SPA-NET: Generalized Permutationless Event Reconstruction with Symmetry Preserving Attention Networks

ML4Jets 2021 "Heidelberg"

arXiv:2010.09206 arXiv:2106.03898 SPA-NET @ GitHub

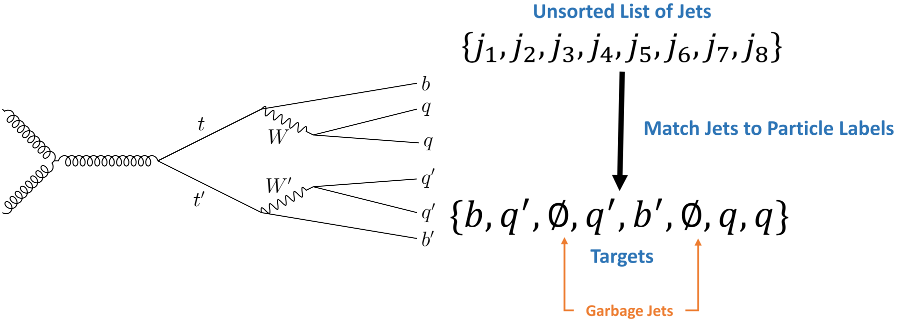
Michael Fenton†, Alexander Shmakov†, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, Pierre Baldi

University of California, Irvine
m.fenton@cern.ch

July 6, 2021

Jet-Parton Matching

- This is an example of a *set assignment* problem: we must take as input a set of objects, and output sub-sets

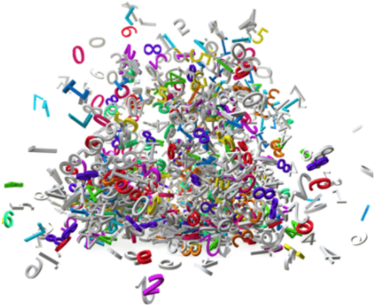


$\rightarrow$ {qqb} and {q'q'b'} subsets are our desired output

> For simplicitly, I will mostly discuss all-hadronic final states, in which all input objects are jets - but this is not a fundamental requirement!

Combinatoric Explosion

- Standard method is to build a "permutation classifier": exhaustively check *every possible permutation* and choose the "best"
  - But this does not scale at all 😕
- Eg $t\bar{t}$:
  - 6j $\rightarrow$ 90 permutations
  - 7j $\rightarrow$ 630 permutations
  - 8j $\rightarrow$ 2520 permutations



- And lets not forget that we often have to do this per event, *per systematic (!)*

$\rightarrow$ "**Combinatoric explosion**" !

Existing Methods

- For all-had $t\bar{t}$, we most commonly use a χ^2 function with known resonance masses (see eg ▸ all-had diff xsec):

$$\chi^2 = \frac{(m_{bqq} - m_t)^2}{\sigma_t^2} + \frac{(m_{b'q'q'} - m_t)^2}{\sigma_t^2} + \frac{(m_{qq} - m_W)^2}{\sigma_W^2} + \frac{(m_{q'q'} - m_W)^2}{\sigma_W^2} \tag{1}$$

- KLFitter does a similar thing, but is even slower and doesn't necessarily outperform χ^2 (at least in all-had)
- Other options include a "Reconstruction" BDT (eg ▸ ttHbb) or DNN (eg ▸ Erdmann et. al)
  - But note that these ML methods are to date only performed in leptonic channels, which have smaller combinatorics

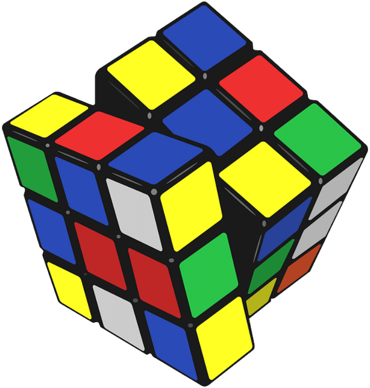
Reducing the Combinatorics

- These methods typically use physics arguments to reduce combinatorics

→ Only consider the leading N jets (ordered in p_T or similar)

→ Consider b-jets and light-jets separately
  - So only a b-jet can be in a b-quark position and vice-versa

- But these types of constraints are not fully efficient; some events become impossible to reconstruct
- And still requires many permutations!



Final states more complex than the 6j $t\bar{t}$ case have mostly been considered intractable until now

Attention Is All You Need ▸ arXiv:1706.03762

- In NLP, Recurrent NNs and LSTMs have been state-of-the-art:
  - Process sentences as a sequence of words to perform translation, prediction, etc.
- "Attention" mechanisms are similar: in a sentence, which words are important context for each other word?
  - Reordering the inputs results in the same reordering of the attention matrices: **permutation invariant**
- Attention based networks have now achieved state-of-the-art performance in many NLP tasks

→ Lets take advantage of this for particle physics!

