Ethics Module

Roshan Padaki
rpadaki@college.harvard.edu

Michael Zhang
michael_zhang@college.harvard.edu

April 10, 2019

1 Model 1

Model 1 relies on learned priors based on the image context. It exploits contextual cues to determine gender-specific words.

(a) This model can easily perpetuate gender biases. The learned priors mean that, given a dataset exhibiting a gender bias, the model will tend to

(b) This model can certainly amplify gender biases. Given a biased dataset, the model will default to the bias in the dataset when contextual cues are weak, and subscribe to the contextual priors when the cues are strong. Hence, the biases at test time might be amplified from the biases at training time.

(c) These biases do constitute harmful stereotypes. In particular, the model is liable to develop an overreliance on its learned priors, which is especially undesirable when testing on unbiased datasets or on new domains.

2 Model 2

Model 2 generates gender-specific words based on the appearance of persons in the scene. The 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.

(a) This model can perpetuate gender biases; supposing the training set exh ibits a biased correlation between appearance and gender pronouns, the model will exhibit the

(b) This model as presented in the readings is designed to correct against amplifying gender bias. Intuitively, their correct method, denoted as an Equalizer, is able to make confident gender predictions under sufficient evidence, and importantly expresses confusion or uncertainty when this is not the case by optimizing an Appearance Confusion Loss. Accordingly it is is unlikely to amplify bias in situations where the evidence does not hold.

(c) Due to the equalizer, these biases are less likely to constitute harmful stereotypes, as the model is designed to not excessively rely on any learned priors.

3 Second model questions

(a) Demographic groups especially vulnerable to harmful biases include groups for which gender presentation (and thus the visual aspects of a person) do not subscribe to the cisgendered correlation with gender identity.

For example, men who present towards feminine are statistically a minority to men who present towards masculine; thus, a model training on visual cues is liable to learn harmful biases against this and other similar populations.

Another example is with racial minorities. Different racial minorities may exhibit different dominant means of gender presentation. However, given that the training data is easily liable to exhibit racial bias, the model might learn the dominant gender presentations of the dominant racial groups, thus potentially rendering racial minorities vulnerable.

(b) One way to correct for these biases is to control for factors such as representation in our training set, to prevent the model from learning from pure statistical biases in the data. That is, when biases such as racial discrimination are identified to result from the dataset exhibiting, say, a biased racial distribution, one fix would be to seek a more uniformly representative dataset. Another intervention that may not necessarily be useful is to increase the weighting of the