Large-Scale Distributed Sentiment Analysis with RNN

Jianzhun Du, Rong liu, Matteo Zhang, Yan Zhao
Harvard University SEAS, Advisors: Ignacio M.Llorente, Zudi Lin

Abstract

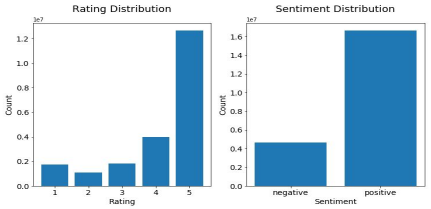
In this projects, **we use Recurrent Neural Networks (RNN) to uncover whether a piece of Amazon review has positive sentiment.**

We employ MapReduce on AWS cluster to clean more than 140 million reviews and transform texts of varying length to fixed-length number format. HDF5 file format is used to load the data without blowing up the memory. After processing, we distribute the workloads across multiple GPUs on an AWS cluster. Large minibatch technique and dynamic load balancing are utilized to speed up the application, written with Pytorch, whose MPI interface with NCCL backend enables communication between nodes.

In addition, we compare various real-life scenario tradeoffs, such as bottlenecks introduced when using a mix of different GPUs and money-speed tradeoff when selecting GPU instances.

Data

We combined score 1, 2,3 into negative class, 4 and 5 into positive class. Below is a picture of class break down before and after we merge classes. (Our dataset are granted by Dr. Julian McAuley).

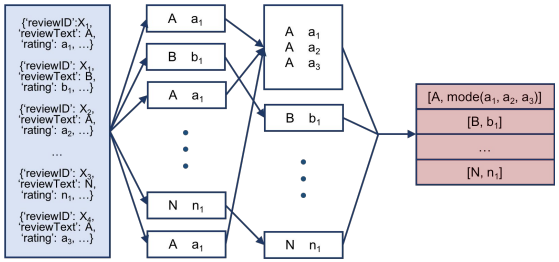


Preprocessing Steps:

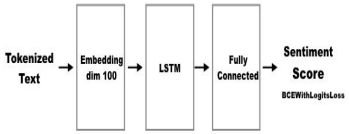
- Remove duplicates of text and keep the mode of the ratings
- Map ratings to binary sentiment indicators
- Remove Stopwords
- Map meaningful and frequent words to numbers according to a dictionary that we generated based on our own dataset
- Truncate or pad text sequences to achieve fixed length

Methodology

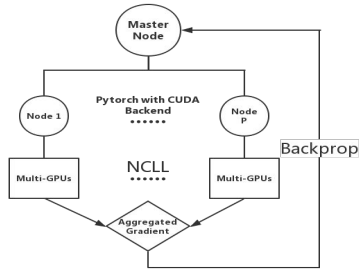
MapReduce Workflow



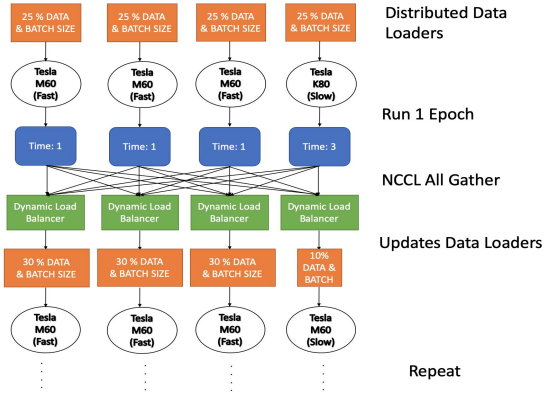
RNN Structure



Pytorch Parallelization



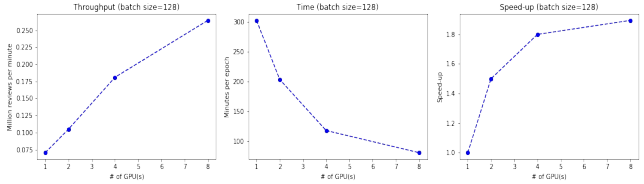
Dynamic Load Balancer



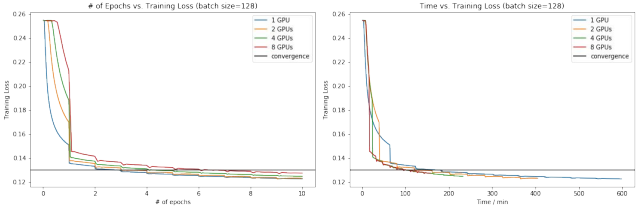
Results

We demonstrate our results from three aspects:

- **Strong Scaling**



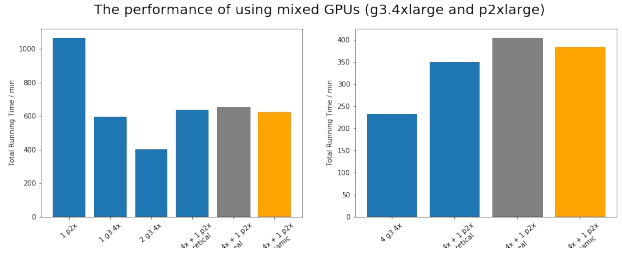
- **The convergence with different number of GPUs**



- **The performance with different distributions of GPUs per node for total 4 GPUs**

# of Node	# of GPU	Time /min	Speed-up
1	4	21.9	2.73
2	2	26.4	2.27
4	1	23.3	2.57

- **The performance with mixed GPUs and our dynamic load balancer**



- **Money-Speed Tradeoff**

Experiment	# p2.xlarge	# g3.4xlarge	# g3.16xlarge	Seconds	Hours	Total Price
Single g3.4xlarge	0	1	0	35920	9.98	11.38
Single p2.xlarge	1	0	0	63923	17.76	15.98
Single g3.16xlarge	0	0	1	13156	3.65	16.64
Two g3.4xlarge	0	2	0	24316	6.75	15.39
Two g3.16xlarge with 2 GPU only	0	0	1	15820	4.39	20.02
Four g3.4xlarge	0	4	0	13973	3.88	17.69
Two g3.16xlarge	0	0	2	9531	2.65	24.17