Utilising Symmetries

- When trying to make something more efficient, a useful question to consider is always; are there symmetries that we can take advantage of?
  - Of course, the answer is YES

Every jet is made equal; we want to consider them without ordering
$$j_1 \leftrightarrow j_2 \leftrightarrow \ldots \leftrightarrow j_N$$

2-body decays (usually) invariant to ordering of decay products
$$W \rightarrow q\bar{q} \leftrightarrow \bar{q}q,\ H \rightarrow b\bar{b} \leftrightarrow \bar{b}b$$

We also don't (usually) care about resonance charge;
$$t \leftrightarrow \bar{t},\ W^+ \leftrightarrow W^-$$

Tensor Attention

´

- We introduce *Tensor Attention*, a generalisation of attention, to encode these symmetries.

- We use the natural permutation invariance of attention to ensure invariance in the inputs $X \in \mathbb{R}^{N \times D}$
  - No more arbitrary p_T ordering!
  - N is the number of jets and D is an arbitrary hyperparameter that defines the size of the latent space representation

- We can encode the symmetry of eg $W \to qq$ (or $H \to bb$, or any other decay with invariance in decay products) by using the NN weights matrices $\theta \in \mathbb{R}^{D \times D \times \ldots \times D}$ to construct:

$$\mathcal{S}^{i_1 i_2 \ldots i_{k_p}} = \sum_{\sigma \in G_p} \Theta^{i_{\sigma(1)} i_{\sigma(2)} \ldots i_{\sigma(k_p)}}$$

$$\mathcal{O}^{j_1 j_2 \ldots j_{k_p}} = X_{i_1}^{j_1} X_{i_2}^{j_2} \ldots X_{i_{p_k}}^{j_{p_k}} \mathcal{S}^{i_1 i_2 \ldots i_{k_p}}$$

  - $\mathcal{O}^{j_1 j_2} = O^{j_2 j_1} \to$ qq invariance!

[1] $\sigma_{(1)}$ represents the symmetry between particles and k_p is the number of decay products from the particle p.
1

Tensor Attention

- Finally, we select the jets from each final state particle by taking the maximum of the k_p-D softmax

$$\mathcal{P}_p^{j_1 j_2 \dots j_{k_p}} = \frac{\exp\left(\mathcal{O}^{j_1 j_2 \dots j_{k_p}}\right)}{\sum_{j_1, j_2, \dots, j_{p_k}} \exp\left(\mathcal{O}^{j_1 j_2 \dots j_{k_p}}\right)}$$

.

- We produce one output per final state particle $\mathcal{P}^1$, $\mathcal{P}^2$, ..., $\mathcal{P}^N$ and embed symmetry between them into the loss function:

$$\mathcal{L} = \min \sum_{i=1}^{m} CE(P_{\sigma_i}, T_{\sigma_i})$$

where CE is cross-entropy and σ represents the symmetry between the final state particles. For example, for $t\bar{t}$:

$$\mathcal{L} = \min\{CE(P_t, T_t) + CE(P_{\bar{t}}, T_{\bar{t}}), CE(P_t, T_{\bar{t}}) + CE(P_{\bar{t}}, T_t)\}$$

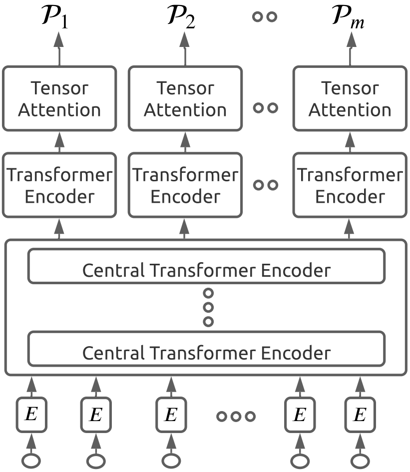
Partial Event Training

- A further complication in training any ML algorithm for event reconstruction is generating sufficient statistics for any training
- The efficiency to generate full events is often low, with one or more partons being lost due to phase space cuts or matching inefficiency
  - eg; Only 32.5% of $t\bar{t}$ events are fully matched; just 6.6% of 4top!
- We add a mask term $\mathcal{M}$ to our loss function to only operate on fully matched particles
  - Fewer events must be generated
  - Improve performance on partial events

$$\mathcal{L}_{min}^{masked} = \min_{\sigma \in G_E} \left(\sum_{i=1}^{m} \frac{\mathcal{M}_{\sigma(i)} CE(\mathcal{P}_i, \mathcal{T}_{\sigma(i)})}{CB\left(\mathcal{M}_{\sigma(1)}, \mathcal{M}_{\sigma(2)}, \dots, \mathcal{M}_{\sigma(m)}\right)} \right)$$

  - where CB is a normalisation to achieve class balance between different event types.

