tensorflow-tracing: A Performance Tuning Framework for Production

Sayed Hadi Hashemi, University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications; Paul Rausch; Benjamin Rabe, University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications; Kuan-Yen Chou, University of Illinois at Urbana-Champaign; Simeng Liu, University of Illinois at Urbana-Champaign and National Center for Supercomputing Applications; Volodymyr Kindratenko, National Center for Supercomputing Applications; Roy H Campbell, University of Illinois at Urbana-Champaign

https://www.usenix.org/conference/opml19/presentation/hashemi

This paper is included in the Proceedings of the 2019 USENIX Conference on Operational Machine Learning (OpML ’19).

May 20, 2019 • Santa Clara, CA, USA

ISBN 978-1-939133-00-7

Open access to the Proceedings of the 2019 USENIX Conference on Operational Machine Learning is sponsored by USENIX.
1 Introduction

The growing popularity of Deep Neural Networks (DNN) within the mainstream [8] has had a rapid transformative effect on clusters and data centers. DNN training jobs are becoming one of the largest tenants within clusters, and often take hours to weeks to complete; and even a slight performance improvement can save substantial runtime costs. Despite this fact, the DNN specific performance tuning tools are yet to keep up with the needs of the new changes in production environments.

On one hand, the existing application-agnostic resource-level tools such as top, Nvidia Nsight (for GPU utilization), IPM (for MPI network monitoring) are too limited to predict or explain the behavior and performance of a job accurately. In DNN applications, there exists a complex relationship among resources. Even though measuring coarse metrics such as bandwidth, latency, and GPU/CPU utilization can draw an overall picture of cluster performance, these metrics are not easily translatable to application-level metrics and do not provide actionable insights on how to handle performance bottlenecks.

On the other hand, the short list of application-aware tools, such as MLModelScope [6], TensorBoard [1], and tf.RunOptions [2], while able to provide actionable insights, are mainly designed for the need of application developers and are not intended for production use. Such tools require substantial modification to applications, and early planning as to what, when and how data should be collected.

In this article, we introduce tensorflow-tracing to fill the gap between these two classes of performance tuning tools. To achieve this goal, tensorflow-tracing addresses the following technical challenges:

- Collecting the application-level runtime metrics, such as the timing of each operation or the iteration time, needs explicitly expressed in the training job source code. To makes it possible to trace ML jobs without requiring any application modification, tensorflow-tracing monkeypatches the tensorflow library at the system level.
- Collecting some metrics is expensive and have a significant overhead on the runtime. tensorflow-tracing treats metrics differently; it collects low-overhead metrics automatically, while expensive ones are collected on demand through an admin interface.
- There is no easy way to exchange runtime metrics among users and admins — tensorflow-tracing facilitates this through a portable file format and supporting tools to explore these metrics offline.

The tensorflow-tracing is publicly available under Apache-2.0 license1. It supports native TensorFlow [3], Horovod [7], and IBM PowerAI [5] applications.

---

1 https://github.com/xldrx/tensorflow-tracer

1 University of Illinois at Urbana-Champaign
2 National Center for Supercomputing Applications
This material is based on work supported by the National Science Foundation under Grant No. 1725729.
2 Design and Implementation

Figure 1 shows the building blocks of tensorflow-tracing:

**MonkeyPatching** In order to provide tracing without code modification, tensorflow-tracing injects a proxy function to the tensorflow library using a monkey-patching scheme to intercepts the calls to certain functions and redirects them to the Data Management module. While monkeypatching the library at the system-level automatically enables tracing for any DNN application, tensorflow-tracing also supports per application patching.

**Data Management** This module is responsible for collecting profiling and tracing data as well as making online decisions as to whether a task should be traced\(^2\). This module is also responsible for serializing/deserializing tracing sessions from/to a file.

**REST/Web Interface** This interface is the main portal for interacting with the system. tensorflow-tracing starts a web server whenever an application is executed which is accessible either through a web browser or a REST API client (possibly from terminal). The interface provides two logical views:

1. **Main Interface** shows the list of tasks and their associated profiling/tracing data. This interface allows request tracing. (Figure 2)
2. **Timeline Interface** visualizes an instance of a task trace as a series of timelines, one for every resources (e.g. CPU, GPU, Network Interface) on each machine. Each box represent an operation in the DataFlow DAG of DNN application. (Figure 3)

**CLI** It loads a tracing sessions offline and enables exploring through a web interface.

3 tensorflow-tracing in action

**Overhead** We observe no performance hit on collecting low-overhead metrics such as iteration times, ‘session.run’ call names and frequencies. We observe less than 3% runtime overhead to iteration time when individual operations in a call are traced. CPU Memory requirements varies for different models. For example: an Inception v3 [9] trace consumes 718KB while next_frame_sv2p [4] consumes 2.4MB.

**Case Study** We have used tensorflow-tracing on different workloads to find the performance issues on application, framework, and infrastructure level. For example, our work TicTac ?? addresses the communication timing issue we found in the tracing of a distributed TensorFlow job with parameter server.

4 Limitation and Future Work

The correctness of the tensorflow-tracing’s distributed traces relies on the precision of the clocks on the different machines. Currently it relies on external sources to synchronize the clocks.

tensorflow-tracing traces the network activities just in the user space. This will miss the events such as packet drops or queue latencies. We are planning to expand this capability by adding network stack events from kernel space.

---

\(^2\)MonitoredSession.run function calls
References


