Project Status Report

Tao JIANG:

1. Implemented a convolutional network based on Pytorch, and trained it on the 30-million Chinese Trademark Database.
2. Deployed the model as an API based on Flask and Gevent on AWS.
3. Extracted about 65 thousand image features, and saved the numpy array as .h5 file.

Zichao DONG:

1. Read papers about active learning. I found that the most effective way to do active learning in image retrieval is finding hard and special examples. Unlike traditional active learning missions, which would focus on finding representative samples, our mission is to find trademarks that are different from that in our current database in order to make our system better.
2. In order to get hard examples, like traditional deep learning based active learning tasks. I will directly choose low score images from the CNN.
3. In order to get special examples, I will use clustering algorithms to get outliers. In other words, people are always try to make fancy trademarks which are not similar to any exist ones, so them would attach more attention to special examples in database.

Lifang LIU:

Set up a multi-node hadoop cluster in AWS, involving a single NameNode and three DataNodes which serve as processing slaves. And then continue to configure the environment on the servers where the project will run.

Zhende ZHUANG:

Problems:

The origin feature file whose format is .h5 is hardly to be input to the Hadoop streaming program because the default setting of Hadoop streaming requires an stdin and the official optional formats do not support .h5 file.

```
hadoop jar hadoop-streaming-2.4.1.jar
    -input input     # means here cannot be an .h5 file
    -output output
    -mapper mapper.py
    -reducer reducer.py
```

Solution:

1. Turn the original .h5 file into .csv file which can be the input of the Hadoop streaming program, and meanwhile generate a subset to act as the test set.
2. Finish some function to preprocess the data from the .csv file
3. Finish the first version of the mapper.py and reducer.py (will be test in a pseudo node Hadoop firstly and later test on the Hadoop cluster)

4. Finish the evaluation method “cos similarity”

The general structure of the mapper and reducer