READING ASSIGNMENT 7
16785-Integrated intelligence in robotics

Aarati Noronha-anoronha

1. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

As this paper points out, machine learning models are trained on datasets which are more often than not inherently biased. And during training, these biases tend to get amplified. For example, images showing an agent cooking have a greater tendency of being identified as women by a model. The same may be extended to an agent shopping. Such biases extend beyond gender to other social aspects as well.

This paper tests on two tasks namely semantic role labelling and multilabel classification tasks. It identifies bias amplification in existing data-sets and applies RBA- reducing bias amplification as a means of calibrating predictions. Zhao et. al applies corpus level constraints on a conditional random field and uses Lagrangian relaxation to optimize the target function.

Whilst being the first of its kind to recognize bias, a lot of improvements could be made. Other ways of measuring bias could be broached. Other domains could be explored like pronoun reference resolution as the author suggests. This algorithm can also be applied to other tasks like segmentation and detection. Another improvement could be trying to learn the user specified margin from the model.

2. Women also Snowboard: Overcoming Bias in Captioning Models

This paper focuses on fair caption generation when gender is slightly occluded or even otherwise. This forces the algorithm to take a closer look at the face of the person rather than pick up on contextual information to predict gender.

It relies on an Equalizer model which includes two types of losses: the appearance confusion loss (ACL) and the Confident Loss (CL). The former loss forces the model to generate neutral words such as person when an image does not have sufficient evidence for gender. Both the words-man and woman have equal probability during training when images of person are blocked out. In order to avoid confusion that comes from not giving into social biases, the CL loss is implemented. It encourages the use of gender specific words given enough evidence of gender in an image.The experiments were carried out on the MSCOCO dataset. Error rate, gender ratio and right for right reasons are the performance metrics. Grad-cam and sliding window saliency maps are positives to take from this paper.

This paper is one of the first to define such a problem. Its approach may extended to other avenues which have gender or social bias such as role labelling and segmentation. Steps must be taken to make the equalizer less cautious during labelling. It should spit out words like man and woman more often than the word person. This involves manipulating the former loss function.