Sea Turtle Conservation Final Report

Training Cammy using CogUniversity

Charles Richardson

Fall 2021

Abstract

Training an AI requires a handful of consideration. Using the proper inputs will set the project off on a good foot, but maintaining proper input throughout can prove to be difficult. For an AI to be consistent and accurate, using dozens or hundreds of inputs is insufficient. Thousands or even millions of inputs is where the results no longer require constant oversight. An AI only knows what it is provided, asking an AI for something is was not taught about will also result in unpredictable behavior. However, after surpassing these two hurdles and creating a clear set of standards for input and output, an AI can assist with any properly implemented problem, saving time, money and stress. This project has highlighted all of this to me, and has brought AI to the front of my mind for future project ambitions.

Intro and Problem Statement

For this project, the objective was to train an AI (Cammy) to understand many properties of a turtle. This was expected to be done with hundreds to thousands of images with labels. To train the AI, we were tasked to use the Nvidia HyperGator SuperComputer and the online platform CogUniversity. The properties that Cammy was expected to learn were as follows:

- General questions about turtles
- Species of turtles
- Weight of turtles

Assumptions

Cammy began knowing nothing about turtles, and took images and tags as input to learn about their properies. As more data was fed to Cammy, it was able to make better predictions on unknown data. Presenting Cammy with non-turtle images could dramatically alter its performance, it is not recommended. Within the each of the properties Cammy was expected to learn, there were strong limits to what Cammy was capable of. For the general questions, they included diet, habitat, size, lifespan, birthing process and navigational capabilities. Anything outside of this range would return unexpected answers. For the species type, Cammy was limited to knowing four. Finally, for the weight, the range was unbounded.

Input Data

Training Cammy took most of the time for this project. Just like teaching a human on a topic, I would have been required to know the information I am providing Cammy so that it can learn it properly. Fortunately, the most of the data were pre-labeled, which made my job much easier.

For the general questions, fourty questions were provided for which I had to provide forty responses. Many of the questions were similar, which helped with answering, as I could simply copy and paste.

The species type input took a bit more time as we were specified to use a labelling software which took some time to get sixty images through. The species identification was made simple as the image files were preceded by a letter indicating their species type. The species of sea turtle were green, loggerhead, ridley and hawkley.

The weight of the turtle was pre-processed into a file which required little heavy lifting. One mass file push into the HyperGator was all Cammy needed to learn this.

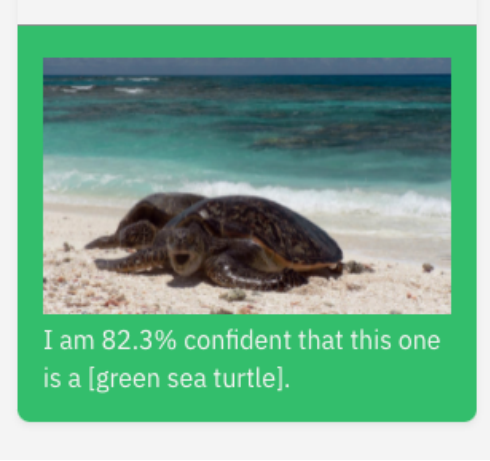
Model Conceptualization

Training Cammy was a tedious process of answering unique prompts in unique ways. This is done to make sure Cammy detects patterns between the prompts and answers and is able to create a reliable model of what is being learned, the downside to this is that it is a time consuming process when done manually.

