Leveraging universality of jet taggers through transfer learning

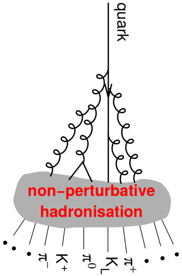
5th Inter-experiment Machine Learning Workshop, 13 May 2022

Frédéric Dreyer, **Radosław Grabarczyk**, Pier Monni

based on arXiv:2203.06210

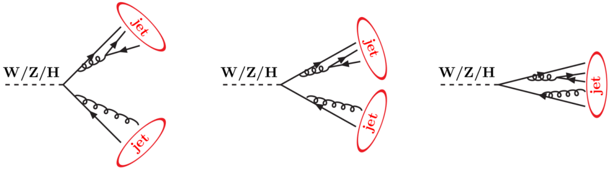
Introduction

- A jet is a collimated bunch of hadrons resulting from QCD showering of coloured particles + hadronisation.



- Jets are prevalent in high energy hadronic collisions. They act as probes of the hard event which is not directly visible.
- Identifying the source of a jet - jet tagging - is crucial for searches for new physics and study of the Standard Model at particle colliders.

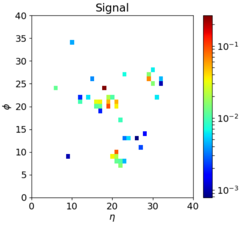
$p_t \lesssim m$ $\qquad$ $p_t \gg m$

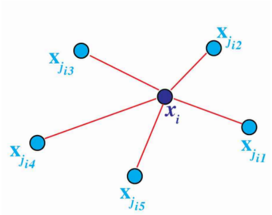


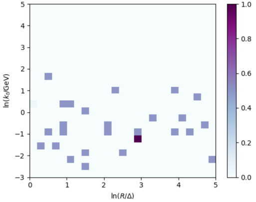
- Boosted jet tagging algorithms are produced by identifying patterns in jet substructure.

- Machine learning techniques outperform analytical discriminants in this task. Many interesting ways to represent a jet: calorimeter images, set of 4-momenta, theory-inpired taggers...

see HEPML-LivingReview for different methods and applications. Image sources: left, middle, right.





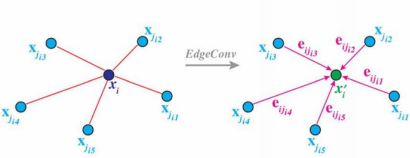


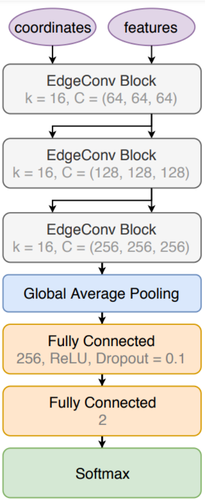
PARTICLENET

ParticleNet

[Qu, Gouskos 10.1103/PhysRevD.101.056019]

- Input: point cloud representing a set of particles; k nearest neighbouring points connected.
- Coordinates of points: particle 4-momenta.





ParticleNet

- State of the art performance.
- Full version of ParticleNet has a large number of parameters and performs a costly nearest-neighbour search after each graph convolution.

Time to train on Oxford cluster:

30h 09min 44s

30 epochs on 10^6 events, NVIDIA GeForce RTX 2080 Ti GPU

LUNDNET

Lund plane representation

To create a Lund plane representation of a jet, use the (Cambridge/Aachen) clustering sequence of the jet to associate a unique Lund tree to each jet.

