Learning Representations for Computer Vision Tasks

Thesis Abstract

Siddhartha Chandra

CVIT, IIIT Hyderabad

Learning representations for computer vision tasks has been the holy grail for the vision community for long. Research in computer vision is dedicated towards developing machines that understand image data, which takes a variety of forms such as images, video sequences, views from multiple cameras, high dimensional data from medical scanners and so on. Good representations of the data intend to discover the hidden structure in it; better insights into the nature of the data can help choose or create better features, learn better similarity measures between data points, build better predictive and descriptive models, and ultimately drive better decisions from data. Research into this field has shown that good representations are more often than not task specific: there is no single universal set of features that solves all the problems in computer vision. Consequently, feature learning for computer vision tasks is not a problem, rather a set of problems, a full fledged field of research per se.

In this thesis, we seek to learn good, semantically meaningful representations for some of the popular computer vision tasks, such as visual classification, action recognition and so on. We study and employ a variety of existing feature learning approaches, and devise novel strategies for learning representations on these tasks. Additionally we compare our methods with the traditional approaches. We discuss the design choices we make, and the effects of varying the parametric variables in our approaches. We provide empirical evidence to show our representations are better at solving the tasks at hand than the traditional ones.

To solve the task of action recognition, we devise a novel PLS kernel that employs Partial Least Squares (PLS) regression to derive a scalar measure of similarity between two video sequences. We use this similarity kernel to solve the tasks of hand gesture recognition and action classification. We demonstrate that our approach significantly outperforms the state of the art approaches on two popular datasets: Cambridge hand gesture dataset and UCF sports action dataset.

We use a variety of approaches to tackle the popular task of visual classification. We describe a novel hierarchical feature learning strategy that uses low level Bag of Words visual words to create “higher level” features by making use of the spatial context in images. Our model uses a novel *Naive Bayes Clustering*

algorithm to convert a 2-D symbolic image at one level to a 2-D symbolic image at the next level with richer features. On two popular datasets, Pascal VOC 2007 and Caltech 101, we demonstrate the superiority of our representations to the traditional BoW and deep learning representations.

Driven by the hypothesis that most data, such as images, lies in multiple non-linear manifolds, we propose a novel non-linear subspace clustering framework that uses K Restricted Boltzmann Machines (K-RBMs) to learn non-linear manifolds in the raw image space. We solve the coupled problem of finding the right non-linear manifolds in the input space and associating image patches with those manifolds in an iterative Expection Maximization (EM) like algorithm to minimize the overall reconstruction error. Our clustering framework is comparable to the state of the art clustering approaches on a variety of synthetic and real datasets. We further employ K-RBMs for feature learning from raw images. Extensive empirical results over several popular image classification datasets show that such a framework outperforms the traditional feature representations such as the SIFT based Bag-of-Words (BoW) and convolutional deep belief networks.

This thesis is an account of our efforts to do our bit to contribute to this fascinating field. We admit that research in this field will continue for a long time, for solving computer vision is still a distant dream. We hope that we have earned the right to say one day, in retrospect, that we were on the right track.