Name: Xiyuan Liu
Class: MSML 603, section 2
Professor: Kemal Davaslioglu
Submission time: Oct 14th 2022

Project proposal

1. Research questions:

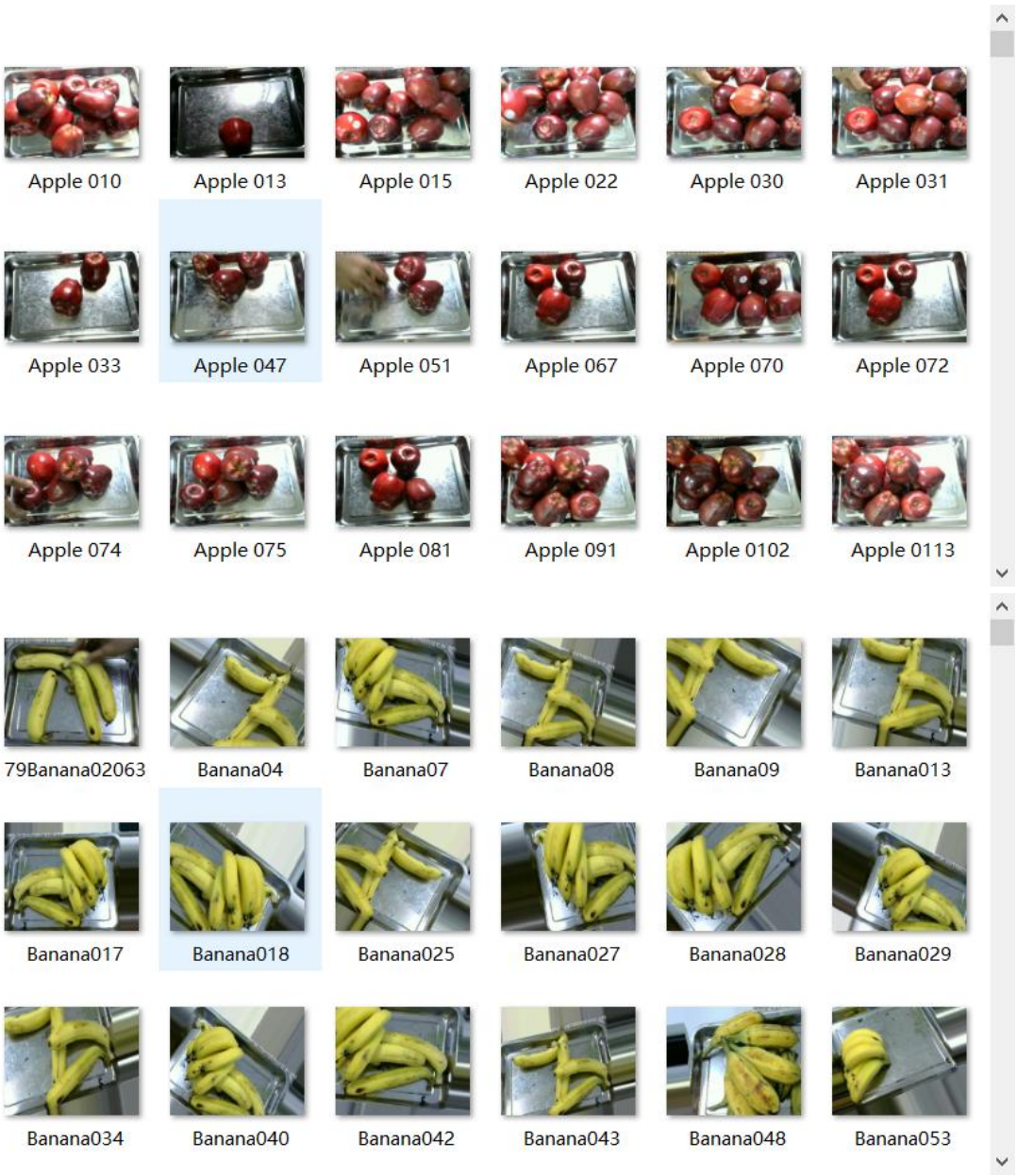
   Image classification problem has been one of the hottest topic in the Artificial Intelligent. Why we need to use AI to deal with image processing? Well, after we enter 21st centuries, the development of technology including internet, improvement of physical camera and data storage technology, the size of data of images growth rapidly. For example, if we are trying to find a particular person on Instagram and the only information we have is a set of images, it is impossible do it by human hand. Also the problems are not only the increasing in the quality of picture, but also the increasing in quantity of the pictures which means a picture may contains more information than before and a person cannot easily recognize it. With algorithms of image classification, people can work more capable and efficient than before.

   For this project, I am trying to do a photo classification problem with application of convolution neural network (CNN). First of all, I will build a CNN model, feed the model with fruits images and see how it works on classification missions. If I can get good score with this model, I will also use other data set to train my model and see how it works with different data set. For example, if I can get a data set of Pokemon, I may use the model to identify them in the episodes. In this step, I do not specific plan to use any specific data yet.

2. Data set description

   My data set is a folder with some fruits photos. Data set of fruit I am using contains 15 classes, including Apple, Banana, Carambola, Kiwi, etc... The size of the data will contain total 12000 samples, 800 samples from each classes. The folder name shows the fruit name of the images in the folder which I will turn it into my labels. I download it from the website and I will upload it to a cloud. The total size of the data set is 1.48 GB. I am planning to use validation to randomly split my data set into training and testing set. 80% of the data in each classes will be used as training data and the remaining will be used as testing data.

   Here are some of the screen shot of my images:



You may notice that there is pattern of this images that the fruit is placing on a metal plate with different posture. I am intentional to pick fruit data set with some pattern. On one hand, I want to control the noise from the environment. On the other hand, I want to see how the model deal with the images that have similarities with different labels. If I can get good score on fruit classification, I will feed the model with another data set which contains more noise and see how the model works in that scenarios.

3. Methodology

As I mentions in research question, I am planning to use convolution neural network (CNN) to do the classification. As one of the most popular neural network model, I believe it is qualified to do the classification mission since CNN will process data with step like kernel, padding, and stride, it works perfectly with images which has so many pixels and each pixel has 3 parameters. By the way, most of AI image identifier is using CNN. I may also test some other models such as SVM and decision tree and see how they perform. Since I am using images as data set, there will not

need to clean the data and erase the missing data specially after padding step in CNN.

4. Performance Metrics

   The first thing I will do is to check the training accuracy and validation accuracy. Then I will also check the training loss and validation loss. I may also check the performance of prediction in each classes and try to find which fruit make the most accuracy damage in the prediction. I not decided yet, but I may build a confusion matrix to see how the model performance.