Aspects of Large Scale Machine Learning at Facebook

Alisson Gusatti Azzolini
azzolini@fb.com

Parv Oberoi
parvoberoi@fb.com

Stuart Bowers
sbowers@fb.com

Hussein Mehanna
hussein@fb.com

Abstract

FBLearner is Facebook's Machine Learning Platform, used by over a dozen teams including Search, Ads and News Feed to train models aiming to deliver more relevant content to users. Tens of thousands of models are trained every week, using trillions of training examples. The system spans different areas, including ML infrastructure, algorithms and applications built on top of the platform. We present here *1)* a set of techniques used to enable training on large datasets of gradient boosted decision trees [1], of one of the platform's most popular models, and *2)* an application built on top of the platform that allows for automated training and feature selection.

1 Introduction

Facebook uses machine learning to improve numerous experiences across our services. To make this possible we leverage a number of technologies ranging from distributed data processing tools to open source machine learning libraries. One particularly exciting piece of technology that we've built to enable machine learning at Facebook is a platform to train and deploy models called FBLearner. FBLearner is currently operating over petabytes of data to train models for News Feed, Search and Ads. Two important factors that have contributed to FBLearner's success are scalable infrastructure and automated training tools.

2 Infrastructure Challenges

An important consideration when training is the preprocessing of data including filtering, augmenting with new features, and annotating events with labels. Here FBLearner heavily leverages Facebook's data warehouse [2], which operates at scales in excess of 300PB. Once data is preprocessed it is common to train on the same data numerous times, experimenting with both features and models. To make this both efficient and performant FBLearner leverages a local data cache implemented as a distributed file system.

FBLearner is capable of training numerous model types as well as calling out to distributed learning infrastructure. One particular model type which has proven very effective is the boosted decision tree algorithm, which is notoriously hard to parallelize. As shown in figure 1, we take a scatter-gather parallelization approach, enabling feature engineers to manually create a top level decision tree that partitions the data into disjoint segments. This splitting is usually done on high importance features that have a clear human interpretation. A separate decision tree model is trained on each data partition and we refer to these resulting models as sub-models. These submodels are accumulated into a single larger model and frequently consumed as features in a large linear model.

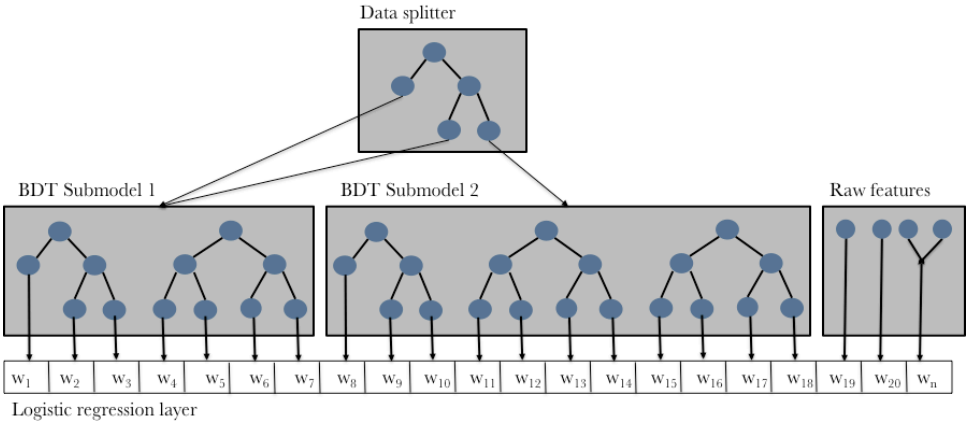

Figure 1. Scatter-Gather approach for training GBDTs on large data

3 Automated Training

Speed of experimentation has proven to be extremely important for long term improvements to model accuracy. To keep pace high many teams have previously set up a weekly rotation in which a single team member would gather new features generated by the team that week, check each for total coverage and quality, then train a model with all available features. This new model can be used to evaluate feature importance in the context of prediction [3], and armed with this feature importance a new set of models can be trained which include only the top k most important features, at various values of k. Using FBLearner's API we have exposed this workflow as a recurring training option allowing team members to see the final models side by side in a dedicated UI and to compare the model accuracy and complexity.

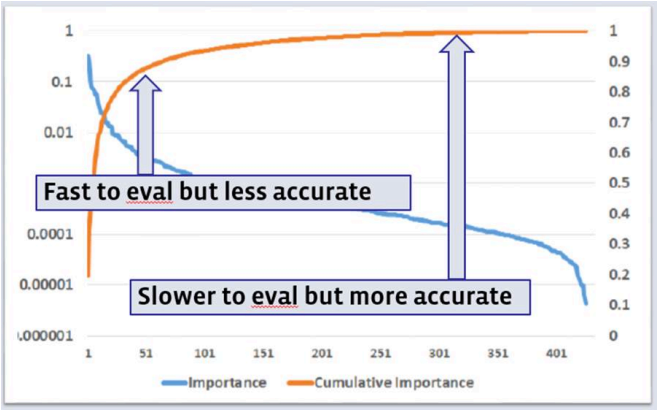

Figure 2. Automated training

References

[1] J.H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29:1189-1232, 1999.

[2] A. Thusoo, S. Antony, N. Jain, R. Murthy, Z. Shao, D. Borthakur, J. Sarma, and H. Liu. Data Warehousing and Analytics Infrastructure at Facebook. In SIGMOD, pages 1013,1020, 2010.

[3] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, J. Candela. Practical Lessons from Predicting Clicks on Ads at Facebook.