Finetuning Foundation Models for joint Analysis Optimization

arXiv:2401.13536

Lukas Heinrich, Nicole Hartman, Matthias Vigl

ACAT 2024, Stony Brook University, Stony Brook, Long Island NY, USA - 12 Mar 2024
Analysis pipeline at the LHC

\[ L = - \frac{1}{4} F_{\mu \nu} F^{\mu \nu} + i \bar{\psi} \gamma^\mu \psi + h.c. + \bar{\psi}_i \gamma^\mu \psi_i + h.c. + \partial_\mu \phi^2 - V(\phi) \]

Theory

- **reconstruction** both reduces dimensionality of the data and gives physics interpretable representation (particles)

Raw Data

O(100M) channels!

### Few parameters

- Few parameters

### Summary Statistics

- Analysis

- Result

Lukas Heinrich, Nicole Hartman, Matthias Vigl
Reconstruction

Lots of (also ML) components - e.g. *tracking, jet tagging* (*ftag*)...

But each **optimised separately** and downstream components are optimised based on the steps prior to it
Analysis optimisation

The optimisation of the sensitivity is primarily the job of the analysis, given a fixed reconstruction - mostly common for all analysis.

Is this the best way to do it?

• a fixed reco loses some information irrevocably
• maybe should rather have a specific reco for every analysis?
Reconstruction = Foundation model

- ML and HEP setups are fortunately very aligned
- But everything is differentiable so can be fine-tuned with gradient descent

Pre-trained on a large dataset

Foundation model

Features

Head

Result

Raw Data

Reconstruction

Particles

Analysis

Result

Raw Data

Lukas Heinrich, Nicole Hartman, Matthias Vigl

Pre-trained on a large dataset

fine-tuned on small labeled dataset
Reconstruction = Foundation model

- ML and HEP setups are fortunately very aligned
- But everything is differentiable so can be fine-tuned w/ gradient descent

Key difference: reconstruction is mostly common and Frozen for each downstream task (analysis)

Q: Could this Finetuning workflow also work in HEP?
A toy end-to-end Analysis

\[ X \rightarrow HH \rightarrow b\bar{b}b\bar{b}^{[1]} \]

Final state with Higgs/QCD Jets

[1]: Duarte Javier, CMS open data [http://opendata.cern.ch/record/12102]
A toy end-to-end Analysis

$X \to HH \to b\bar{b}b\bar{b}$\([1]\)

Final state with Higgs/ QCD Jets

Jet representation

Backbone FM

Transformer (ParT)\([2]\)

Xbb

Xbb flavour tagging (supervised)

Jet representation

Jet representation

Jet representation

Jet representation

Jet representation

\[1\]: Duarte Javier, CMS open data [http://opendata.cern.ch/record/12102]

\[2\]: Huilin Qu, Congqiao Li, and Sitian Qian, “Particle Transformer for Jet Tagging,” (2022), arXiv:2202.03772
Q: Do high-dim embeddings hold more (useful) info than Xbb+HL Features?

Analysis would typically use **frozen** Xbb + HL Features (jet 4-momenta)
Analysis head

The head is trained for S/B discrimination with Jet representations from backbone as inputs

Variable number of jets per event + Permutation Invariance -> DeepSets

Q: Does fine-tuning the jet representation help?
Frozen workflow

Backbone trained on Xbb task and then frozen
DeepSets + binary classification trained on S/B

Transformer

Jet representation

DeepSets

Result

Train and freeze

Jet representation

Jet representation

Jet representation

Xbb

S/B

Lukas Heinrich, Nicole Hartman, Matthias Vigl
Fine-tuned workflow

Backbone pre-trained on Xbb task

Then fine-tuned (end-to-end) on S/B
From scratch training

No backbone pre-training

Backbone + head trained from scratch on S/B
Architecture autonomy

- Scalar + HL: Standard HEP
- Vector + HL: ML-assisted HEP
- Vector: ‘Hits to Higgs’

Representational Autonomy:
- Frozen
- Fine-tuned
- From scratch

Structural Autonomy:
Well-known patterns from ML seem to hold also in HEP

- Fine-tuning workflow improves both **performance** & **data efficiency** (10-100x wrt standard hep)

- **Domain adaptation**: Pre-training on a different dataset (JetClass[^3]) helps

Well-known patterns from ML seem to hold also in HEP

• Fine-tuning workflow improves both performance & data efficiency (10-100x wrt standard hep)

• High-dim embeddings also seem to be useful in the frozen case

• Domain adaptation: Pre-training on a different dataset (JetClass[3]) helps

---

Conclusions

1) Fine-tuning workflow for end to end analysis works and is useful even for simple examples

Compared to standard HEP approach:

- 2x in **background rejection**, 10-100x **in data efficiency**
- There might be more to gain in more complex topologies

2) Key question now: what’s the best pre-training (e.g. supervised or self-supervised)?

SSL approaches are also being explored:

- e.g. “**Masked Particle Modeling**”, yesterday, today, and tomorrow talks
- self-supervised training doesn’t need labels: can pre-train on real data!
  - Huge amount of pre-training possible

3) …and calibration

Backup
From scratch training eventually surpasses frozen models, it’s just slow
From scratch training eventually surpasses frozen models, it’s just slow
Xbb is learned when solving the downstream task even without actual jet labels.
High dim embeddings help for frozen jet representations
Dimensionality becomes less important when training end-to-end
Setup: CMS open data and ParT

Jets are clustered using the anti-kT algorithm with R=0.8 from particle flow (PF) candidates

Constituents features:
- up to 100 PF per jet
- 17 features per PF

High-level features:
- Jet 4-momenta
- Xbb scores from ParT

Particle transformer for FTAG [arXiv:2202.03772]

Training: QCD vs Higgs jets

CMS open data: Duarte Javier, [http://opendata.cern.ch/record/12102]
ParT

(a) Particle Transformer

(b) Particle Attention Block

(c) Class Attention Block

[arXiv:2202.03772]
CMS open data

• CMS simulated dataset:
  • Sample with jet, track and secondary vertex properties for H(bb) tagging ([http://opendata.cern.ch/record/12102](http://opendata.cern.ch/record/12102))
  • Sample with jet, track and secondary vertex properties for H(bb) tagging (http://opendata.cern.ch/record/12102)
  • meant for jet tagging, up to 100 pf cand per jet - 17 feats each
  • signal samples: 11 mass points - M_x from 600 GeV to 4500 GeV, bkg: QCD multijet
  • 'fat jets' (fj) 4-momenta and (old) Xbb score