Fast Inference for Machine Learning in ROOT/TMVA

Kim Albertsson, Sitong An [4], Lorenzo Moneta, Stefan Wunsch, Luca Zampieri
EP-SFT CERN

https://root.cern

Our overview talk **Machine Learning with ROOT/TMVA**
Monday 4 Nov 11:15 at Hall G (Track 6) by Stefan Wunsch

TMVA in modern ML ecosystem

- TMVA: Toolkit for Multivariate Analysis
  - Root-integrated Machine Learning environment
  - Supports the training and evaluation of a variety of Machine Learning algorithms since 2005
- ML Landscape is fast evolving
  - Development of Deep Learning
  - Maturation of powerful DL frameworks outside of HEP
- TMVA evolves to better serve ML-HEP community
- Focus on a **fast and robust ML inference system** for easy deployment of trained ML/DL models to HEP data in production

1. Fast Inference Engine for Decision Trees

ROOT 6.20 Experimental

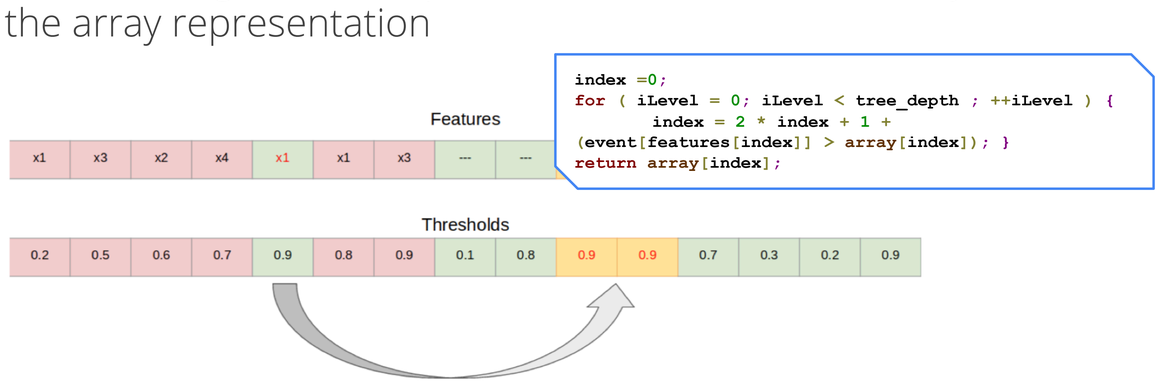
- Initiated as CERN Summer Student Project by Luca Zampieri [1]
- Decision-tree ML algorithms widely popular in HEP and in data science
- Special need for application in HEP
  - Low-latency inference critical for some use cases i.e. HLT
  - Focus on event-loop inference rather than batch inference

➢ Just-in-time compilation

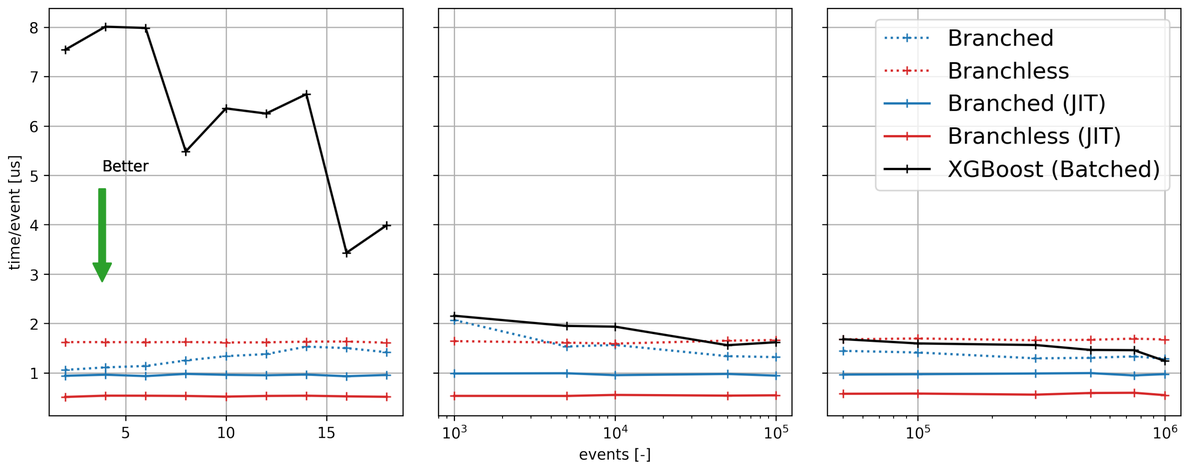
- With Cling [2], the interactive compiler in ROOT
- Compiles hard-coded evaluation logic parsed from the model
- This allows us to exploit compiler optimization dynamically

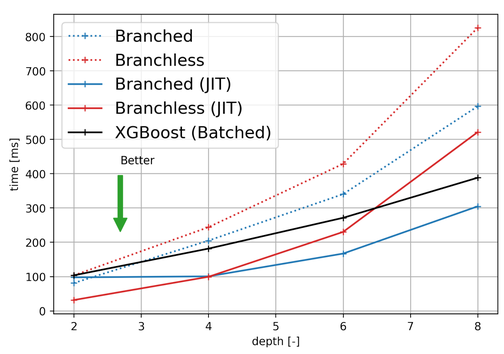
➢ "Branchless" representation of trees

- Unroll the tree into a long array with topological ordering
- Fill in missing values in sparse trees to create full binary trees in the array representation