Spa-Net



- Input: **unordered** list of jets
  - We input the full jet 4-vector plus a boolean *b*-tag variable
  - Trivially can include further info, eg substructure, q/g tagging, ...
- Output: one head per particle
  - In the cases where twoheads predict the same jet in their assignments, we keep the more confident output and select the best non-colliding prediction from the other

Benchmark Datasets

- To benchmark performance, we use three test all-hadronic topologies
  - $t\bar{t}$, $t\bar{t}H, H \rightarrow b\bar{b}$, $t\bar{t}+t\bar{t}$
- MG5_aMC@NLO+Pythia8+Delphes (ATLAS card)
- $\geq 6/8/12$ jet w/ $p_T > 25$ GeV, $|\eta| < 2.5$, $\geq$2b-tags
- Exclusive geometric matching of partons to jets with $\Delta R < 0.4$

Benchmark Results : $t\bar{t}$

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	SPA-NET Efficiency Top Quark	χ^2 Efficiency Event	χ^2 Efficiency Top Quark
All Events	$== 6$	0.245	0.643	0.696	0.461	0.523
	$== 7$	0.282	0.601	0.667	0.408	0.476
	≥ 8	0.320	0.528	0.613	0.313	0.395
	Inclusive	**0.848**	**0.586**	**0.653**	**0.387**	**0.457**
Complete Events	$== 6$	0.074	0.803	0.837	0.664	0.696
	$== 7$	0.105	0.667	0.754	0.457	0.556
	≥ 8	0.145	0.521	0.662	0.281	0.429
	Inclusive	**0.325**	**0.633**	**0.732**	**0.426**	**0.532**

- Event Fraction; % of events in that category
- Event Efficiency: % of events in that category that are perfectly reconstructed
- Top Efficiency: % of fully-matched top quarks that are perfectly reconstructed
- All Events : all events containing at least one fully matched particle
- Complete Events : only events where all particles are fully matched

Benchmark Results : $t\bar{t}H$

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	SPA-NET Efficiency Higgs	SPA-NET Efficiency Top	χ^2 Efficiency Event	χ^2 Efficiency Higgs	χ^2 Efficiency Top
All Events	== 8	0.261	0.370	0.497	0.540	0.056	0.193	0.092
	== 9	0.313	0.343	0.492	0.514	0.053	0.160	0.102
	≥ 10	0.313	0.294	0.472	0.473	0.031	0.150	0.056
	Inclusive	**0.972**	**0.330**	**0.485**	**0.502**	**0.045**	**0.164**	**0.081**
Complete Events	== 8	0.042	0.532	0.657	0.663	0.040	0.220	0.135
	== 9	0.070	0.422	0.601	0.596	0.019	0.152	0.079
	≥ 10	0.115	0.306	0.545	0.523	0.004	0.126	0.073
	Inclusive	**0.228**	**0.383**	**0.583**	**0.572**	**0.016**	**0.153**	**0.087**

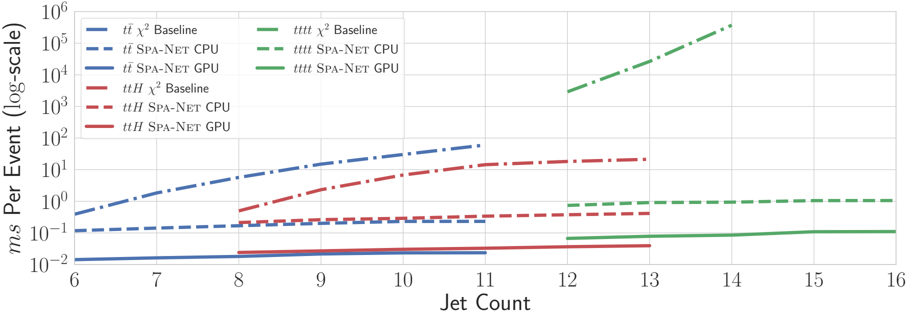
- Full $t\bar{t}H$ reconstruction never performed in all-had in any publication we found
- Trained on events with > 2b-jets, evaluated on > 4b-jet events for consistency with χ^2
  - Efficiencies only a few percent lower when looking at trickier > 2b-jet events

