Megan Shah

August 13, 2020

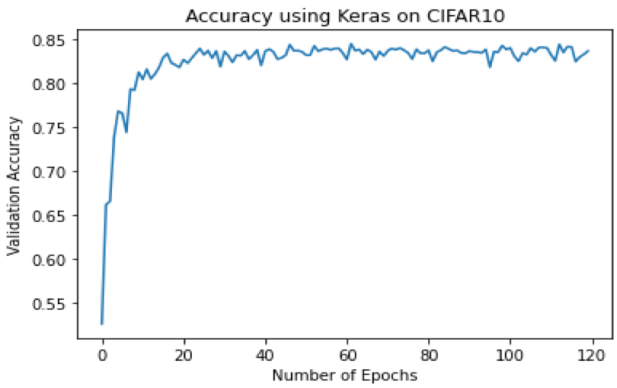
UCF CRCV HSAP 2020 Report

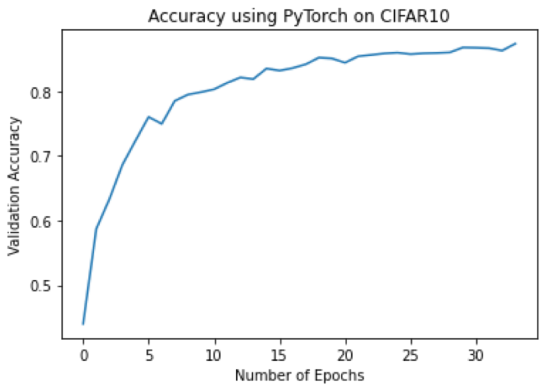
During the summer of 2020, I have had the pleasure of being one of two high school interns at the University of Central Florida Center for Research in Computer Vision (UCF CRCV). When I applied to the High School Apprenticeship Program (HSAP), I had little experience with machine learning and computer science and only some basic knowledge in Python. I was looking forward to understanding more about what computer vision entailed, its uses, and how it could be applied to biomedicine; however, by the end of the program, I realized that I had learned so much more than I had hoped. Through the guidance of my mentors, Dr. Shah and Robert Browning, I have been introduced and challenged in the field of computer vision.

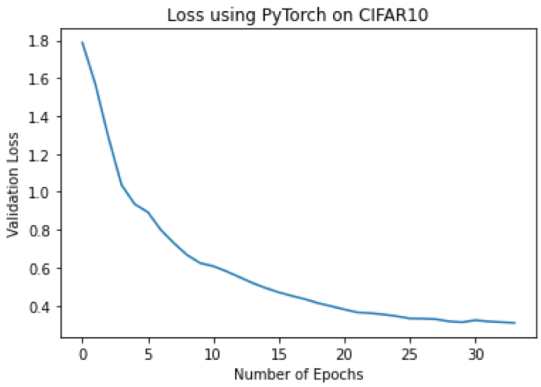
As an introduction to the internship, for the first two weeks of the program, I was given preliminary lectures regarding conceptual knowledge that was needed prior to conducting research. This included learning about *Convolutional Neural Networks*, image convolution using kernels, and image processing. Through these discourses, I learned the fundamentals of computer vision, as well as its importance in the 21st century. I was also given lectures explaining *PyCharm* installation using *Anaconda*, Python commands, and packages that would be used to consolidate, analyze, and display data. For the length of the first week, I worked to run Assignment 0, which entailed running a given code in Python using a *Keras* package to display the *MNIST* dataset, consisting of pixelated images of numbers. Upon finishing this assignment, I worked to further understand the *Keras* package using Assignment 1. This assignment involved creating multiple layers of *Convolutional Neural Networks* in *Keras* to sort the *CIFAR10* dataset. Using 120 epochs, I experimented using both the *SGD* and *Adadelta* optimizers to see which optimizer would result in a higher validation accuracy and lower validation loss. Ultimately, I found that using SGD optimizer I was able to achieve a much higher validation accuracy and lower loss, but running the epochs took around 13 seconds longer. Moreover, I was able to attend the phD defense of Rodney Lalonde, in which I learned about the benefits of *capsule networks* in medical imaging.

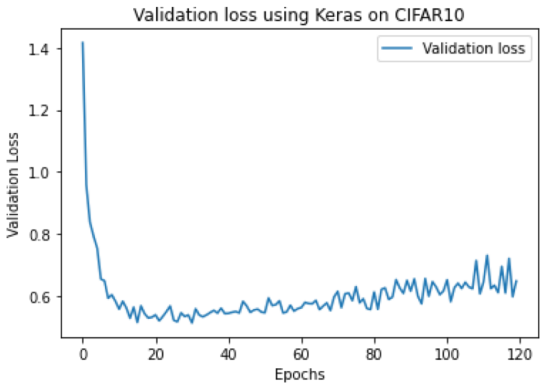
In the second week of the internship, I was introduced to *Deep Learning* and to another Python package called *PyTorch*, an open source machine learning library specifically designed for computer vision and *Deep*

Learning techniques. During the second week, I also worked on developing code for Assignment 2 to run the same *CIFAR10* dataset, but using *PyTorch*. Although the intended function and result was the same in both Assignment 1 and Assignment 2, the code was significantly different because of the differences in the coding style. I also used *MatPlotLib*, a Python package, to develop code to graph my validation accuracies and losses in both assignments, displayed below. I found that *PyTorch* required a lower number of epochs to achieve the same level of stability. Throughout the week, I also was able to attend three presentations and one phD defense by Harish Raviprakash where I learned about brain image analysis and object segmentation. The first lecture was given by Aidean Sharghi on *Video Textual Synopsis,* where I gained a conceptual understanding of video summarizations and how programmers use user preferences to generate a specific summarization with a hierarchical model. I also attended a presentation by Aisha Urooj in which I learned about *attention*, or attaching an "emotional significance" to certain objects based on user preferences, and developing models for image and video captioning. The last presentation that I attended was with Krishna Regmi. During this lecture, I was given a conceptual understanding of the geolocation project that I would be collecting data for. This project involves converting ground level videos of cities to aerial views and vice versa. I was explained the criteria for collecting both ground and aerial videos as well as an overview of what processing them after collection would entail.









