Project Proposal: Image Classification - Corn Kernels

Jon Tanner Nelson

The Project Generally:

I am aiming to gain some early experience in image classification, as computer vision is of interest to me personally. Towards this end I have found a community driven Kaggle competition which has already run its course where classification of corn kernel images is the objective. Given the nature of a completed Kaggle competition including community discussion, and posted solutions I believe this project is an ideal entry point into image classification for me, as I will have a rich source of feedback to consult which can aid in overcoming novel obstacles. Obviously I would refrain from viewing and/or using the code of others until I have an at least semi-functional solution, as developing my own workflow and knowledgebase is a key learning outcome I wish to strive towards.

Further, there is an associated tutorial form the competition organizer Rob Mulla, for an image processing library OpenCV,which can stand to aid in building up the basic knowledge necessary to begin this task. Seeing as there are several organized resources to support my work on this problem I think my time spent will be highly productive and informative, as I will not be entirely unguided in my approach to a problem type which is novel to me. Additionally I would look to this classic computer vision task linked within the final project power point as a separate source of guidance within this project when needed.

The competition hosts a relatively large data set of labeled images of corn kernels which is freely available on Kaggle. The data set has 14,323 images within the training set, all of which are labeled as one of four categories: pure, broken, discolored, and silkcut. Thus even without access to the 3,480 images within the testing set available I would still be able to self test the accuracy of my model using the provided training set as my total data set.

Proposed work:

1. Downloading the data from the linked Kaggle page,
2. Installing the OpenCV (cv2) and glob library on my local machine,
3. Inspecting the images manually to build some intuition about what features may be helpful to focus on in my ML architecture; further this may help with determining if any preprocessing could be helpful,

4. Preprocessing images, splitting into testing and training set, possibly reducing the size of images to enable processing of the data within reasonable time frames, or converting to grayscale or reducing dimensionality in terms of the color channels if one or two color channels obviously capture the most critical variance within the data, further if needed other forms of dimensionality reduction or data compression could be considered if compute time becomes an issue,
5. Feature extraction could be a potentially valuable step to take, where raw data can be processed to return further data which differentiates images and thus improves classification accuracy. I would need to research the methods but I have learned that options such as histogram of oriented gradients (HOG), or local binary patterns (LBP) may be useful feature extraction techniques,
6. Selecting and training classifiers, depending on further research I would narrow into which classifier suites the project best, however I would like to try using multiple classifiers in order to really explore the options that exist. As of now the two classifiers which I would think to use first are the KNN or Support Vector Machine (SVM) classifiers, however I would engage in further research before settling on my first choice once model development has begun,
7. Finally evaluation of various classifiers, hyperparameters within those classifiers, and preprocessing techniques would be performed in order to identify a best initial classifier to fine tune,
    a. (calculate the accuracy, precision, recall, and F1-score to evaluate the performance of the classifier)
8. Fine tuning the top classifier would be the objective until project completion, aiming to achieve the best results manageable within the time frame.

Timeline:

All proposed activities are to be performed leading up to the date and deadline they are linked to.

Report 1 – 4-9-23:
Data download, library instillation, image-set inspection, data splitting, testing of 2-3 classifiers early on to determine what further preprocessing and feature extraction may be necessary for computational feasibility as well as improving accuracy. This would involved some research into classifier strengths and weaknesses as well as best practices with them with respect to image classification tasks.

Report 2 – 4-16-23:
Perform novel preprocessing and feature extraction techniques, research feature extraction techniques ad hoc, rerun early attempts at classification with as large portions of the total data as is possible in order to evaluate efficacy of novel preprocessing and feature extraction techniques.

Lightning Talk – 4-25-23:
Iterate on most performant classifier via hyperparameter tuning and possibly further feature extraction and preprocessing tuning. Prepare lightning talk presentation explaining the overall process of research, implementation and iteration.

Final Report – 4-30-23:
Perform any final iterations which may improve performance, check my solution against that of top performers on Kaggle to learn where I may improve further in the future with similar tasks.

References:

1. https://www.kaggle.com/competitions/kaggle-pog-series-s01e03/overview
2. https://www.youtube.com/watch?v=kSqxn6zGE0c
3. https://www.kaggle.com/competitions/digit-recognizer/overview