A zealous parallel gradient descent algorithm

Gilles Louppe and Pierre Geurts
Department of EE and CS, University of Liège, Belgium

Abstract

Parallel and distributed algorithms have become a necessity in modern machine learning tasks. In this work, we focus on parallel asynchronous gradient descent and propose a zealous variant that minimizes the idle time of processors to achieve a substantial speedup. We then experimentally study this algorithm in the context of training a restricted Boltzmann machine on a large collaborative filtering task.

Mini-batch gradient descent

Minimize $\mathbb{E}_z[C(\theta, z)]$ where $C$ is some (typically convex) cost function and the expectation is computed over training points $z$. In mini-batch gradient descent, this is achieved using the update rule

$$\theta_{t+1} \leftarrow \theta_t - \alpha \sum (C(\theta, z_i))_{\theta}$$

where $\alpha$ is some learning rate and $b$ is the number of training points in a mini-batch.

Zealous parallel gradient descent algorithm

Procedures followed by each individual thread

- $\Delta \theta$: pending updates of $\theta$;
- $b$: current mini-batch;
- $\text{pid}$: unique identifier.

Global state

- $\theta$: vector of parameters of the model;
- next: pid of the next thread allowed to update $\theta$;
- counter: array of integers, such that $\text{counter}[i]$ corresponds to the number of pending updates of thread $i$.

Policy functions

- function trylock(pid)
  - counter[pid]++;
  - return next = pid;
- function next(pid)
  - counter[pid] ← 0;
  - next ← arg max(counter)

Critical section

- $\theta \leftarrow \theta - \alpha \Delta \theta$;
- $\Delta \theta \leftarrow 0$;

next(pid);

Conclusions and future work

- Significant speedup over the asynchronous parallel gradient descent algorithm.
- Future work: corroborate the results obtained in this work with more thorough experiments.
- Updates of $\theta$ may become too much delayed if the number of cores becomes too large, which can impair convergence.
- Future work: Explore strategies to counter the effects of delay. Derive theoretical guarantees on the convergence of the algorithm.

References and acknowledgements


Gilles Louppe and Pierre Geurts are respectively research fellow and research associate of the FNRS Belgium. This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.