Week 4 Report

Weitian Li weitian.li@rutgers.edu

1. Reproduce HWGQ with TensorFlow (If you want to join in one paper now, you should do this.) Here is my implement.

完成情况：caffe 的代码不是很理解，然后对于 tf 版的代码进行了回传的改写。

```
import tensorflow as tf

def get_hwgq(bitA):

    def quantize(x, k):
        # in order of
        assert k in [2,3,4,5], 'Does not support %d bits' % k
        code_book={
        '2':[0.5380, 0., 0.5380*(2**2-1)],
        '3':[0.3218, 0., 0.3218*(2**3-1)],
        '4':[0.1813, 0., 0.1813*(2**4-1)],
        '5':[0.1029, 0., 0.1029*(2**5-1)]
        }
        delta, minv, maxv = code_book[str(k)]
        #print(delta,minv,maxv)
        @tf.custom_gradient
        def _quantize(x):
            return tf.to_float(x>0.)*(tf.clip_by_value((tf.floor(x/delta +
0.5)+tf.to_float(x<0.5*delta))*delta, minv, maxv)),\
                    lambda dy:
dy*(tf.to_float(x>minv)*tf.to_float(x<maxv)+tf.to_float(x>maxv)*tf.log(x-maxv))
        return _quantize(x)

    def fa(x):
        if bitA == 32:
            return x
        return quantize(x, bitA)

    return fa
```

增加了一个 long-tail 的 ReLU 回传，还没有进行实验不知道能提高多少，原文说差别不是很大。

2. Write HWGQ layer in HWGQ with Cuda.

完成情况：还在阅读相关论文的资料。