11785 Project Proposal

Yuxuan Sun (yuxuans)
Yichen Pan (yichenp)

Introduction

In large e-commerce companies, the well categorization of products is becoming a more and more challenging task since the list of products is rapidly growing, and the number of categories is extremely large. Apart from inefficient manually labelling, traditional machine learning methods can be applied to automatically label the products and predict their categories. Cdiscount.com, the France's largest non-food e-commerce company, is also facing the categorization problem due to its rapid growing lists of products. Here we decide to build a deep neural network classification framework to automatically categorize the images.

Scope and Background Research

The main task in this project is about the so called Extreme Classification, where we need to deal with multiclass involving an extremely large number of labels. Different from normal classification tasks, extreme classification task may involve a number of difficulties. First, the computational complexity has been an issue for a long time. Second, data scarcity remains a question despite the vastness of available data. Additionally, the statistical dependence/correlation of the labels poses challenges and opportunities for learning approaches. All the properties of extreme classification problems (e.g. data and feature distribution) and their specificities across different fields are still not well understood. Moreover, valid evaluation measures in this setting have not been fully and well settled.

Due to the fact that datasets in extreme classification exhibit fit to power-law distribution,which indicates that a large fraction of labels have very few positive instances in the data distribution, most state-of-the-art approaches for extreme multi-label classification are mainly focused on capturing correlation among labels by embedding the label matrix to a low-dimensional linear sub-space. Specifically, by assuming that the training label matrix is low-rank, the effective number of labels can be reduced by projecting the high dimensional label vectors onto a low dimensional linear subspace. Still, leading embedding approaches have been unable to deliver high prediction accuracies or scale to large problems as the low rank assumption is violated in most real world applications.

Dataset

The dataset contains a list of 7,069,896 dictionaries, one per product. Each dictionary contains a product id, the category id of the product, and between 1-4 images:

- Almost 9 million products: half of the current catalogue
- More than 15 million images at 180x180 resolution
- More than 5000 categories