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.
We assume that the training data are a random sample of images on the internet.

a. Can this model perpetuate gender biases? How?

1. Yes, this model can perpetuate gender biases. The model will learn to associate men and women with the contextual cues that appear most often in their photos. These cues will reflect biases between men and women (i.e. men photographed more often at work and women photographed more often at home). When making predictions, this model will have higher error rates when men and women appear in atypical contexts with respect to their gender, thus perpetuating these biases.
2. Yes, this model can also perpetuate gender biases. Just as the first model will learn associations between gender and contexts, this model will learn to associate gender with appearance. When a man or women does not fit this archetype (i.e. women in business clothes or men wearing make-up), error rates will be higher. The model with consequently perpetuate these biases. This issue is not ameliorated by the equalizer. Since many aspects of one's appearance is associated with gender, we expect that the equal probabilities will not be employed often.

b. Can this model amplify gender biases? How?

Both of these models can amplify gender biases. When models are deployed at scale, the decisions they make can have societal impact. A person making biased decisions with societal implication will lead to these biases being amplified – an algorithm is no different.

c. If the answer is yes, do these biases constitute harmful stereotypes? Why?

To the extent that all gendered language will fail to account for individuals who do not conform to the gender binary, these models (not just their biases) will be harmful. With respect to the biases specifically, they are harmful to the extent that perpetuating these biases hurt individuals of that gender. For instance, perpetuating and amplifying gender biases related to employment will limit economic opportunities for women.

For the second model, answer the following questions:

a. Mention two demographic groups who are rendered vulnerable to harmful biases.

Women are vulnerable to biases that may be learned from data because historically, they have had less opportunity, respect, and safety – historical data will likely reflect this inequity. Individuals who do not conform to a gender binary may be particularly vulnerable because captioning images with gendered language will, by definition, fail to recognize their identity.

b. Can you prevent the software from incorporating these biases? How?

One approach is to incorporate regularizers in one's loss function to penalize the model for perpetuating harmful biases. Another approach is to encourage the model to create captions that do not include gender-specific words in the first place, although this is easier in English than in other languages. This could be accomplished via post-processing with a rule-based system (i.e. she/he → they, man/woman → person, etc.) or by limiting a model's vocabulary.