Megan Shah

August 13, 2020

During the third week, I was able to use the information I had learned in Krishna Regmi's presentation to collect over 140 ground and aerial videos for the geolocation data collection project. I classified and organized the videos using an Excel spreadsheet in which I included the YouTube ID, publisher name, and date that the video was published. In addition, I attended a virtual live presentation given by Dr. Aidean Sharghi, in which he further elaborated on some of the topics he had mentioned in his pre-recorded phD defense that I watched in Week 2. This presentation was much more in depth and allowed me to fully engulf myself in the intricacies of video summarization. I learned about synopsys and summarization techniques and how long hours of footage can be condensed into textual or frame-by-frame summary using user-query information. I also learned about hierarchical structured models as well as their benefits and drawbacks. By the middle of the third week, I started reading a research paper about video description techniques called "Video Description: A Survey of Methods, Datasets, and Evaluation Metrics," in which I learned about video description and its applications in human-robot interaction, image captioning, video captioning, and dense video captioning. I also began to understand what terms like attention and transformation methods entailed.

In weeks 4 and 5, I attended a virtual graduate student workshop hosted by Dr. Shah, where I learned more about the benefits of attending graduate school, teaching and research positions offered by professors to help graduate students gain experience, as well as the plethora of fellowships available for graduates to continue researching topics of their choice. I also continued with data collection for the geolocation project and collected a total of over 250 ground and aerial videos. Building off of the reading I did last week, I continued to read the research paper about video description in which I began to grasp what machine translation and image captioning metrics like BLEU, ROUGE, and METEOR meant. I acquired an in-depth background of the three phases of Video Description architecture-- the Classical, Statistical, and Deep Learning phases. The Classical phase involved object recognition, human action detection, and lastly sentence generation, while the Statistical phase was useful for extremely large datasets, like open domain datasets.

By the time of week 6, I was introduced to a new project that I would be working on called the UTRAP project during a presentation given by Rohit Gupta and Xiaoyu Zhang. This project was designed to fool a

computer vision model to mis-classify an object using a white box adversarial technique. After installing *Adobe AIR* and *BULKR*, I was able to mass select and download around 2000 relevant images in 7 main classes. During this week, while also continuing to collect data for the geolocation project and reading the research paper, I attended the Research Experiences for Undergraduates (REU) Spotlight presentations and poster sessions. Through the session, I was able to watch brief overviews of 11 undergraduate students' summer computer vision projects. As with last week, I also continued on with the research paper and learned about the importance of encoding and decoding in the *Deep Learning* literature phase. The paper discussed the three popular methods of encoding and decoding, using a combination of *Convolutional and Recurrent Neural Networks*, as well as newer *deep reinforcement networks*, which have become increasingly popular through companies like Google's Deep Mind. *Deep reinforcement models* are considerably different from their counterparts because the model's interactions with the environment are designed to be state-based, or dependent on previous interactions and because the models do not overfit, overgeneralize, and there is no lack of datasets.

In the last week of my internship, I culminated my geolocation data collection, finished reading the research paper, and concluded the new UTRAP collection of images. Throughout week 7, I continued collecting images for the UTRAP project mentioned in week 6. This project involved collecting 1000 images from four main classes—warplanes, tanks, submarines, and cannons—and I generated around 20 key terms for each class. I was also able to finish reading "Video Description: A Survey of Methods, Datasets, and Evaluation Metrics," in which I learned about the datasets that computer scientists use to train and validate their data, primarily in deep learning. The main datasets used are primarily broken up into four main categories: cooking, movies, videos in the wild, and social media. It is these datasets that determine the pace of research in computer vision. The cooking dataset (YouCook, MP-II Cooking, TACoS) consists of various closed and open domain datasets, which all contain annotated videos with various actions involving cooking and the manipulation of cooking ingredients. Movies datasets (M-VAD, MPII-MD) focus on extracting audio descriptions and short clips from Hollywood movies. Social media datasets (VideoStory, ActivityNet) involve multi-sentence descriptions and bounding boxes, while, lastly, videos in the wild (MSVD, Charades, VTW) use human annotated clips with complex

language to summarize indoor and outdoor activities. I also got a brief overview on some evhe opportunity to form so many wonderful connections. Using the knowledge that I have acquired this summealuation metrics that are performed over machine generated captions/descriptions. Finally, I concluded my data collection for the geolocation project by collecting 40 new ground and aerial videos which contributed to the 400 total videos I collected.

This internship has been a massive learning opportunity for me and I would like to extend my utmost gratitude to my mentors and presenters for their patience, thorough explanations, and guidance. Although it was challenging at times, I have found my experience at the University of Central Florida Center for Research in Computer Vision to be extremely informative and rewarding. Despite the different set of circumstances this year, my knowledge in deep learning and computer vision has been amplified tremendously and I am looking forward to continuing my research in computer vision under Dr. Shah in the upcoming fall and spring.