Deep Learning Based Predictions for Smart Buildings

Technical Report
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Danny Nguyen

May 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: Danny Nguyen

Approved: _____ Date _____
Haiying Shen, Department of Computer Science

Technical Report

For the 2019 Spring semester independent research I began by working with Haoyu for the first few weeks. His area of research dealt with the Hadoop framework and its efficient data storage (HDFS) and parallel application running (YARN/MapReduce). Hadoop is very important for the management of big data with many node clusters being able to increase the processing power. The Hadoop ecosystem is useful for the IoT and the processing of the mass amounts of sensor data from smart devices/buildings. The installation of Hadoop was quite challenging for me as I was unfamiliar with the Linux environment: the management of environment variables, bash files and command line usage. Once I successfully installed a single node cluster, I began to work on processing some sample data and producing the word count using MapReduce. It was at this point I was informed that I would not be able to assist Haoyu efficiently on his large-scale project, but was able to work on the data processing component.

For the remainder of the Semester I then worked with Huiying on the Smart Building project. The project consisted of using many sources of data in order to provide some buildings at UVA the ability to set the temperature based on occupants' preferences, predict the number of occupants' in a room to set up the cooling and heating automatically, and predict the events to be held in conference rooms. Some of the data sources used include: the UVA academic calendar, CS office calendar, weather data, and UVA event list calendar. It was my job to process the raw data from these sources, namely a CS faculty member's calendar, in order to produce structured data for the machine learning algorithm to predict occupancy data and other factors of prediction.

The data pre-processing tasks involved using python in conjunction with excel to process the data. A large task at hand was learning the basics of python throughout the semester as I had never used python before. At the same time as learning python, I learned about several python packages to assist in the processing of the spreadsheets, openpyxl, pandas, and numpy. My first task involved using excels sort/title tools and a simple VBA script to assign the time interval, day type, and other csv data to the corresponding event. I have attached the head of the excel file below.