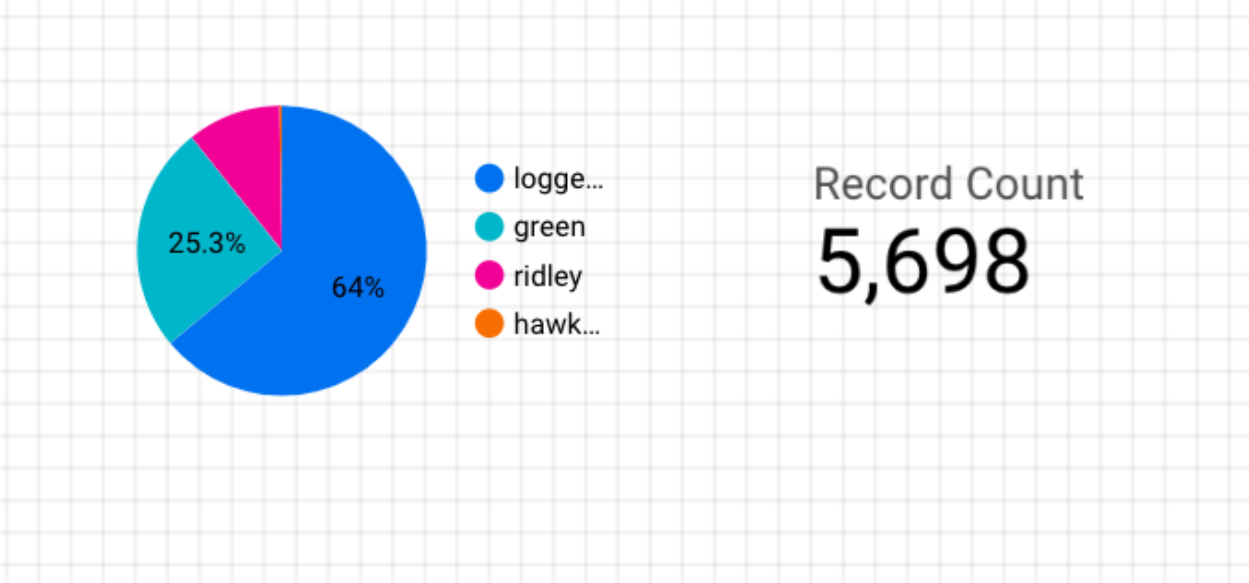
The way Cammy was able to learn from pictures and labels was through regression analysis. The mathematics behind this topic are outside the scope of this class, but the purpose is to detect as many variables (patterns) as possible and find similarities between them across multiple instances, to determine what makes up the provided object.

Model Presentation

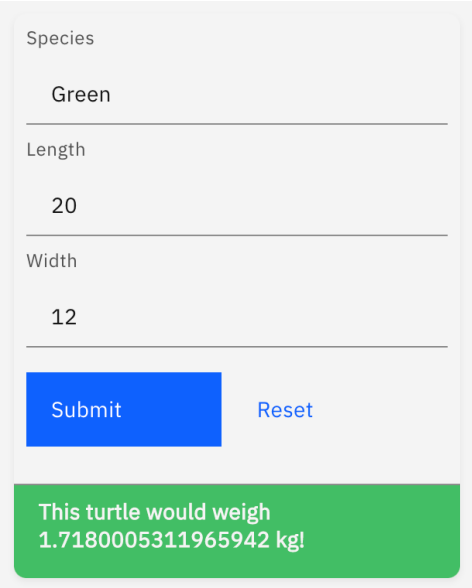
Training an AI takes a large amount of data. Provided with sixty images, Cammy was able to be about 80% certain that unfamiliar images of green sea turtles were truly sea turtles, however, when provided an image of a poop emoji, it was over 50% confident it was a loggerhead! Therefore, with sixty images, we cannot trust that Cammy is able to accomplish many insights productively.



A correct prediction, yet only 82% confident it is a green sea turtle. The low confidence here could be a result of two turtles being in the image, when Cammy is only provided one turtle per image.



The model weight data was much more extensive than the images, outnumbering it by nearly 100-fold. Equipped with this extensive set of data, Cammy provided much more accurate results for acceptable data points



An acceptable weight for a well fed green sea turtle of its size.

Verification and validation

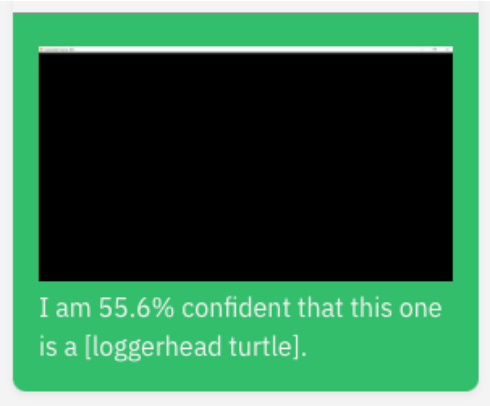
Verifying the output of Cammy was straightforward for the first 2 parts of the project. The questions were easy to understand and detect anomalies. The test images were pulled from

google images which gave me access to the species type before testing it. Therefore, verifying that Cammy was working correctly and validating the output was not difficult for these parts.

