CSC 215-01 Artificial Intelligence (Spring 2020)

Mini-Project 3: House Price Estimation using Both Visual and Textual Data

Due at 4 pm, Wednesday, March 18, 2020

Demo Session: class time, Wednesday, March 18, 2020

1. Problem Formulation

Most existing automatic house price estimation systems rely only on some textual data like its neighborhood area and the number of rooms. In this project, we practice with algorithms to extract visual features from house photos and combining them with the house's textual information.

The goal of this project is to build **a novel house price estimation system by using both textual and visual inputs**, other than only using textual information such as area, neighborhood, and number of bedrooms.

Goal: Build your best deep learning model and show its RMSE and lift chart on test data.

2. Dataset

https://github.com/emanhamed/Houses-dataset

This dataset contains both visual and textual information.

- Each house is represented by four images for bedroom, bathroom, kitchen and a frontal image of the house.
- The dataset folder contains 2140 images, 4 images for each house.

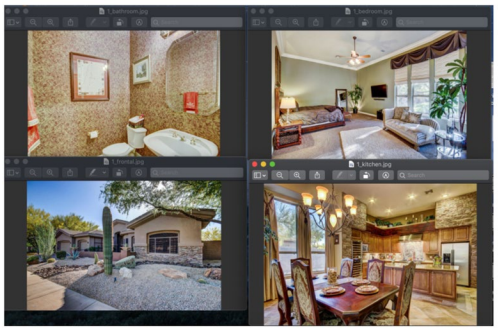
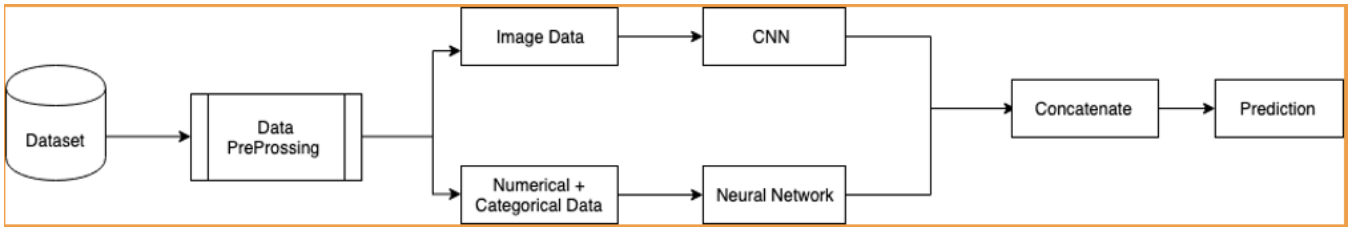


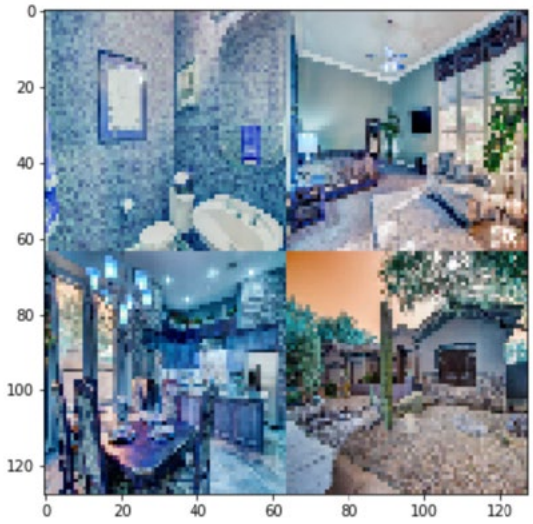
Fig. 1 Example of Images of house

- The dataset has a text file that contains the textual metadata of the dataset. Each row in the file represents the number of house in order. The numbers represent number of bedrooms, number of bathrooms, area of the house, zipcode and the price.

3. Model Design



We will use Functional Keras API. For each house, for simplicity, let's create a combined image using the 4 house images (bedroom, bathroom, frontal and kitchen, as shown below. You can find the code to create such images in the cheat sheet. To speed up training, you may use Google Colab GPU.



4. Requirements

- To remove the outliers first. For example, you may only keep the houses with a price between 100K and 900K
- Split data for training and testing. Use training data to train your models and evaluate the model quality using test data
- Encode categorical features and normalize numeric features.
- You must use EarlyStopping when training neural networks using Tensorflow.

5. Grading breakdown

You may feel this project is described with some certain degree of vagueness, which is left on purpose. In other words, **creativity is strongly encouraged**. Your grade for this project will be based on the soundness of your design, the novelty of your work, and the effort you put into the project.

Use the evaluation form on Canvas as a checklist to make sure your work meet all the requirements.

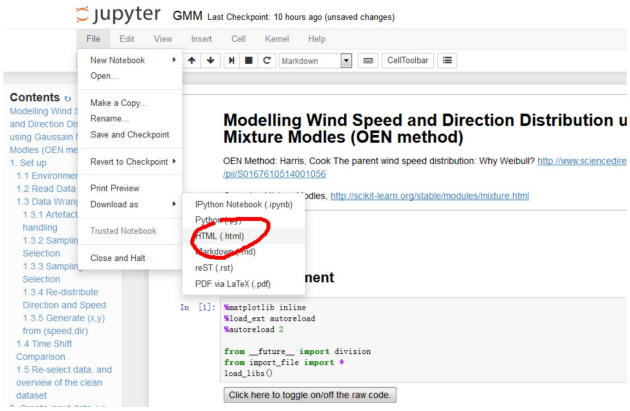
Implementation	**70 pts**
Your report	**20 pts**
Additional features	**10 pts**

6. Teaming

Students must work in teams of 2 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserve the right to assign different grades to team members depending on their contributions. So you should choose partner carefully!

7. Deliverables

(1) The HTML version of your notebook that includes all your source code. Go to "File" and then "Download as". Click "HTML" to convert the notebook to HTML.



5 pts will be deducted for the incorrect file format.

(2) Your report in PDF format, with your name, your id, course title, assignment id, and due date on the first page. As for length, I would expect a report with more than one page. Your report should include the following sections (but not limited to):

- **Problem Statement**
- **Methodology**
- **Experimental Results and Analysis**
- **Task Division and Project Reflection**

In the section "**Task Division and Project Reflection**", describe the following:

- who is responsible for which part,
- challenges your group encountered and how you solved them
- and what you have learned from the project as a team.

10 pts will be deducted for missing the section of task division and project reflection.

(3) A text file "additional features.txt" to describe and claim credit for additional features if there is any.

All the files must be submitted **by team leader** on Canvas before

4:00 pm, Wednesday, March 18, 2020

NO late submissions will be accepted.

8. In-class Demo:

Each team member must demo your work during the scheduled demo session. The following is how you should allocate your time:

- Model design (1 minute)
- Findings/results (1 minute)
- Task division (1 minute)
- Challenges encountered and what you have learned from the project (1 minutes)
- Additional features (1 minute)

Failure to show up in demo session will result in **zero** point for the project.

9. Hint:

Cheatsheet is provided for data/image preprocessing.

10. Think beyond the project

- Can you use transfer learning?
- How can we further improve the prediction accuracy? Example, does it help if we treat four images separately without combing them together as input? Is there a better way to remove outliers? How can we use zipcode better?
- A research paper: Eman H. Ahmed, Mohamed N. Moustafa, House price estimation from visual and textual features. Can you compare your best model with their model?

  https://arxiv.org/pdf/1609.08399.pdf