Automation for distributing computations in tensorflow

Łukasz Dudziak

The idea

- Currently – distributing is done manually by a programmer
  - Recent article from Google about using RL for device placement
  - What about many-machines systems?

- Try to find a way to automatically distribute computations
  - NOT using RL (not suitable for mini-project)
  - Focus on neural networks training

What to do?

- Investigate scalability of TF
  - Overhead of network communication (GRPC)
  - When it is reasonable to split computations between multiple machines?

- How to split computations?
  - Develop a policy/heuristic to do that
  - Trade-offs between quality of the placement and latency of calculating it

- Compare to other frameworks

What has been done?

- Nothing yet. ☺

The goal.

- To have a low-overhead mechanism to automatically distribute computations.
  - At least as good as naïve version (don't distribute if overhead would kill performance)
  - Suitable when you don't want to wait ~1 day for an optimal placement produced by RL (suboptimal is fine)

The end.