Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

Now a days, as we blindly train machine learning models without understanding what these models are learning, and just focusing on the accuracy of the downstream tasks, authors in their state of the art paper made a shocking revelation that even training word embeddings on google-news dataset exhibit gender bias to a great extent which is unacceptable. As learning word-embeddings is the starting point of many NLP downstream tasks it could amplify the biases and possess huge implications and risk.

Generally, word embeddings of similar words tend to be closer and follow simple arithmetic operations. This property is very helpful and has many applications in natural language to find hidden relationships. On further analysis, authors found that these word embeddings are biased and filter some gender neutral words as gender specific which could be very degrading and risky (eg computer programmer and homemaker can be both man and woman).

Authors exploit the property that gender neutral words should be equidistant with the gender definitive words but that is not the case with the learned word embeddings. Using this property, Tolga et al. showed that word embeddings having unequal distances imply gender bias and also propose a debiasing algorithm to mitigate it by identifying the gender subspace and performing neutralize or equalize operation. Authors through their experiment also showed that bias found in the word-embeddings strongly correlated with the bias found in the society.

I very much liked the title of the paper- it’s very catchy and intriguing to read the paper. I also wonder about other types of biases that exist and whether they can be identified and mitigated. On the other hand, the debiasing algorithm proposed by the authors was slightly weak and ineffective and could be improved. In Future, I would like to work on automating bias detection from the language models to build better moral models.

In conclusion, it was an eye opening paper which created a lot of buzz in the news because this paper showed machine learning has a long way to go and they also make errors.