (Lorentz symmetry applied, competitive performance),…

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong[a,e,1] Qi Meng[b] Jue Zhang[b] Huilin Qu[c] Congqiao Li[d] Sitian Qian[d] Weitao Du[a] Zhi-Ming Ma[a] Tie-Yan Liu[b]



JHEP07(2022)030

Top Tagging

Model	Accuracy	AUC	$1/\varepsilon_B$ ($\varepsilon_S = 0.5$)	$1/\varepsilon_B$ ($\varepsilon_S = 0.3$)
ResNeXt	0.936	0.9837	302 ± 5	1147 ± 58
P-CNN	0.930	0.9803	201 ± 4	759 ± 24
PFN	0.932	0.9819	247 ± 3	888 ± 17
ParticleNet	0.940	0.9858	397 ± 7	1615 ± 93
EGNN	0.922	0.9760	148 ± 8	540 ± 49
LGN	0.929	0.9640	124 ± 20	435 ± 95
LorentzNet	**0.942**	**0.9868**	$\mathbf{498 \pm 18}$	$\mathbf{2195 \pm 173}$
ParT	0.940	0.9858	413 ± 16	1602 ± 81
ParT-f.t.	**0.944**	**0.9877**	$\mathbf{691 \pm 15}$	$\mathbf{2766 \pm 130}$

q/g discrim.

Model	Accuracy	AUC	$1/\varepsilon_B$ ($\varepsilon_S = 0.5$)	$1/\varepsilon_B$ ($\varepsilon_S = 0.3$)
ResNeXt	0.821	0.8960	30.9	80.8
P-CNN	0.827	0.9002	34.7	91.0
PFN	–	0.9005	34.7 ± 0.4	–
ParticleNet	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3
EGNN	0.803	0.8806	26.3 ± 0.3	76.6 ± 0.5
LGN	0.803	0.8324	16.0	44.3
LorentzNet	**0.844**	**0.9156**	$\mathbf{42.4 \pm 0.4}$	$\mathbf{110.2 \pm 1.3}$
ParT$_{exp}$	0.840	0.9121	41.3 ± 0.3	101.2 ± 1.1
ParT-f.t.$_{exp}$	**0.843**	**0.9151**	$\mathbf{42.4 \pm 0.2}$	$\mathbf{107.9 \pm 0.5}$

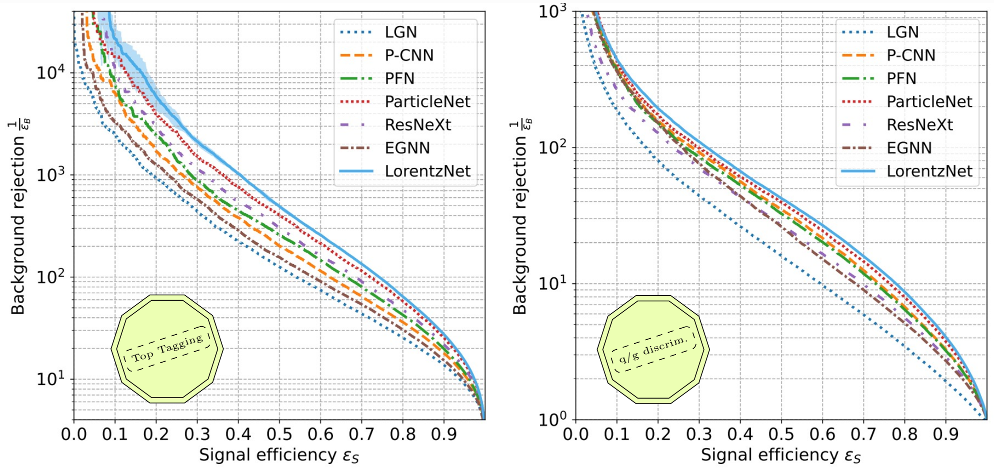


Figure 2. A comparison of ROC curves between LorentzNet and other algorithms on top tagging dataset (left) and quark-gluon dataset (right).

End of the Journey?

