1 Paper Overview

"Clipper: A Low-Latency Online Prediction Serving System", written by Daniel Crankshaw, Xin Wang, et al. is a system paper in arXiv'17. Machine learning is being deployed in a growing number of applications which demand real-time, accurate, and robust predictions under heavy query load. Within this paper, a general-purpose low-latency prediction serving system, Clipper, is proposed. This system servers as the intermediate for end-users and various machine learning frameworks. With supported caching, batching and adaptive model selection, this system fulfills the requirement of online machine learning framework serving with low latency and robust predictions [2]. In general, the contributions of this paper is summarized as follows:

1. The modular architecture for simplified model deployment is introduced.
2. Several enhancements, such as caching, batching, and adaptive model selection techniques, are included, which reduce latency and improve prediction for both throughput and accuracy.
3. The system evaluation is performed on four common machine learning benchmark and a comparison between clipper and TensorFlow serving are completed.

1.1 Problem Summary

The focus of this system is unique for the time it published. Focusing on model deployment and serving, there are several possible extensions in terms of batching, caching and personalized recommendation, which compensates the shortage of other serving system at that time. [2].

1.2 Related Works

- Clipper: A low-latency online prediction serving system [2].
- Tensorflow: A system for large-scale machine learning [1].

2 Paper Strengths

The focus of this system is unique for the time it published. Focusing on model deployment and serving, there are several extensions in terms of batching, caching and personalized recommendation, which compensates the shortage of other serving system at that time. The challenge-solution writing style is easy for reader to identify their contributions.

1. The focus of this system is rarely explored.
2. The contribution of this paper is illustrated by comparison with other system.

3 Paper Weaknesses

It is a great effort for heterogeneous model deployment, but this model deployment might not be necessary. Different machine learning platform originally targets at solving different problems, i.e. TensorFlow is for neural network and HKT is for speech recognition. There is no need for developers to develop system in different platform and perform integration. The popularity of this system reflects that one main platform might be enough for problem address. And there are possible some specific optimization for uniform eco-system, such as TensorFlow model and TensorFlow Serving. And there are improvement in prevailing machine learning system, such as TensorFlow serving. TensorFlow currently enables dynamic model query. Although it is interesting to the dynamic control of batch size, this technique is not unique but a borrowed notion from control theory. The combining of multiple models might also introduce internal computation burden, comparing with one model. There is extra effort to perform model selection by introducing weights. In order to train the weighted scale (Exp3 and Exp4), multiple models require to run simultaneously, which is possible a heavy burden, comparing with data-intensive A/B testing. Moreover, since all the evaluation are performed

in a single server and present large-scale machine learning systems are in distributed manner, there is possible scaling and communication concerns for this system. From the experiments, the performance of TensorFlow Serving are better than this proposed system.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J Franklin, Joseph E Gonzalez, and Ion Stoica. Clipper: A low-latency online prediction serving system. In *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, pages 613–627, 2017.