SOW

Proposal of the research topics of collaboration with Intel DAAL team.

1. Harp model computation w.r.t. Tensorflow and preliminary results

Research on the design of computing engine for large-scale learning system is one of the core research directions of our group.
The work starts with selecting a representative algorithm with high impacts, parallelizing the algorithm, and comparing its design and implementation among different state-of-the-art systems. All these lead us to a deeper understanding of the characteristics of the algorithm and the corresponding system needed.

Tensorflow is emerging as a popular framework that successfully keeps growing its community during the last two years. Based on the concept of computation graph, it provides a general interface to let the user to build complex models, and at the same time, it encapsulates all the complexity of communications, synchronizations and parallelism among different devices inside the execution engine. End users can focus on the modeling part and get the computation executed automatically on multiple devices.
Tensorflow comes with strong support for deep learning and later claims to support more general machine learning algorithms. Many new learning algorithms are implemented into the latest release. However, how efficient the existing execution engine is for the general machine learning algorithms is still a question mark.

Gradient Boosting Tree is one widely used algorithm for regression and classification problems. It keeps to be the top 1 algorithm among the winners in the kaggle contests. In the last two years, many endeavors have been put to improve its performance and scalability in academy, see Xgboost[1], TFBT[2], LightGBM[3], DimBoost[4]. Intel DAAl also adds it as one of the new algorithm in the latest release, following the implementation of [1]. With optimization, the DAAL GBT kernel gains 1.2x to 9.5x speedup than Xgboost in one Xeon Platinum server[5].
Current DAAL GBT kernel only supports batch mode. Extensions to support distributed model could be an useful task. And the different existing implementations provide a good testbed to the research on the design of learning system. As Harp/HarpDAAL proposes using synchronized computation model for learning algorithms, which is quite different to the popular existing parameter server based systems, including Tensorflow(TFBT), Microsoft DMTK(LightGBM), Tencent Angel(DimBoost), a new implementation of GBT on HARP can be a promising task.

[1]T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794

[2]N. Ponomareva et al., “TF Boosted Trees: A Scalable TensorFlow Based Framework for Gradient Boosting,” in Machine Learning and Knowledge Discovery in Databases, 2017, pp. 423–427.
[3]G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3149–3157.
[4]J. Jiang, B. Cui, C. Zhang, and F. Fu, “DimBoost: Boosting Gradient Boosting Decision Tree to Higher Dimensions,” in Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 1363–1376.
[5] Optimized Intel Mathematical libraries for HPC and Machine Learning, https://www.dislab.org/GraphHPC-2018/slides/GraphHPC-2018_3_Israfilov_Optimized-Intel-Mathematical-libraries-for-HPC-and-Machine-Learning_ru.pdf, Access on 07/20/2018

2. IndyCar work and preliminary results

Research on real-time machine learning algorithms and system is another topic of our group, and in this direction, the application driven approach is critical.

Now we are building collaboration with the IndyCar company. The IndyCar Series is the premier level of open-wheel racing in North America. Computing System and Data analytics is critical to the game, both in improving the performance of the team to make it faster and in helping the race control to make it safer. They need a system which is capable of detect anomaly events during the game in a real-time fashion, then preventive action can be done to potentially improves the safety and performance of the game.

Anomaly detection is a heavily studied area of data science and machine learning. It refers to the problem of finding patterns in data that do not conform to expected behavior [4]. Detection of anomalies, especially temporally in real-time streaming data, has significant importance to a wide variety of application domains, as it can give actionable information in critical scenarios. In such streaming applications, data are observed sequentially and the processing must be done in an online fashion, i.e., the algorithm can not rely on any look-ahead procedures. Real-time requirements also bring stringent computational constraints, and the phenomenon of concept drift, the statistics of incoming data change over time. All these issues make the problem very challenging,[11].
Traditional time-series modeling and forecasting models can be utilized to detect temporal anomalies. Approaches based on ARIMA are capable and effective for data seasonal patterns [3, 8]. Techniques based on relative entropy[10], graph[2, 6] are also utilized to detect temporal anomalies. Another mainstream approach is to build simulation model with explicit domain knowledge for domain-specific applications, such as in [12, 9]. However, model-based approaches are limited for lack of generalizability. A novel approach was proposed in [1] to use Hierarchical Temporal Memory(HTM)[7, 5] networks to robustly detect anomalies on real-time data streams. HTM networks have some important features for this application. It is an online algorithm, it has good prediction performance, and it is capable to

continuously lean and model the spatial and temporal characteristics of the inputs, which is promising to resolve the problem of concept drifting. HTM based approach becomes to be one the state-of-the-art real-time anomaly detectors.

Intel DAAL has some outlier detection kernels supporting batch processing. Considering to add the HTM based online anomaly detection algorithm should be an interesting task. Furthermore, as HTM is a special kind of neural network on top of sparse data representation and operations, CPU architecture might be more efficient to support it than GPU platforms. In this way, it is one interesting instance of the research topic on helping people to select appropriate computing architecture according to the characteristics of different learning algorithms.

[1] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262:134–147, Nov. 2017.
[2] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey.Data Mining and Knowledge Discovery, 29(3):626–688, 2015.
[3] A. M. Bianco, M. G. Ben, E. J. Martnez, and V. J. Yohai. Outlier Detection in Regression Models with ARIMA Errors using Robust Estimates. Journal of Forecasting, 20(8):565–579.
[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. ACM Comput. Surv., 41(3):15:1–15:58, July 2009.
[5] Y. Cui, S. Ahmad, and J. Hawkins. Continuous online sequence learning with an unsupervised neural network model. Neural computation, 28(11):2474–2504, 2016.
[6] S. Guha, N. Mishra, G. Roy, and O. Schrijvers. Robust random cut forest based anomaly detection on streams. In International Conference on Machine Learning, pages 2712–2721, 2016.
[7] J. Hawkins and S. Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. Frontiers in neural circuits, 10:23, 2016.