1. Undo the last clustering step, defining two subjets j_1, j_2 ordered in transverse momentum.
2. Save the kinematics of the current declustering step i as a tuple $\mathcal{T}^{(i)} = \{k_t, \Delta, z, m, \psi\}$

$$\Delta \equiv (y_1 - y_2)^2 + (\phi_1 - \phi_2)^2, \quad k_t \equiv p_{t1}\Delta,$$
$$m^2 \equiv (p_1 + p_2)^2, \quad z \equiv \frac{p_{t1}}{p_{t1}+p_{t2}}, \quad \psi \equiv \tan^{-1}\frac{y_2 - y_1}{\phi_2 - \phi_1}.$$

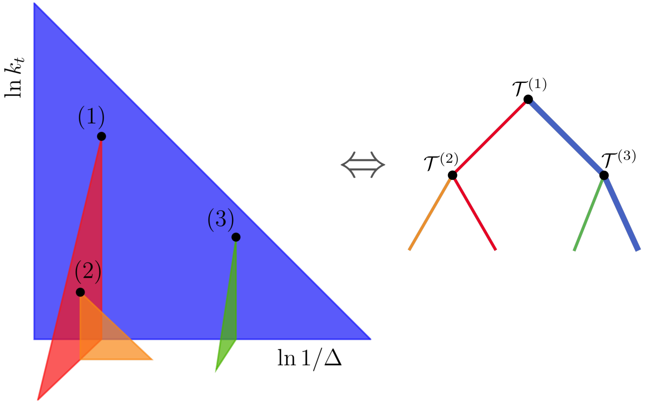
3. Repeat this procedure on both j_1 and j_2 until they are single particles.

Cambridge/Aachen clustering: pairwise recombination of particles with smallest Δ separation.

[Dreyer, Salam, Soyez, JHEP 1812 (2018) 064]

Lund plane representation

- Each jet is thus mapped onto a tree of Lund declusterings from its clustering sequence.
- We can use this Lund tree as an input to a graph neural network.

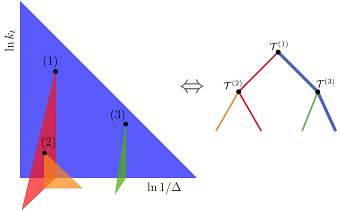


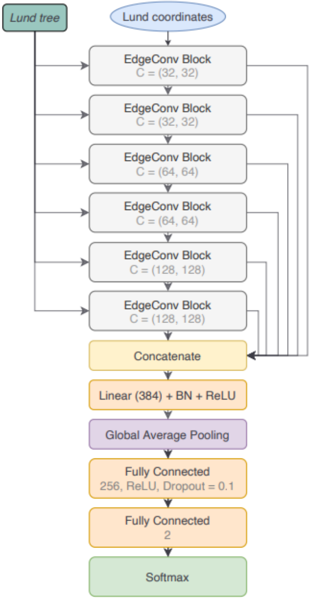
LundNet

[Dreyer, Qu JHEP 03 (2021) 052]

- Input: declustering tree graph with each node representing a splitting; fixed graph structure.
- Tuple of kinematic variables as coordinates of each node.

LundNet-5 :
$\mathcal{T}^{(i)} = (\ln k_t, \ln \Delta, \ln z, \ln m, \psi)$
LundNet-3 :
$\mathcal{T}^{(i)} = (\ln k_t, \ln \Delta, \ln z)$





LundNet

- Better performance than `ParticleNet` for top tagging.
- No nearest neighbour search, higher level kinematic information gives considerable speedup.

Time to train on Oxford cluster:

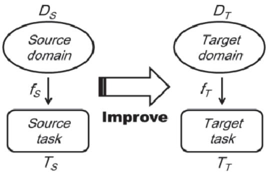
3h 48min 14s

30 epochs on 10^6 events, NVIDIA GeForce RTX 2080 Ti GPU. Preprocessing time **not included**.

TRANSFER LEARNING

Transfer learning

- ML jet tagging models are costly in terms of data needed for training/take a long time to train.
- Resultant models are task-specific, but early convolutional layers learn general features.
- If domains are similar enough, one can use the learned parameters of a network pretrained on a *source task* to improve learning for the *target task*.



