MATERIALL DATA APPLICATION PROJECT

Abstract. In this document we outline the Materiall Data Application Project for the Data-X at Berkeley Applied Data Science course.

1. Introduction

When browsing housing data, two basic data types to understand for further processing in the data pipeline are text and images. A key question we ask prior to building a personalized recommendation model is what is the universal space of attributes? Identifying a basic set of attribute values is a precursor to applying any recommendation engine. Textual information can provide a minimal subset of attributes that can be used, but often we find that data to be incomplete or in need of augmentation. Images provide another source of data from which known attributes can be extracted as well as the identification of additional features. The task for the students of the Data-X program is to build image classification models which will help extract attributes important to the home buying process.

2. ROI on Materiall Data Application Project

Image Classification continues to be one of the central problems in Computer Vision and Machine Learning. The task is to predict a single label or a distribution over labels for a given image. The task can further be broken down into either a supervised learning problem, where our models are trained using labeled data, or an unsupervised learning problem, where we do not use labeled training data. The current project focuses on building image classification models to aid in the home buying process via the Materiall Personalized Recommendation Engine.

3. Problem Statement

The specific image classification models to be implemented can be understood in steps. As a first pass, students will implement an image classification model that can predict the room in which the image was taken. Students will have the ability to train their models first on a set of labeled images and then deployed and tested. An important component of the first step is the identification of a suitable training set within the catalogue of bay area homes. The next problem will focus on object detection per room type. The idea here is to detect important attributes of a home which may already be catalogued (such as number of bedrooms, number of bathrooms, etc.,) as well as other attributes which may not be listed but inferred from the objects detected (for example lighting in a room may be a function of the number of windows detected or kitchen space could be a function of the number of distinct objects present in the image). Finally as a last step we will aim to build an image classifier for the extracted attributes.

4. Methodology

In this section please outline in detail the experiments you are running and provide code snippets to understand implementation. Some background information about the methodology used would also be helpful. Here the entire workflow should be documented. For example: Image Preprocessing, Detection of an Object, Feature Extraction and training, Classification of the Object.

5. Findings

Please report your experimental findings in this section. Provide data outputs and visuals.

6. Discussion

Feel free to engage in a discussion of your results and any new learnings that emerged from the experiment.

7. References

Please List all references used for both analysis of experimental results as well as methodology used. In particular any code snippets used from external sources should also be documented.