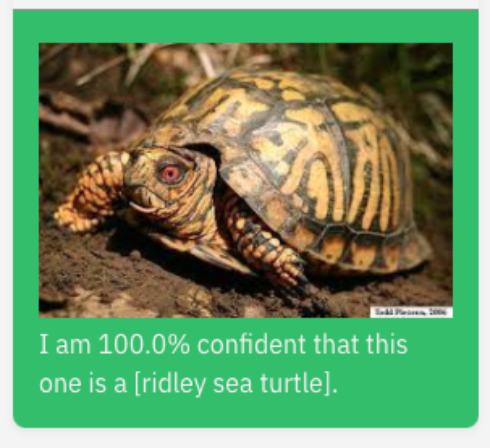
Confirming the weight output took a bit more effort. I used similar values to those provided in the file, then compared the weights to see if they aligned with Cammys predictions.

Experimentation

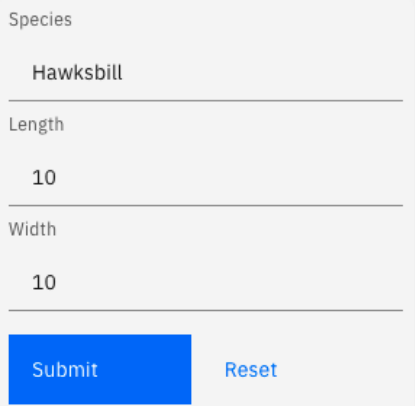
Testing outlying data on Cammy revealed how important it is to put error detection into AI to prevent erroneous results. Here was the outcome of the experiments:



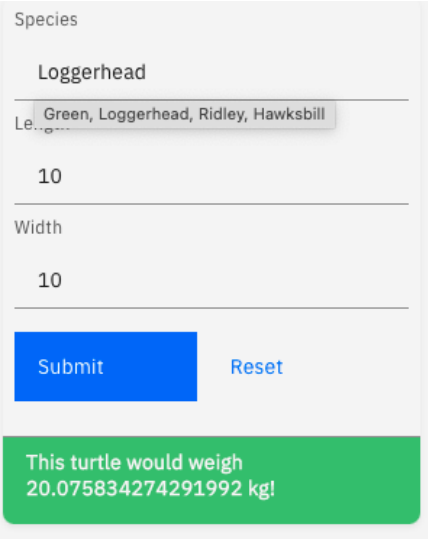
Purely incorrect prediction here, yet outside of the scope of Cammy's knowledge. Cammy was only trained to know about four species of turtles. It was not trained to know about non-sea turtle images. This results in erroneous results like this one.



The high confidence of this prediction must come from the similarities between species. For this is an Eastern Box Turtle. Similar to the previous image, Cammy was not trained to know about anything apart from the four species mentioned above.



Negative weight defies the laws of physics, as does a 10x10 turtle. Cammy did not receive data of any turtle that was as long as it is wide, which is why this offsetting result is returned.



The same data as above but a different species. The result of this could be from different datasets for the different species. Another possibility is that since height isnt included, Cammy assumed that the turtle stopped eating seagrass and started eating icecream, which contributed to massive weight gain!

Summary and recommendations

Through my experience, I found that the best way to train Cammy for a specific target audience would be to have that target audience train Cammy herself. In other words, have a few representatives of the audience fill out the answers to the questions in a way that would have pertain to the entire target audience. Furthermore, AI needs tons of training to be accurate. Dozens of data points are not enough. Hundreds are not either. Thousands may suffice, sometimes. Ultimately, tens or hundreds of thousands of data points is what is ideal. While millions to billions of data points are optimal for pristine accuracy.

Any tool that makes working with AI easier is a great addition to the industry. Using AI for jobs over people makes jobs faster, easier and simpler, in most cases. Therefore, the tools that are developed to automate the use of AI are without a doubt, great.

When large amounts of data are pooled into Cammy's system, it is important to consider the implications. In the case of sea turtles, there is not much that can be done to hurt the integrity of the system. However, considering the previous project for this class, where we had to differentiate trash and marine life, biases in that data could result in many animals losing their lives to a "smart" system.

References

University of Florida, “Coguniversity,” *CogUniversity*. [Online]. Available: https://cu2-uf-cap1.mybluemix.net/#/en-us/research/cogability. [Accessed: 21-Nov-2021].

Appedices

No further details to include.