Types of transfer:

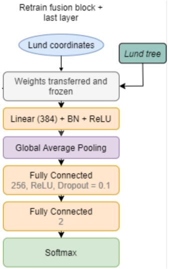
- Top tagging at different p_T cuts.
- W tagger ↔ Top quark tagger.

Exploiting universal features of QCD

- Universality of QCD suggests most information learnt in training process is common to different signals and experimental setups.
- Can use transfer learning to develop fast and data-efficient jet taggers from existing models.

We consider two transfer learning methods:

- Fine-tuning: retrain all weights with a learning rate 10x lower (3x for W)
- Frozen: keeping the EdgeConv frozen and retraining the final dense layers



Transfer learning scenario

Target task:

Discrimination between boosted top quark jets (fully hadronic decay) and QCD backgrounds with a $p_t > 500\,\mathrm{GeV}$ cut.

Source tasks:

- `LundNet-3`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.
- `LundNet-5`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut, W boson tagger with $p_t > 500\,\mathrm{GeV}$ (hadronic decay).
- `ParticleNet`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.

In all cases the jets are defined using the anti-k_t algorithm with $R = 0.8$, and there is a rapidity cut of $|y| < 2.5$; generated with `Pythia` 8.223; 500k signal (top), 500k background (QCD) jets in training data set.

For each source task we try both the fine-tuning and frozen methods.

Transfer learning scenario

Target task:

Discrimination between boosted top quark jets (fully hadronic decay) and QCD backgrounds with a $p_t > 500\,\mathrm{GeV}$ cut.

Source tasks:

- `LundNet-3`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.
- `LundNet-5`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut, W boson tagger with $p_t > 500\,\mathrm{GeV}$ (hadronic decay).
- `ParticleNet`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.

In all cases the jets are defined using the anti-k_t algorithm with $R = 0.8$, and there is a rapidity cut of $|y| < 2.5$; generated with `Pythia` 8.223; 500k signal (top), 500k background (QCD) jets in training data set.

For each source task we try both the fine-tuning and frozen methods.

Transfer learning scenario

Target task:

Discrimination between boosted top quark jets (fully hadronic decay) and QCD backgrounds with a $p_t > 500\,\mathrm{GeV}$ cut.

Source tasks:

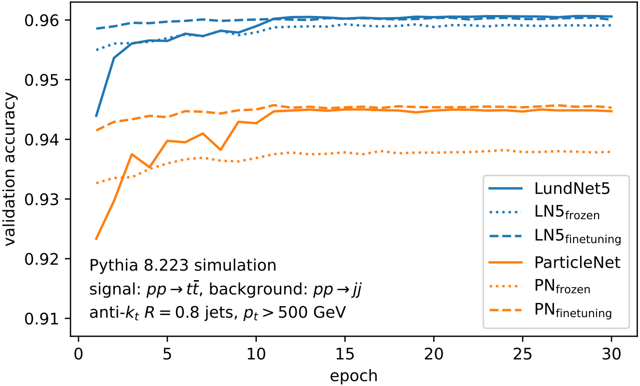
- `LundNet-3`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.
- `LundNet-5`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut, W boson tagger with $p_t > 500\,\mathrm{GeV}$ (hadronic decay).
- `ParticleNet`: top tagger with $p_t > 2\,\mathrm{TeV}$ cut.

In all cases the jets are defined using the anti-k_t algorithm with $R = 0.8$, and there is a rapidity cut of $|y| < 2.5$; generated with `Pythia` 8.223; 500k signal (top), 500k background (QCD) jets in training data set.

For each source task we try both the fine-tuning and frozen methods.

Observation 1: computational complexity of new models

All models were trained for 30 epochs, but fine-tuning allows to decrease the number of epochs significantly with a small impact on performance:



Training data set consists of 500k events for signal and background.