- Tree traversal is now a maths operation - cheaper than if branch

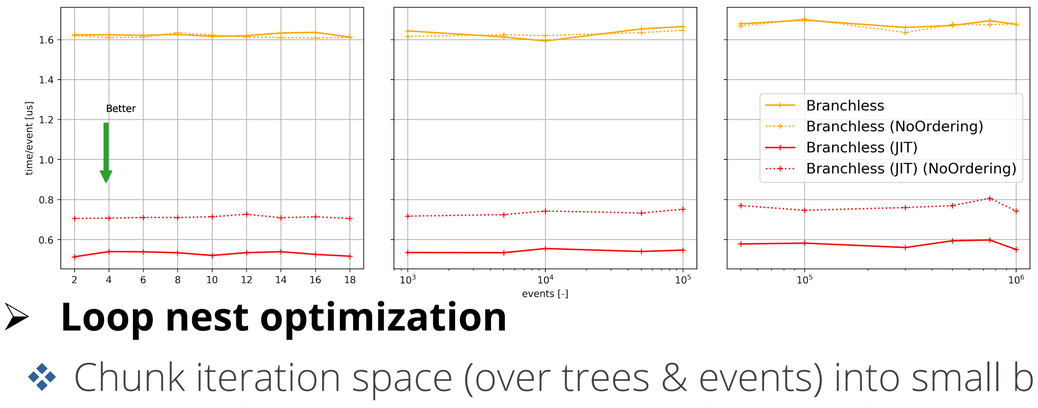




- Branchless implementation assumes shallow, nearly-full trees
- Most decision-tree ML algorithms produce these
- Branched implementation to be integrated

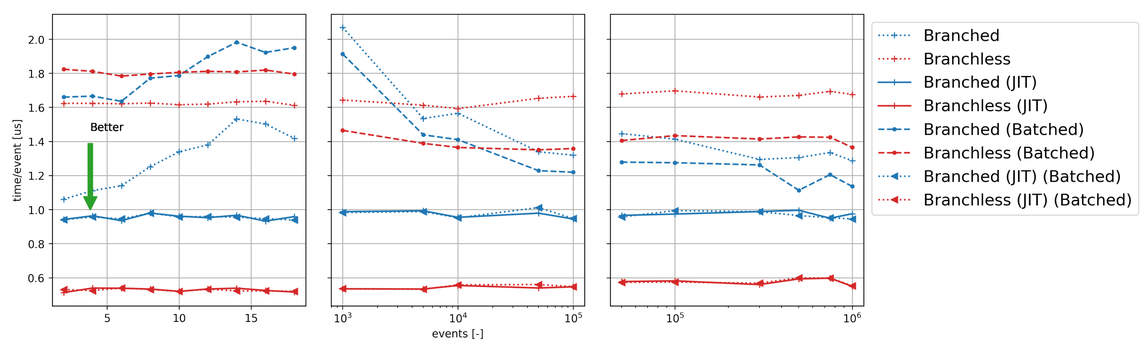
➢ Tree ordering

- Order trees by feature & cut value of first (root) node
- Improve dynamic branch prediction, reduces branch misses



➢ Loop nest optimization

- Chunk iteration space (over trees & events) into small blocks
- Improve data/instruction locality, reduces cache misses



- Loop nest optimization improves non-JITted implementations only

```
for(i=0; i<num_events; i+=batch_size)
  for (auto tree: trees)
    for(j=0; j<batch_size; j++)
      tree.inference(events[i+j]);
```

2. Inference of ONNX Deep Learning Models

Work in Progress

- **ONNX** [3] is an open format for DL models
  - Supports most popular DL operators/layers
  - Convertors available from major DL framework to ONNX
- **ONNX runtime**: an open source inference engine
  - Supports by industry in fast development
  - Highly optimized for low-latency inference
  - Multiple backends and optimization methods supported

Development in TMVA: an inference interface for ONNX models, designed for HEP applications

✓ **ONNX operator-based infrastructure**
- ONNX model exploration and manipulation
- Allow potential customized optimization

❑ **Interface with open-source ONNX runtime**
- Convenient interoperability with ROOT data
- Support for implicit multi-threading inference

❑ **Code generation from ONNX model**

- Generate inference script from model for target backend
- No external dependency on ONNX runtime
- Allow easy integration of DL Algorithms into existing C++ analysis frameworks

Working on ML/DL for High Energy Physics?
Help us to help you - come and talk to us about what **you** would like to see in ROOT/TMVA!

Acknowledgement and References

1. L. Zampieri, 2019, *Fast inference engine for Decision Trees,* CERN-STUDENTS-Note-2019-183
2. V Vasilev, Ph Canal, A Naumann, and P Russo, 2012, *Cling–The New Interactive Interpreter for ROOT 6. In Journal of Physics: Conference Series, Vol. 396.*
3. ONNX., 2019, *Open Neural Network Exchange,* https://onnx. ai/
4. S.An. gratefully acknowledges the support of the Marie Sklodowska-Curie Innovative Training Network Fellowship of the European Commission Horizon 2020 Programme, under contract number 765710 INSIGHTS. https://insights-itn.eu/

s.an@cern.ch/luca.zampieri@epfl.ch