	Title	Start	End	start_date	start_time	end_data	end_time	interval	week	weekend/weekday?		misc	misc	misc	misc
2	CS Faculty Meeting	4/2/18 12:00	4/2/18 13:00	4/2/2018	12:00:00 PM	4/2/2018	1:00:00 PM	1	Monday	weekday		12:00:00	PM		
3	Aaron Block	4/2/18 13:00	4/2/18 14:00	4/2/2018	1:00:00 PM	4/2/2018	2:00:00 PM	1	Monday	weekday		1:00:00	PM	1:00:00	PM
4	dMIST	4/3/18 13:00	4/3/18 14:00	4/3/2018	1:00:00 PM	4/3/2018	2:00:00 PM	1	Tuesday	weekday		1:00:00	PM	2:00:00	PM
5	Grizzly Peak	4/4/18 13:00	4/4/18 14:00	4/4/2018	1:00:00 PM	4/4/2018	2:00:00 PM	1	Wednesday	weekday		1:00:00	PM	2:00:00	PM
6	helena call	4/5/18 13:00	4/5/18 14:00	4/5/2018	1:00:00 PM	4/5/2018	2:00:00 PM	1	Thursday	weekday		1:00:00	PM	2:00:00	PM
7	Prep	4/6/18 13:00	4/6/18 14:00	4/6/2018	1:00:00 PM	4/6/2018	2:00:00 PM	1	Friday	weekday		1:00:00	PM	2:00:00	PM
8	CS/ECE 4457	4/7/18 13:00	4/7/18 14:00	4/7/2018	1:00:00 PM	4/7/2018	2:00:00 PM	1	Saturday	weekend		1:00:00	PM	2:00:00	PM
9	YAAHOO - Common House	4/8/18 13:00	4/8/18 14:00	4/8/2018	1:00:00 PM	4/8/2018	2:00:00 PM	1	Sunday	weekend		1:00:00	PM	2:00:00	PM
10	Office Hours	4/9/18 13:00	4/9/18 14:00	4/9/2018	1:00:00 PM	4/9/2018	2:00:00 PM	1	Monday	weekday		1:00:00	PM	2:00:00	PM
11	Weekly Meeting	4/10/18 13:00	4/10/18 14:00	4/10/2018	1:00:00 PM	4/10/2018	2:00:00 PM	1	Tuesday	weekday		1:00:00	PM	2:00:00	PM
12	Yonghwi	4/11/18 13:00	4/11/18 14:00	4/11/2018	1:00:00 PM	4/11/2018	2:00:00 PM	1	Wednesday	weekday		1:00:00	PM	2:00:00	PM
13	Professoring with Ben	4/12/18 13:00	4/12/18 14:00	4/12/2018	1:00:00 PM	4/12/2018	2:00:00 PM	1	Thursday	weekday		1:00:00	PM	2:00:00	PM
14	Josie	4/13/18 13:00	4/13/18 14:00	4/13/2018	1:00:00 PM	4/13/2018	2:00:00 PM	1	Friday	weekday		1:00:00	PM	2:00:00	PM
15	Brenda	4/14/18 13:00	4/14/18 14:00	4/14/2018	1:00:00 PM	4/14/2018	2:00:00 PM	1	Saturday	weekend		1:00:00	PM	2:00:00	PM
16	Prep	4/15/18 13:00	4/15/18 14:00	4/15/2018	1:00:00 PM	4/15/2018	2:00:00 PM	1	Sunday	weekend		1:00:00	PM	2:00:00	PM
17	CS/ECE 4457	4/16/18 13:00	4/16/18 14:00	4/16/2018	1:00:00 PM	4/16/2018	2:00:00 PM	1	Monday	weekday		1:00:00	PM	2:00:00	PM
18	buy fractus antennas	4/17/18 13:00	4/17/18 14:00	4/17/2018	1:00:00 PM	4/17/2018	2:00:00 PM	1	Tuesday	weekday		1:00:00	PM	2:00:00	PM
19	Exam	4/18/18 13:00	4/18/18 14:00	4/18/2018	1:00:00 PM	4/18/2018	2:00:00 PM	1	Wednesday	weekday		1:00:00	PM	2:00:00	PM
20	Link Lab All Hands	4/19/18 13:00	4/19/18 14:00	4/19/2018	1:00:00 PM	4/19/2018	2:00:00 PM	1	Thursday	weekday		1:00:00	PM	2:00:00	PM
21	Smart Cities Meeting	4/20/18 13:00	4/20/18 14:00	4/20/2018	1:00:00 PM	4/20/2018	2:00:00 PM	1	Friday	weekday		1:00:00	PM	2:00:00	PM
22	Rodrigo	4/21/18 13:00	4/21/18 14:00	4/21/2018	1:00:00 PM	4/21/2018	2:00:00 PM	1	Saturday	weekend		1:00:00	PM	2:00:00	PM
23	Daniel	4/22/18 13:00	4/22/18 14:00	4/22/2018	1:00:00 PM	4/22/2018	2:00:00 PM	1	Sunday	weekend		1:00:00	PM	2:00:00	PM
24	CS Faculty Meeting	4/23/18 13:00	4/23/18 14:00	4/23/2018	1:00:00 PM	4/23/2018	2:00:00 PM	1	Monday	weekday		1:00:00	PM	2:00:00	PM
25	SIF Smart Infrastructure	4/24/18 13:00	4/24/18 14:00	4/24/2018	1:00:00 PM	4/24/2018	2:00:00 PM	1	Tuesday	weekday		1:00:00	PM	2:00:00	PM
26	Nirupam	4/25/18 13:00	4/25/18 14:00	4/25/2018	1:00:00 PM	4/25/2018	2:00:00 PM	1	Wednesday	weekday		1:00:00	PM	2:00:00	PM
27	Grizzly Peak	4/26/18 13:00	4/26/18 14:00	4/26/2018	1:00:00 PM	4/26/2018	2:00:00 PM	1	Thursday	weekday		1:00:00	PM	2:00:00	PM

Another task I was given was to match the label from one sheet to the corresponding event on the other and write it to the spreadsheet. I first combined the two workbooks into one xlsl file and imported the data into pycharm; next, using the openpyxl software I parsed through the data and mapped the corresponding labels. Sample code is attached below.

```
import openpyxl
wb = openpyxl.load_workbook('match_startandard.xlsx')
#print(wb.sheetnames)
Sheet1 = wb['Sheet1']
brad = wb['bradjc5_cal']

#numData['C1'].value = 8
S1 = Sheet1.max_row
bra = brad.max_row
#print(ND)
#print(brad.cell(row=1,column=1).value)

for x in range(2,bra):
    for y in range(2,S1):
        if Sheet1.cell(row=y,column=1).value == brad.cell(row=x,column=7).value:
            brad.cell(row=x,column=8,value=Sheet1.cell(row=y,column=2).value)


#brad.cell(2,8).value = Sheet1.cell(2,2).value


wb.save('match_startandard.xlsx')
```

