FACEBOOK AI

Paper review

2018

Linna HU, Hongyu Huang, Jiashuo Sun

Ams560

2018/3/11

**Review of Facebook AI**

**Summary**

At Facebook, machine learning provides key capabilities in driving many essential products and services. Because of the diversity of models in practice, Facebook needs diverse efforts that span machine learning algorithms, software, and hardware design.

Aspects of datacenter infrastructure that support machine learning at Facebook:

Software: ① open-source machine learning frameworks. ② internal “ML-as-a-Service” flows. ③ distributed training algorithms.

Hardware: leveraging a large fleet of CPU and GPU platfroms for training models.

**Contributions**

1. Machine learning is applied across most services, and computer vision represents only a small fraction of the resource requirements.
2. Facebook relies upon an incredibly diverse set of machine learning approaches.
3. Tremendous amount of data are funneled through machine learning pipelines, and this creates engineering and efficiency challenges far beyond the computer nodes.
4. Facebook relies on CPUs for inference, and both CPUs and GPUs for training.
5. The worldwide scale of people on Facebook and corresponding diurnal activity patterns result in a huge number of machines that can be harnessed for machine learning tasks such as distributed training at scale.

**Comments**

When designing the hardware and software infrastructure, Facebook considers these factors: the importance of co-location data with compute, the importance of handling a variety of Machine Learning workloads, and the opportunities that arise from spare capacity from diurnal compute cycles. It’s hardware and software design support machine learning at global scale.

Facebook builds high-performance, energy-efficient systems for machine learning that meet the demands of the abundant ML-based applications. It evaluates and prototypes novel hardware solutions, while simultaneously keeping an eye on the upcoming, near and longer-term algorithm changes, and the potential impact on system-level design.