CS 287: Ethics Module
Assignment

Suppose you are asked to produce an image captioning software that employs machine learning. One of the challenges you face in the process is related to the generation of gender-specific caption words. To perform this task, you ought to make a choice between two models.

(1) The first model relies on learned priors based on the image context. It exploits contextual cues to determine gender-specific words.
(2) The second model generates gender-specific words based on the appearance of persons in the scene. This model incorporates an equalizer, which ensures equal gender probability when gender evidence is occluded and confident predictions when gender evidence is present. Further, it limits gender evidence to the visual aspects of persons.

For each of these models answer the following questions.

a. Can this model perpetuate gender biases? How?
b. Can this model amplify gender biases? How?
c. If the answer is yes, do these biases constitute harmful stereotypes? Why?

For the *second model*, answer the following questions:

a. Mention two demographic groups who are rendered vulnerable to harmful biases.
b. Can you prevent the software from incorporating these biases? How?