Another example of a data processing task was finding the location of a certain event given keywords to map onto such as “in” or “at. Attached is the code.

location.py - C:\Users\Danny\Documents\python\location.py (3.7.2)

File Edit Format Run Options Window Help

```
import openpyxl


#filelocation = 'C:\Users\Danny\Documents\locations.xlsx'
wb = openpyxl.load_workbook('locations.xlsx')
# print(wb.sheetnames)
#Sheet1 = wb['Sheet1']
brad = wb['bradjc5_cal']

# numData['C1'].value = 8
#S1 = Sheet1.max_row
bra = brad.max_row
# print(ND)
# print(brad.cell(row=1,column=1).value)

for x in range(2, bra):
    s = brad.cell(row=x,column=1).value
    if s.find("in ")!=-1:
        brad.cell(row=x, column=9, value=s[s.index("in ")+3:])
    if s.find("at ")!=-1:
        brad.cell(row=x, column=9, value=s[s.index("at ")+3:])
    if s.find("Link Lab")!=-1:
        brad.cell(row=x, column=9, value="Link Lab")

wb.save('locations.xlsx')
```

I worked mainly on creating automated scripts to process the calendar data, and provided an API for weather data prediction. However, I did extensive research on literature review and the creation of the summaries for several articles, which I will briefly outline. The first set of articles were related to the work of the smart building project, namely occupancy detection.

The papers aimed to identify which ambient sensor is the most effective in recognizing human presence. Four different types of sensors from two manufacturers were deployed to collect the following data reliably: illumination, temperature, humidity, levels of carbon dioxide, pressure and sound from within one staff office. It was concluded that the top three dominant sensors to recognizing indoor human occupancy are CO2 rate, illumination level and sound rate sensor, and the most dominant one is CO2 rate. Decision Trees can be used to explore the relationship between sensor types, features, and occupancy, so they were used to perform the classification and to explore the relationship between different types of sensors.

The second set of articles did not deal directly with the goals of the smart building project but with factors directly affect natural disaster evacuation behavior. Evacuation behavior refers to how residents act after receiving evacuation orders or information about imminent risks. This article focuses on the connection between evacuation choices and social capital, as measured by metrics describing Facebook users' social ties (social media data). Another studied the case of Florida's 2004 hurricane season. The paper used logistical analysis to study the affect of demographic factors (age, gender, location) on evacuation behavior, but more interestingly focused on choice of evacuation lodging (family or friends, hotel or motel, public shelter) and storm strength as factors. The last article studied the evacuation behavior in South Korea with the novelty of using evacuation drill knowledge as a factor. The influence of demographic and socio-economic characteristics on evacuation behavior was also considered. The results reveal that an evacuation drill experience is positively related to making a decision to evacuate.

The last tasks of the semester involved using the machine learning sklearn package, in order to predict occupancy data. First, I used pandas data frames to split the "Start Time, Day Status, Day, Event Name, Location" columns in dan.csv (given data) as the input variables X and "Count" as the Y- which needs to be predicted. Next, I used the "train_test_split″ method to divide dan.csv(given data) into test data sets and training data sets (X/Y_train and X/Y_test). I then constructed a Decision Tree Classifier to use the fit method to predict the count using the training and testing sets. Lastly to check how accurate the prediction of count was, I ran accuracy_score on Y_test and the prediction which yielded a score close to 1, indicating it was a good predicter. The code is attached below.

```
import numpy as np
import openpyxl
import xlrd
import pandas as pd
df = pd.read_excel('dan.xlsx', sheet_name='dan')
#print(df.head())
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

dfX = df[df.columns[0:5]]
dfY = df[["Count"]]
#print(dfY.head())
X_train, X_test,Y_train,Y_test = train_test_split(dfX,dfY, test_size=0.2)

#print(len(Y_test))
#print(len(Y_train))
#print(len(dfY))

clf = DecisionTreeClassifier(random_state=0)
crossval = cross_val_score(clf, dfX, dfY, cv=10)

count_pred = clf.fit(X_train, Y_train).predict(X_test)
print(count_pred)
print(accuracy_score(Y_test, count_pred))
print(crossval.mean())
```

In conclusion, I worked briefly with the Hadoop ecosystem, but transitioned to performing tasks to assist the smart building project including: pre-processing occupancy data, doing literature review, and touching on using machine learning to predict occupancy.