DIANA Fellowship Proposal

Matthew Feickert

1 Project description

Within only a couple of years, the rapid development of software libraries for numerical computations through data flow graphs (e.g., TensorFlow [1], Theano [2], and MXNet [3]) has led to a fundamental change of paradigm in machine learning software. These libraries are designed around the concept that a numerical program can often equivalently be expressed as a graph, where nodes represent mathematical operations and edges represent the data communicated between them. Most notably, these libraries allow one to automatically deploy computation over one or more CPUs or GPUs within a single API, which makes it easy to maximize performance without specialized software expertise.

While these frameworks have originally been developed for the purpose of deep learning research, they are usually general enough to be applicable in a wide variety of other domains. For this reason, the objective of this DIANA project is to conduct a feasibility study to answer whether statistical models used in particle physics (typically built with RooFit [4] and HistFactory [5]) could equivalently be expressed as computational graphs, to assess their capabilities and limits, and to determine how those frameworks would scale in terms of data and model parallelism. In addition, the study should also determine whether existing probabilistic programming frameworks based data flow graphs (e.g., Edward [6] and tensorprob [7]) are applicable for particle physics statistical models. Where appropriate, the study should finally identify shortcomings, on which further software efforts could be dedicated.

2 Roadmap

- April — May, 2017
  - Design of representative benchmarks models in RooFit.
  - Study of data flow graphs frameworks (e.g. TensorFlow, Theano, MXNet) or of probabilistic programming frameworks (e.g. Edward, tensorprob).
  - Implementation of the benchmark models in one of the studied frameworks.
  - Benchmarks evaluating their data and model parallelism.
- June, 2017
  - Technical report and recommendations for particle physics use cases.
  - Implementation of the study and benchmark models in a secondary framework (if time permits).

- July — August, 2017
  - Upstream software contributions to address the identified limitations (if any).
  - Development of/Contribution to a probabilistic framework (if time permits).

3 Deliverables

- Establish benchmarks that are representative of particle physics use cases.
- Technical report investigating the potential and limits of data flow graphs frameworks for particle physics statistical models.
- Contributions to a probabilistic framework (if time permits).
- Tutorial on data flow graphs targeted for physicists.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

[2] Theano Development Team collaboration, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas et al., *Theano: A Python framework for fast computation of mathematical expressions*, *arXiv e-prints* **abs/1605.02688** (May, 2016) .

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang et al., *Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems*, *CoRR* **abs/1512.01274** (2015) .

[4] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, *eConf* **C0303241** (2003) MOLT007, [physics/0306116].

[5] ROOT collaboration, K. Cranmer, G. Lewis, L. Moneta, A. Shibata and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, .

[6] D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang and D. M. Blei, *Edward: A library for probabilistic modeling, inference, and criticism*, *arXiv preprint arXiv:1610.09787* (2016) .

[7] I. Babuschkin, "tensorprob: A probabilistic programming framework based on TensorFlow." https://github.com/tensorprob/tensorprob, 2016.