6.867 Project Proposal

Landon Carter, Courtney Guo, Jingyu Li

November 2015

1 Description

In this work, we wish to implement semi-supervised image clustering. For this, we will train an autoencoder neural net with a relatively small set of labelled examples, representing only a subset of the clusters. From the hidden layer of this neural net, we will extract a compressed feature representation to use in EM clustering. The set of features to feed into the autoencoder will be a pre-processed set of features taken from OpenCV. This is particularly interesting, because we wish to separately cluster images that we do not already have examples of. An example would be training on labelled images of sunsets and people, and asking that the algorithm distinguish images of dogs and computers.

As an auxillary, from each cluster we will select the example which is closest to the mean of the cluster, so that this example can be shown to a human for labelling. This example should be the best representation of each cluster, and so would be perfect for "tagging" purposes.

2 Challenges

- It may be difficult to obtain representative features.
- We need to guarantee we select a diverse and representative set of images for each category. Otherwise, we will not be able to extrapolate well to unseen data.
- Runtime may be a concern. Training a neural network takes significant time, so it is important to reduce images to a reasonable number of features (as inputs to the neural network). This is important because the search space will simply be too large if we have too many neurons.

3 Distribution of Work

Though we will each contribute to all aspects of the work and seek to fully understand all aspects, the features each of us will focus on are listed below:

- Landon Carter – Implement EM clustering algorithm, construct datasets.
- Courtney Guo – Design and implement neural network.
- Jingyu Li – Reduce images to a large set of features (which will be inputted into the neural network and used to compose final features) using OpenCV.

4 Timeline

- November 21 – Background research, implement clustering algorithm
- November 28 – Neural network that gives features of images
- December 1 – Combine neural network with clustering and train for good representative image features

- December 5 – Be able to give representative images of specific categories
- December 8 – Finish write-up