Draft Proposal
Estimating Gender Bias in Machine Learning

Team Members

- Rahul Ethiraj [USC #3765791028]
- Sithara Kamalakkannan [USC #6524330088]

Inspiration

- http://norman-ai.mit.edu/
- http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
- https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Objective

The aim of this project is to determine if Gender Bias exists in the predictions of Machine Learning Algorithms based on the data it has been trained on, and to see to what extent it may or may not reflect the bias that we as humans have.

Candidate data sets for training include the FaceScrub dataset that has over 100K images of men and women, or the IMDB-Wiki dataset that includes 500K+ images with gender and age labels. Once the dataset has been determined, it will be processed to fit a machine learning algorithm.

Potential methods to analyze our problem description is for the dataset to be trained using Generative Adversarial Networks (GAN). This will be done using the HyperGan package on TensorFlow.

Below are the steps of our algorithm:

- The generator returns a random image from the huge dataset.
- This generated image is fed into the discriminator alongside a stream of images taken from the actual dataset.
- The discriminator takes in both men and women images and returns probabilities, a number between 0 and 1, with 1 representing a prediction of men and 0 representing women.

So we have a double feedback loop:

- The discriminator is in a feedback loop with the ground truth of the images, which we know.
- The generator is in a feedback loop with the discriminator.

Results will be obtained by either treating this as a Classification Problem and determining with what bias our GAN predicts men and women or by making our GAN generate images of men and women and detect how the algorithm “thinks” a man or woman would look like.