Project: A Further Investigation on Running Time

Last updated: December 30, 2019
Goal

- Investigating why for the previous project some MATLAB operations are in inefficient
Consider the following MATLAB code to run the same operation on CPU and GPU

```matlab
function test

m = 10000;

for gpu_use = 0:1

    A = gpu(randn(m,m), gpu_use);
    B = gpu(randn(m,m), gpu_use);
    a = gpu(randn(3*m,1), gpu_use);

end
```

b = gpu(randn(1,3*m), gpu_use);

f1 = @() A*B;
f2 = @() a+b;

if gpu_use == 1,
gputimeit(f1)
gputimeit(f2)
else
    timeit(f1)
    timeit(f2)
function M = gpu(M, gpu_use)

if gpu_use == 1
    M = gpuArray(M);
end

Results:
>> test
ans =
   4.8204
ans =
   0.5093
ans =
   3.6401
ans =
   0.0212
We conduct this experiment because both operations are used in our stochastic gradient implementation.

For example, in `padding_and_phiZ.m` we have:

```matlab
idx = reshape(net.idx_pad{m} +
    [0:num_data-1]*a*b, [], 1);
```

In it the sum of a column and a row vector is conducted.

This code can be run on MATLAB only. Neither `timeit` nor `gputimeit` is supported on Octave.
Complexity of the two operations

$10^{12}$ and $9 \times 10^8$

We do not expect a 1000-fold time difference because we already know that matrix products by optimized BLAS get better data locality.

But the difference between CPU and GPU is surprising.

From CPU to GPU, the matrix product is shortened by less than half.
But for the sum of a column and a row vectors, GPU is 25 times faster

Let’s try to investigate why this happened

To begin let’s write a C code to do the vector sum

Use multi-core and do all possible optimization

See how much you can do better (or worse) than MATLAB

We know that apparently CPU is not fully utilized

Can you try to measure the cache misses?
Can you do some investigation to explain why on GPU the vector sum can be done so efficiently?

FYI, for matrix products, we have checked non-squared matrices. The speedup from CPU to GPE may be slightly better (but only slightly better)
Students with the following IDs (last three digits): ??

please do a ??-minute presentation (??-minute the contents and ??-minute Q&A)