Proposal

Team members: Junchi Tian, Yiming Dong, Jingshu Song

In the final project, I decide to use Kannada handwritten digits dataset from Kaggle (https://www.kaggle.com/c/Kannada-MNIST/overview). This dataset is similar to the most famous computer vision dataset: handwritten digits recognition provided by Yann LeCun. Instead of using Arabic digits, this dataset uses Kannada digits. So, this is multi-classification model and will use to recognize Kannada numeric from 0 to 9.

Kannada Digits from 1 to 10

Kannada is a language spoken predominantly by people of Karnataka in southwestern India. The language has roughly 45 million native speakers and is written using the Kannada script. One of reasons we choose this dataset because it is not as simple as Arabic dataset, also not hard to implement deep learning algorithm by us. Another important reason is that this model will be helpful to Karnataka people and scholars who interest in Kannada culture.

In this dataset, there are 785 images in train and 785 images in test, with 28 pixels in height and 28 pixels in width, for a total of 784 pixels. We consider use ImageDataGenerator to generate more images to increase the size of train set to get a better performance. So deep learning is needed.

We will use Keras as framework and CNN as deep network. We choose Keras because of its simplicity and readability. Also, this dataset is not very large so that we don't need to worry about the speed. We choose CNN because it will show high performance than simple MLP with two dimensional images.

The necessary reference material and detailed technical information can be found here: https://arxiv.org/abs/1908.01242 and https://github.com/vinayprabhu/Kannada_MNIST. The author of paper provides this dataset to Kaggle.

The evaluation metric for this contest is the categorization accuracy, or the proportion of test images that are correctly classified. For example, a categorization accuracy of 0.97 indicates that people have correctly classified all but 3% of the images. We will try to achieve accuracy over 0.98.

We truly apologize for submitting this proposal late. We chose this dataset two weeks ago and we've already built our first model with some errors. But we spent most of time on Exam 2 and final project in NLP class in the past 2 weeks. In this weekend, we will correct errors and try to improve performance and completed our report.