DIANA Fellowship Proposal

Matthew Feickert

Abstract

The statistical software used for the last several years in high energy physics is facing scalability challenges. In addition to processing speed, which is being addressed with GPU-based fitting approaches, we also face memory limitations as the combined statistical models grow in size. Thus, it is critical to investigate more distributed models for these computations.

The rapid development of software libraries for numerical computations through data flow graphs (e.g., `TensorFlow` and `Theano`) 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. While originally developed for the purpose of deep learning research, they are general enough to be applicable in a wide variety of other domains. Under the mentorship of Gilles Louppe and Vince Croft, Matthew Feickert will conduct a feasibility study to answer whether statistical models used in particle physics could equivalently be expressed as computational graphs, assess their capabilities and limits, and determine how those frameworks would scale in terms of data and model parallelism.

1 Project Description

One of the focus areas for DIANA [1] is to "establish infrastructure for a higher-level of collaborative analysis, building on the successful patterns used for the Higgs boson discovery". A large component of this focus is statistical software. `RooFit` [2] is one of the primary tools used now but it is facing scalability challenges. In addition to processing speed, which is being addressed with GPU-based fitting approaches, we also face memory limitations as the combined statistical models grow in size. Thus, it is critical to investigate more distributed models.

Within only a couple of years, the rapid development of software libraries for numerical computations through data flow graphs (e.g., `TensorFlow` [3], `Theano` [4], and `MXNet` [5]) 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, conducted by Matthew Feickert under the mentorship of Gilles Louppe and Vince Croft, is to conduct a feasibility study to answer whether statistical models used in particle physics (typically built with `RooFit` and `HistFactory` [6]) 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` [7] and `tensorprob` [8]) are applicable for particle physics statistical models. Where appropriate, the study should finally identify shortcomings on which further software efforts could be dedicated.

Chien-Chin Huang is a computer science PhD student supported via DIANA. He is investigating the model and data parallelism in systems such as `TensorFlow`. A bottle neck in the current work is a set of benchmark physics problems that he can use for these scalability tests. This DIANA fellowship would help remove that bottleneck and accelerate work to connect other DIANA projects like `Histogrammar` [9].

2 Deliverables

- A public GitHub repository with the benchmark model code and instructions for use.
- A document providing the implementation-independent definition of the binned benchmark template that are representative of particle physics use cases.
- Scripts preparing the benchmarks using `HistFactory` to establish `RooFit` benchmark.
- A technical report investigating the potential and limits of data flow graphs frameworks for particle physics statistical models.
- A tutorial on data flow graphs targeted for physicists.
- Contributions to a probabilistic framework (if time permits).

3 Roadmap

- April — May, 2017
  - Design of representative benchmarks models:
    - A template for binned models that is parametrized in terms of number of events, number of bins, number of channels, number of signal/background components, number of parameters of interest and nuisance parameters.
    - Establish a precise mathematical formulation that is implementation-independent. This will be based on the `HistFactory` schema.
    - Document the template model.
  - Implement this template for the benchmark models with a `HistFactory` script (probably in `Python`). This will provide the `RooFit` benchmark.
  - Study of data flow graphs frameworks (e.g., `TensorFlow`, `Theano`, `MXNet`) or of probabilistic programming frameworks (e.g., `Edward`, `tensorprob`).
- June, 2017
  - Implementation of the benchmark models in one of the studied frameworks.
  - Benchmarks evaluating their data and model parallelism.
- July — August, 2017
  - Technical report and recommendations for particle physics use cases.
  - Upstream software contributions to address the identified limitations (if any).
  - Development of/Contribution to a probabilistic framework (if time permits).

4 Proposed Timeline

- May 1st through May 14th:
  - ☐ Create a GitHub repository for the project.
  - ☐ Design a template for binned models that is parametrized in terms of number of events, number of bins, number of channels, number of signal/background components, number of parameters of interest and nuisance parameters.
  - ☐ Begin documentation of the template model on the GitHub repository.
- May 15th through May 28th:
  - ☐ Establish a precise mathematical formulation that is implementation-independent based on the `HistFactory` schema.
  - ☐ Begin study of data flow graphs frameworks or of probabilistic programming frameworks.
  - ☐ Implement the template for the benchmark models with a `HistFactory` script.
- May 29th through June 11th:
  - ☐ Conclude from study of data flow graphs frameworks or of probabilistic programming frameworks which framework to pursue.
  - ☐ Begin implementation of the benchmark models in the selected framework.
  - ☐ Begin to write technical report.
- June 12th through June 26th:
  - ☐ Finish implementation of the benchmark models in the selected framework.
  - ☐ Apply benchmarks evaluating framework data and model parallelism.
- June 27th through July 10th:
  - ☐ Finish technical report.
  - ☐ Create tutorial on data flow graphs targeted for physicists.
- July 11th through July 31st:
  - ☐ Upstream software contributions to address the identified limitations (if any).
  - ☐ Development of/Contribution to a probabilistic framework (if time permits).

References

[1] P. Elmer, K. Cranmer, M. Sokolof and B. Bockelman, *Data-Intensive Analysis for High Energy Physics (DIANA/HEP)*, June, 2014.

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

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

[4] 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) .

[5] 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) .

[6] 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*, .

[7] 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) .

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

[9] J. Pivarski and A. Svyatkovskiyi, "Histogrammar." https://github.com/histogrammar, 2017.