Austin Vornhagen

CSCI 8300

12/07/2023

Final Report

Discussion:

a. I created the dataset by combining a Kaggle dataset with images I found on Google. I annotated the dataset using the website Labelbox. I created my own ontology that I used to annotate the images. All images were annotated manually. The code is written in Python. I used the Python libraries Json, NumPy, cv2, PIL, io, TensorFlow, and keras. All the code is executed in Google Colab. I had to upgrade to the PRO version to take advantage of the A100 and V100 GPUs and have more system and GPU RAM available to train and execute all the models.
b. My program imports the data from a .ndjson file, a new line delimited JSON. I then parse the data just to get what the models will use and store the data in Python lists. Each step then has additional unique preprocessing to format the data for the models. I am using neural networks for all the models. The classification models use convolutional, batch normalization, max pooling, dropout, and dense layers. The pricing model is more of a deterministic function because I did not have any data on pricing. I am using preprocessing functions to perform feature engineering. I am not sure if I could have combined all the models into 1 since I was having issues with memory and compute. The data is sometimes normalized between 0 and 1 to better train the models. The displayed results are done with matplotlib.
c. The models are running without error but I would not say they are working as in they are not accurately predicting where the damage is on the exterior and interior of the car and how much it will cost to repair. I did not realize how long annotating the images takes. Initially, I planned to annotate 1000+ images but it was taking me 2-5 minutes per image to complete the annotation. That means it will take between 33 and 83 hours to annotate 1000 images and that is after I spend the time to find that many images. I encountered issues with getting the data to a format that is consumable by the neural networks. I am still having trouble getting the neural networks to produce correct results. One major constraint is the RAM available through Google Colab. I have a hard stop at the RAM limit enforced by Google Colab. I decided to pay money to train and execute these models. Even if I do everything right with the models, I do not think I have an adequate amount of data to tell if they are working correctly or get the results I want.
d. I do not think there are any assigned discussion problems.