**Graded Discussion 4**

Bias is a significant concern when training machine learning models. In a technical sense, bias is the error arising from capacity limitations of the machine learning model. In other words, the limited ability of a model to fit the training data.

An important, and often overlooked or ignored, source of bias is the limitations in the training data themselves. A common source of such bias is having an important class or category that is significantly under-represented in the training data. These situations occur in applications ranging from social media, autonomous vehicles to fairly mundane industrial applications. The results can range from embarrassing (my dog is not a skunk!) to dangerous (autonomous vehicle not recognizing a child).

Think about some case you have either recently encountered or read about where biased training data has led to dangerous or unethical behavior by a deep neural network agent. What problem did you perceive? What steps can you think of that might mitigate this problem?

Make a post of about 100-300 words here in Piazza. Then post a comment of suggestion on another students post of about 100-200 words.  Posts are due by May 10. Do not post respond in Canvas.

One of the systems that lately has gained more visibility is the potential for bias in facial recognition software used in many sectors of society now days. There are several issues that have been brought to the attention of the public, one of them is described by PhD Joy Buolamwini (MIT Computer Scientist) while working on a project using one of the most popular and widespread commercial facial recognition software and found that the software was not able to detect her face as she is an African Canadian black woman. Initially as explained by herself in her TED talk (How I'm fighting bias in algorithms | Joy Buolamwini), she thought it was a fleeting/small bug in the software, but later in other situations she encountered the same situation and after looking into the problem found that one of the issues was the bias in the model given by the fact that the model was trained in a dataset that did not include many (significant) people with particular characteristics such as darker skin tone. Her research uncovered large gender and racial bias in AI systems sold by companies like IBM, Microsoft and Amazon, with detection error rates of no more than 1% for lighter skinned men and 35% for darker skinned women.

Machine learning is being used for facial recognition, but its uses as seen in class extends beyond computer vision. Who gets hired or fired? Who gets a loan or does not get a loan? Who gets admitted into college? All these questions are now getting inside the real of machine learning and it is important to start paying attention to any bias in the algorithms that could be expressed by discriminatory practices that are uncovered by algorithmic and mathematical justifications. This is a very complicated problem and there is no immediate right solution, but one of them as expressed by Doctor Buolamwini is to pay close attention to the data ingested by the model and understand the implications of data imbalances (bias) that could be seen in our algorithms inputs and look for ways to compensate for that if it is not possible to just obtain new data.

How I'm fighting bias in algorithms | Joy Buolamwini (<https://www.youtube.com/watch?v=UG_X_7g63rY&list=WL&index=3>)

“AI Racial Gender Bias” <https://time.com/5520558/artificial-intelligence-racial-gender-bias/>

“Racial Discrimination in Face of Recognition Technology” <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>

“Facial Recognition Technology is Both Biased and Understudied” <https://www.media.mit.edu/articles/facial-recognition-technology-is-both-biased-and-understudied/>

“What is Responsible AI” <https://towardsdatascience.com/what-is-responsible-ai-548743369729>