Parallel Backpropagation for Multilayer Neural Networks

Palash Goyal[1] Nitin Kamra[1] Sungyong Seo[1] Vasileios Zois[1]

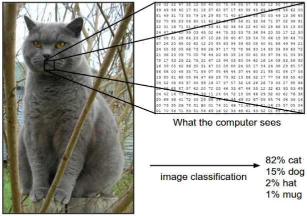
[1]Department of Computer Science
University of Southern California

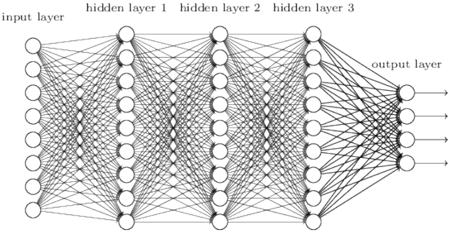
May 5, 2016

Outline

Problem statement

- Prevailing neural network architectures are implemented using several hidden layers, with each one consisting of thousands to millions of neurons in order to generalize well on diverse inputs.
- Can you imagine how many parameters need to be trained for every iterations?





Problem statement(contd.)

- Gradient Descent is the most commonly used optimization algorithm used to train neural networks in supervised settings.
- Having the gradient as a sum of partial gradients with respect to individual training examples opens up the possibility of parallelizing the gradient computation efficiently.

Gradient Descent

- Suppose we have some training set (x_i, y_i) for $i = 1, \cdots, N$:

$$\mathcal{L}_{MSE}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i;\theta))^2$$

- Then, the parameters are updated as follow:

$$\theta_j = \theta_j - \alpha\frac{\partial\mathcal{L}_{MSE}(\theta)}{\partial\theta_j}$$

Batch vs. Stochastic

- If the gradient is computed by using whole dataset (N), it is Batch Gradient Descent.
  - This is great for convex, or relatively smooth error manifolds.
  - Gradient is less noisy (averaged over a large number of samples).
  - For large or infinite datasets, batch gradient is impractical.
- Stochastic Gradient Descent(SGD) computes the gradient using a single example.
  - SGD works well for error manifolds that have lots of local maxima/minima because the somewhat noisy gradient helps to escape local minima into a region that hopefully is more optimal.
  - SGD may go "zig-zag" to a local minimum because of highly noisy gradient.

Alternative: Mini Batch

- Mini Batch Gradient Descent (MGD) is to compute the gradient against more than one training example at each step.
- M is the mini batch size.

$$\mathcal{L}_{MSE}^{k}(\theta) = \frac{1}{M} \sum_{i=s_k}^{s_k+M} (y_i - f(x_i;\theta))^2$$

$$\theta_k = \theta_k - \alpha \frac{\partial \mathcal{L}_{MSE}^{k}(\theta)}{\partial \theta_k}$$

- Parallelization : Distributing training examples across various processors and letting them compute a partial gradient over their own training examples.

Outline

Pthreads - POSIX threads

- **Low level multithreading API** - Threads are created and each task is assigned to each thread in parallel.
- **TODO** - ABOUT Basic functions to show how Pthread is used in our project

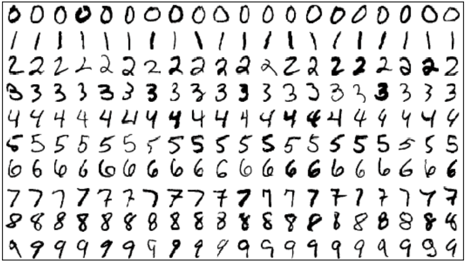
Outline

CUDA

- **TODO** - General concepts and how CUDA is used in our project

MNIST dataset

- MNIST, handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples.
- The digits have been size-normalized and centered in a fixed-size image to 28x28 image.
- 10,000 examples in 60,000 training set are used as validation set and left 50,000 images are used for training.
- The original labels values are 0 to 9 but it is vectorized by one-hot encoding.



Experiment parameters

- **Network structures**
  - # Layers :
  - # Nodes :
  - # Bias :
  - What else?
- **Activation function**
  - Sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$
$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$

- **Hyperparameters**
  - Learning rate α :
  - Regularization : None
  - # Epochs : 50?
  - Size of Mini Batch : 1000?
  - What else?

Parallel experiments

- **PThreads**
  - About PThreads setting
- **CUDA**
  - About CUDA setting
- **theano**
  - About theano

Results

- **Running time vs. Accuracy FIGURE**

Analysis

- Discussion