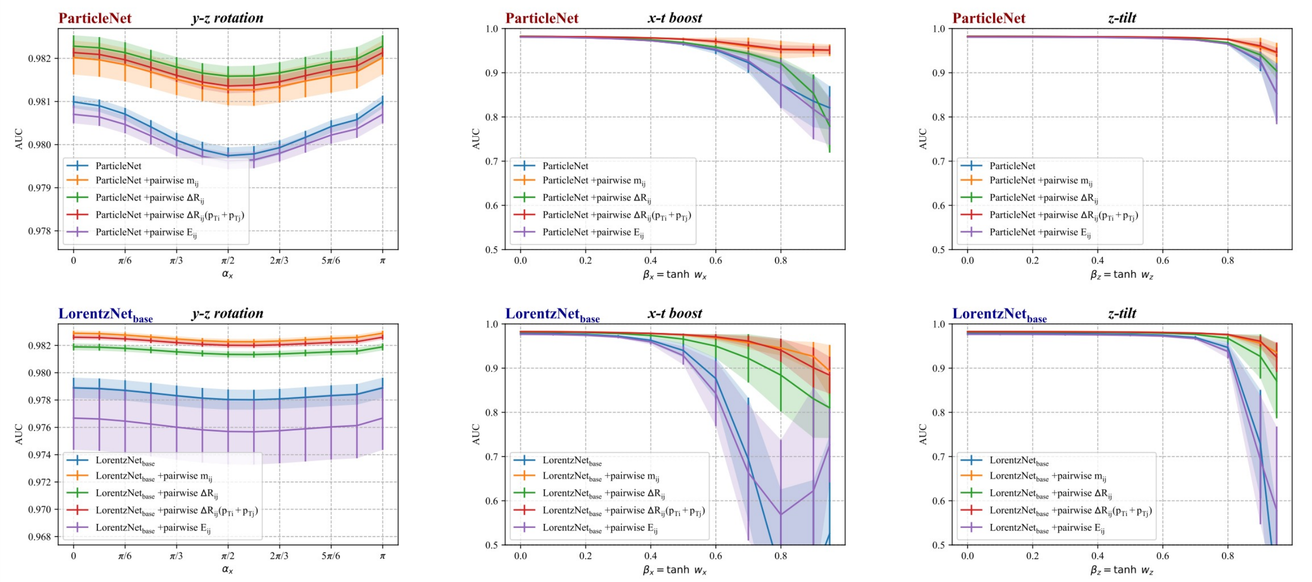
- There is still room for further improvement:
  - “Physics” should be a term added to the “SOTA equation”
    - Pair-wise features from particle interactions enhance ParT’s performance
    - Improvements has been seen in physics inspired models:
      - LundNet, LorentzNet (Lorentz symmetry applied, competitive performance),…
      - Lorentz symmetry is indeed important!

Does Lorentz-symmetric design boost network performance in jet physics?

Congqiao Li,[1,*] Huilin Qu,[2] Sitian Qian,[1] Qi Meng,[3] Shiqi Gong,[4] Jue Zhang,[3] Tie-Yan Liu,[3] and Qiang Li[1]

Our answer: **YES**!

Adding Lorentz symmetry inspired features (either pairwise or elementwise) will enhance performance of NN



https://arxiv.org/abs/2208.07814

More details: ML4Jets 2022 Next Week!

End of the Journey?

- There is still room for further improvement:
  - "Physics" should be a term added to the "SOTA equation"
    - Pair-wise features from particle interactions enhance ParT's performance
    - Improvements has been seen in physics inspired models:
      - LundNet, LorentzNet (Lorentz symmetry applied, competitive performance),…
      - Lorentz symmetry is indeed important!
  - Education from ML community:
    - We have already benefited a lot from ML community (transformer, pair-wise features, pretrain, etc.)
    - Attempts on novel techniques (Multi-modal transformers, neighbor embedding…) are promising to explore!
  - A practical perspective: model complexity vs efficient computation
    - Larger and Larger ML models call for model compression techniques.

Back Up
Input Features

Table 5. Particle input features used for jet tagging on the JETCLASS, the top quark tagging (TOP) and the quark gluon tagging (QG) datasets. For QG, we consider two scenarios: $\mathrm{QG}_{\mathrm{exp}}$ is restricted to use only the 5-class experimentally realistic particle identification information, while $\mathrm{QG}_{\mathrm{full}}$ uses the full set of particle identification information in the dataset and further distinguish between different types of charged hadrons and neutral hadrons.

Category	Variable	Definition	JETCLASS	TOP	$\mathrm{QG}_{\mathrm{exp}}$	$\mathrm{QG}_{\mathrm{full}}$
Kinematics	$\Delta\eta$	difference in pseudorapidity η between the particle and the jet axis	✓	✓	✓	✓
	$\Delta\phi$	difference in azimuthal angle ϕ between the particle and the jet axis	✓	✓	✓	✓
	$\log p_{\mathrm{T}}$	logarithm of the particle's transverse momentum p_{T}	✓	✓	✓	✓
	$\log E$	logarithm of the particle's energy	✓	✓	✓	✓
	$\log \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}(\mathrm{jet})}$	logarithm of the particle's p_{T} relative to the jet p_{T}	✓	✓	✓	✓
	$\log \frac{E}{E(\mathrm{jet})}$	logarithm of the particle's energy relative to the jet energy	✓	✓	✓	✓
	ΔR	angular separation between the particle and the jet axis ($\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$)	✓	✓	✓	✓
Particle identification	`charge`	electric charge of the particle	✓	—	✓	✓
	`Electron`	if the particle is an electron (`|pid|==11`)	✓	—	✓	✓
	`Muon`	if the particle is an muon (`|pid|==13`)	✓	—	✓	✓
	`Photon`	if the particle is an photon (`pid==22`)	✓	—	✓	✓
	`CH`	if the particle is an charged hadron (`|pid|==211 or 321 or 2212`)	✓	—	✓	✓[a]
	`NH`	if the particle is an neutral hadron (`|pid|==130 or 2112 or 0`)	✓	—	✓	✓[b]
Trajectory displacement	$\tanh d_0$	hyperbolic tangent of the transverse impact parameter value	✓	—	—	—
	$\tanh d_z$	hyperbolic tangent of the longitudinal impact parameter value	✓	—	—	—
	σ_{d_0}	error of the measured transverse impact parameter	✓	—	—	—
	σ_{d_z}	error of the measured longitudinal impact parameter	✓	—	—	—

[a] `(|pid|==211) + (|pid|==321)*0.5 + (|pid|==2212)*0.2`
[b] `(|pid|==130) + (|pid|==2112)*0.2.`