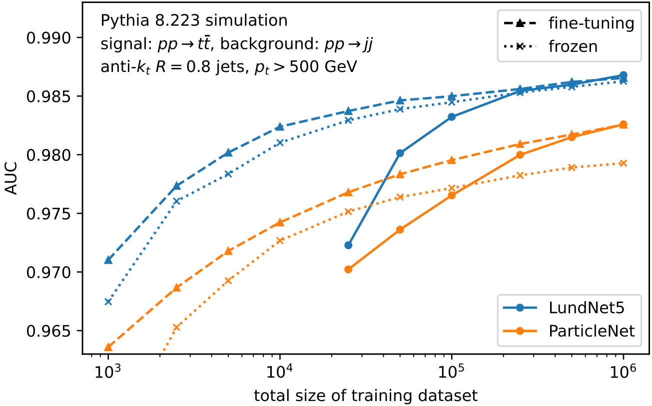
How does decreasing the number of training samples affect performance?

Training data set consists of 500k events for signal and background.

How does decreasing the number of training samples affect performance?

Observation 2: training data set size dependence

We trained models transferred from `LundNet-5` top quark source task and `ParticleNet` source for different dataset sizes.
Both fine-tuning and frozen models show that a dramatic reduction in training data does not affect performance strongly:



By how much does transfer learning decrease the time of training?

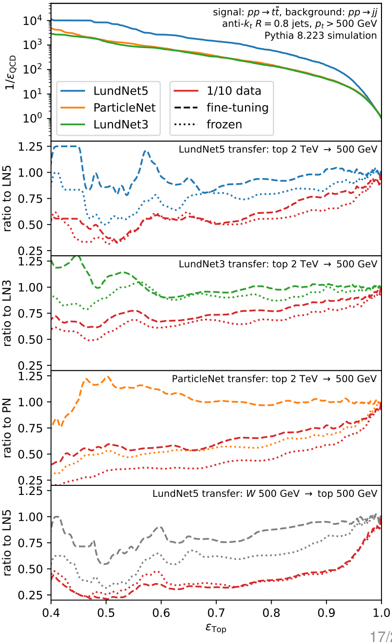
Observation 3: timings

	Training time [ms/sample/epoch]	Total for 10^6 samples [hh:mm:ss]	Total for 10^5 samples [hh:mm:ss]
LundNet5	0.46	03:48:15	00:22:43
$\mathrm{LN5}_{\mathrm{frozen}}$	0.15	01:17:02	00:07:36
$\mathrm{LN5}_{\mathrm{finetuning}}$	0.46	03:48:32	00:22:45
ParticleNet	3.60	30:09:44	02:59:17
$\mathrm{PN}_{\mathrm{frozen}}$	2.16	18:13:21	01:47:37
$\mathrm{PN}_{\mathrm{finetuning}}$	3.60	29:59:46	03:01:04

30 epochs total, NVIDIA GeForce RTX 2080 Ti GPU. Time of preprocessing for `LundNet` is **not included**.

Final observation: tagger performance

- Fine-tuning models have a slight edge over frozen models, but both achieve very high accuracies.
- `LundNet`-frozen models perform relatively better than `ParticleNet`-frozen models.
- Transfer learning between different sources of signal is also successful, see bottom panel in the figure.



Conclusion

- We investigated the ability of `LundNet` and `ParticleNet` to learn the universal features of QCD and transfer them to a different task
- Defined two transfer learning methods, fine-tuning of all weights and retraining dense layers with frozen edge convolutions.

Reliable taggers can be derived from different models with an order of magnitude less data and training time.

https://github.com/fdreyer/lundnet

BACKUP SLIDES

LundNet preprocessing

On a personal CPU, for a format compatible with `ParticleNet` (jet represented as a set of 4-momenta), preprocessing takes 4.6ms per jet.

$$4.6\text{ms} \cdot 10^6 \approx 1 \text{ h } 15 \text{ min}$$

$$4.6\text{ms} \cdot 10^5 \approx 8 \text{ min}$$

ROC curve ratios for all models