Benchmark Results : 4-top

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	SPA-NET Efficiency Top Quark
All Events	== 12	0.219	0.276	0.484
	== 13	0.304	0.247	0.474
	$\geq$ 14	0.450	0.198	0.450
	Inclusive	**0.974**	**0.231**	**0.464**
Complete Events	== 12	0.005	0.350	0.617
	== 13	0.016	0.249	0.567
	$\geq$ 14	0.044	0.149	0.504
	Inclusive	**0.066**	**0.191**	**0.529**

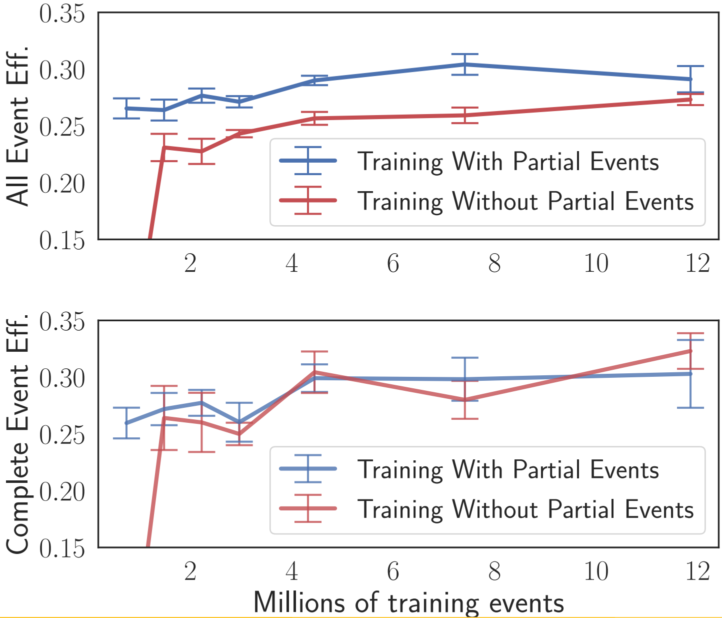
- χ^2 is intractable for this topology!

CPU Time



- The SPA-NET calculation scales only as $\mathcal{O}(N^k)$, where k is the size of the largest resonance group
  - 2 for W, Z, or $H \to b\bar{b}$, 3 for top, 4 for $H \to VV \to qqqq$
- On the other hand, χ^2, KLFitter, or a Reco BDT/NN scale as $\mathcal{O}(N^f)$, with f the total number of partons in the event
  - 6 for $t\bar{t}$, 8 for $t\bar{t}H$, 12 for $t\bar{t}t\bar{t}$
- A further factor 10 is gained by running inference on a GPU

Impact of Partial Event Training



The SPA-NET Package : ▸ GITHUB

- We have released a user friendly package to implement SPA-NET for arbitrary final states
- Simple config file:

```
[SOURCE]
mass = log_normalize
pt = log_normalize
eta = normalize
phi = normalize
btag = none

[EVENT]
particles = (t1, t2)
permutations = [(t1, t2)]

[t1]
jets = (q1, q2, b)
permutations = [(q1, q2)]

[t2]
jets = (q1, q2, b)
permutations = [(q1, q2)]
```

Input features, normalisations

Target topology

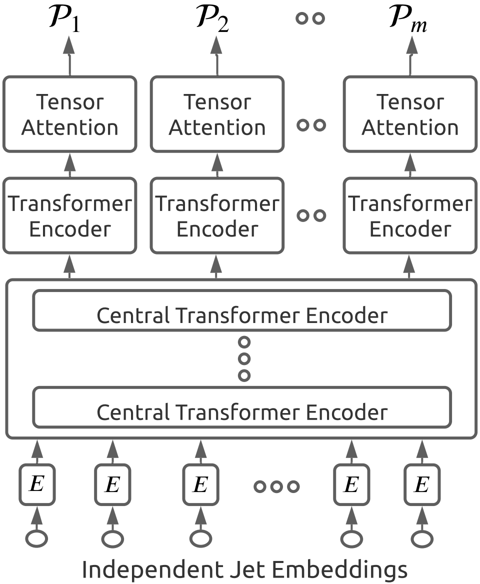
Symmetry between particles

Particle decays and symmetries

- Our three benchmarks are just examples; the applications are not limited to top events or to all-jet events
  - eg 1l $t\bar{t}$ / $t\bar{t}H$ / $t\bar{t}t\bar{t}$, $H \to VV \to qqqq$, VVV, ...

Summary

- We have developed a highly efficient, highly effective event reconstruction technique for arbitrary all-jet final states (and beyond!)
- Improved performance relative to baseline methods
- Previously intractable topologies are now feasible to reconstruct for the first time
- Full details in arXiv:2010.09206 and arXiv:2106.03898
- Code at GitHub



Backup