MLDB Presentation
SLIDE: Sub-LInear Deep Learning Engine

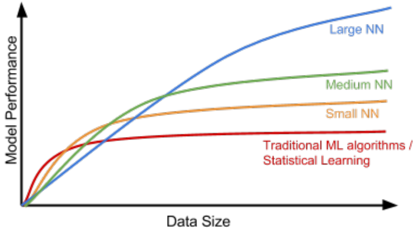
Beidi Chen Tharun Medini James Farwell Sameh Gobriel
Charlie Tai Anshumali Shrivastava

June 16, 2020

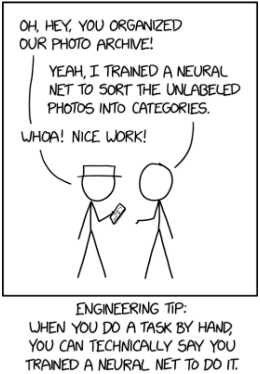
Overview

1. Motivation
   - Existing Approaches
   - Problem Setting
2. Contributions
3. Locality Sensitive Hashing
   - Sampling Approach to LSH
   - Additional Optimizations
4. Implementation
5. Results

Motivation



(a) Model Performance wrt Dataset size



(b) Fun xkcd comic

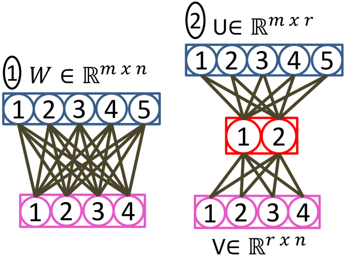
Trends

- Large datasets $\rightarrow$ More Data
- Big models (Eg, 17B parameter NLP models)
- Improvements in optimizations and gradient descent
- Matrix multiplication is a computational bottleneck
- Many approaches exists such as GPUs

Existing Approaches

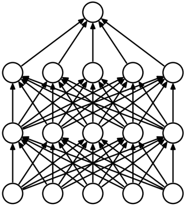
Low Rank structure

- $W \in \mathbb{R}^{m \times n}$ is weight matrix
- W has a low-rank structure $W = UV$
- $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$, where $r \ll \min(m, n)$
- Equivalent representation with I activation function is better
- $\mathcal{O}(mn)$ becomes $\mathcal{O}(mr + rn)$
- Better storage of parameters as well
- But still needs dense gradient update, cannot parallelise asynchronously

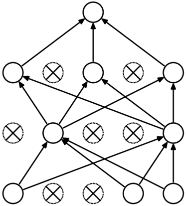


Dropout and Sparsity

- Well known regularization method for Neural Networks
- With probability p neurons in each layer is turned off
- Used during training to ensure model generalizes
- Sparsity above 50% tends to begin hurting performance



(a) Standard Neural Net



(b) After applying dropout.

Adaptive Dropout

- [Ba and Frey, 2013]

Problem Setting

Contributions

Main Contributions

- C++ OpenMP implementation for "standard" CPU
- Sparsity inspired, LSH based backpropagation algorithm
- Rigorous evaluation with TF-GPU and CPU
- Further optimizations using Hugepages and SIMD instructions

Locality Sensitive Hashing

Sampling Approach to LSH

Additional Optimizations

Implementation

Results

Questions or Comments

Ba, L. J. and Frey, B. (2013).
Adaptive dropout for training deep neural networks.
In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3084–3092, Red Hook, NY, USA. Curran Associates Inc.