Justin Mateo
CSCE 580
7 December 2025

CSCE580 Final

Q1a:

a) Towards Enhancing Road Safety in South Carolina Using Insights from Traffic & Driver-Education Data. Sai Teja Paladi

b) The paper concludes that two way divided roads with barriers should be increased in urban counties and that educational programs such as Alive at 25 should be more widespread in rural counties. An example for the first action would be to consider an urban county in SC, such as Richland which the paper states has one of the highest collision rates. A certain road in it may have frequent head-on collisions and fender-benders due to it being a two way road that is non-divided. After implementing a barrier converting it to a two-way divided road, the number of collisions decreases over the following year.

At the same time a rural county such as Aiken, which on the collision heatmap has a moderate number of collisions, is seeing a continued rise in teen and young adult car collision incidents. Participation in the Alive@25 program has also been decreasing along the same timeline, which may correlate with the increase in collisions. Expanding awareness of the Alive@25 program in an attempt to boost participation could result in lower collision rates for the county following an increase in participation.

c) The conclusion of the paper is supported in my examples because it shows that the infrastructure improvements such as installing barriers on non-divided roads, could directly reduce collision risks in urban areas. Similarly, an increase in awareness and promotion of the Alive@25 program in rural areas could lead to fewer collisions in those communities, supporting the paper's observation that driver education participation is declining in these areas.

Q2 Preface:

Notes: I tried using MiniCPM-V-4_5, LlaVA, and GPT-4o through HuggingFace transformers and the OpenAI API with no success. I ended up prompting GPT directly using their interface. I outline my attempts in the colab notebook but I will also copy it below for reference.

Also, the attendance sheet for Class 23 is missing, and there are two class 25 images:

Nov13-20251118_094251.jpg

20251120_152601.jpg

So analysis returns 26 classes, when really it is 27

From my colab notebook:

NOTE:
I attempted to use MiniCPM-V-4_5, LlaVa through huggingface transformers library, and finally GPT-4o through the OpenAI API.

MiniCPM-V-4_5: was not able to prompt the model

LlaVa: Ran out of GPU resources (it used 22GB of RAM which maxed out the L4, the paid GPU through student account) I attempted to remedy this by quantizing the images using bitsandbytes, but both attempts failed to run with the error that the library on this runtime was not the latest version, despite trying to manually update it here.

GPT-4o: It said I hit the API limit despite it only being 27 images

I removed the code for MiniCPM and LlaVa when I tried to switch to GPT, but since that is not working either, I have decided to leave the code in to show there was an attempt.

Due to the time constraints, I am resorting to prompting GPT directly using their interface.

Q2 Answers:

a) I did not prepare the actual images, but I did clean the csv data output by the model. First I imported all of the csvs to the notebook through google drive, and I combined them all into one csv for ease of analysis. Next I checked the data for duplicate entries, empty values, and the formatting of the date. These dates were in string format so to make it easier to analyze I converted the dates to datetime format.

b) I ended up using the pre-trained GPT model by prompting it directly using their interface. I prompted a few images at a time with the prompt “extract csv files from these attendance sheets with: name, username, attendance index number, and the class number and date”. It output csv files that I manually skimmed through to check for inaccuracies, and the model was very accurate.
Initially I attempted to use MiniCPM-V-4_5 through huggingface’s transformers library, but was unable to prompt the model. Using the same library I attempted to use LlaVa, and this model maxed out the GPU’s RAM. To remedy this I attempted to quantize the

images using bitsandbytes, but this failed due to an error saying that the library on the colab notebook was not the latest version, despite me trying to manually update it. Lastly I attempted to use GPT-4o through OpenAI's API but it somehow maxed out the API limit, which is why I ended up prompting GPT directly through their interface.

c) The answers to this question are on my google colab notebook under the header "Q2c on attendance data: a,b,c,d"

d) If I had more time I would have preprocessed the images more to try and improve the performance. This would include but is not limited to converting images to greyscale, denoising, testing different image resolutions to find a balance between size and accuracy, and more. I would try different prompts to see what gives the most accurate extraction from the images, although GPT was very accurate with some typos. Similarly to ProjectB I could try to compare different models' accuracy in extracting the data. Finally I could try to fine tune a model specifically for these attendance sheets, although the dataset would be pretty small.