1. Can this model perpetuate gender biases? How?

   (a) (First model) Yes. Contextual cues have well-documented gender-biases—similar to those we observe in `word2vec`. For instance, the model might associate bosses with men and secretaries with women.

   (b) (Second model) Yes (though perhaps to a lesser extent). Gender biases can propagate through appearance. For instance, the model might learn the statistical association between wearing jewelry and being female. More perniciously, if the training data is everyday images, biases can also be learned since everyday images may not be a representative sample of (non-biased) human interaction. For instance, if women are disporportionately told to smile in the workplace, as a result of well-documented gender biases, and the training data consist of images of various workplaces, the model might associate smiling with femininity.

2. Can this model amplify gender biases? How?

   Both models can amplify gender biases if the non-causal statistical relationship the model learns is influential in making decisions downstream. For instance, the first model might underly some software to generate automatic templates for newspaper articles. The second might underly certain software to caption images. Both may amplify biases since biase propagate to downstream decisions that have real-world consequences.

3. If yes, do these biases constitute harmful stereotypes? Why?

   Yes. Associating women with secretaries and men with bosses certainly constitute harmful stereotype, as these positions have different social standing and are nonequivalent (learning the association between men and businessmen and women and businesswomen, on the other hand, is more admissible, as the positions are gendered but equivalent counterparts). With the second model, certain relationship between appearance and gender might be harmful stereotype, as the training data may contain harmfully stereotypical representations of gender (e.g. disportationately male businessmen and female secretaries). But certain associations also tend to be more innocuous, like perhaps associating smaller physique with women and larger physique with men.

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

   Women and people of color

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

   There has been some literature (Sendhil Mullainathan and Jon Kleinberg), whose point is that bias is machine learning models is fine so long as downstream decisions optimally debias; however, debiasing might be difficult to accomplish. One might also have humans assess the bias of certain language, and then have a model that learns to score text that might perpetuate gender biases, which can be a preprocessing-esque debiasing step for further learning.