deeplearning.ai

NLP and Word Embeddings

Debiasing word embeddings

Some of the ideas for diminishing or eliminating these forms of bias in word embeddings

bias mean gender, ethnicity, sexual orientation bias.

The problem of bias in word embeddings

Man:Woman as King:Queen

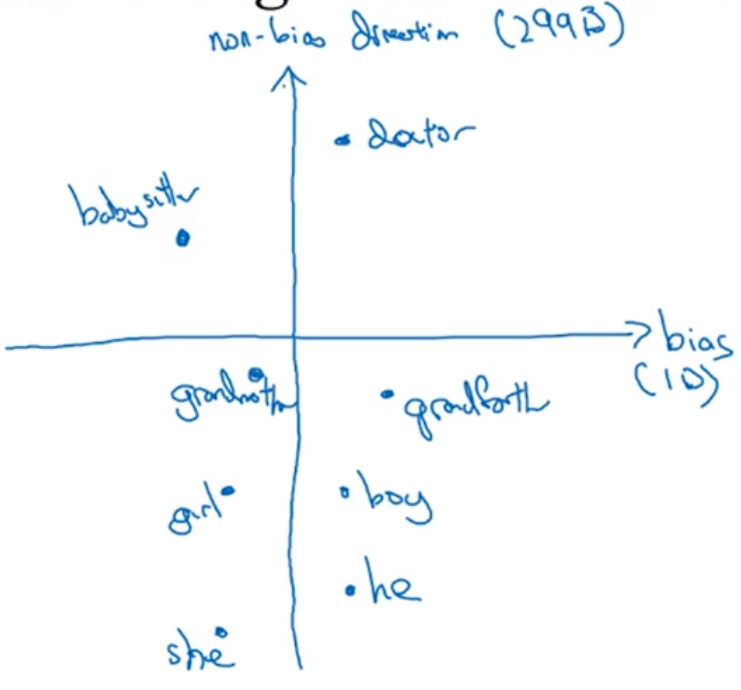
Man:Computer_Programmer as Woman:Homemaker

Father:Doctor as Mother:Nurse

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Addressing bias in word embeddings



1. Identify bias direction.

$e_{he} - e_{she}$

$e_{male} - e_{female}$

average

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

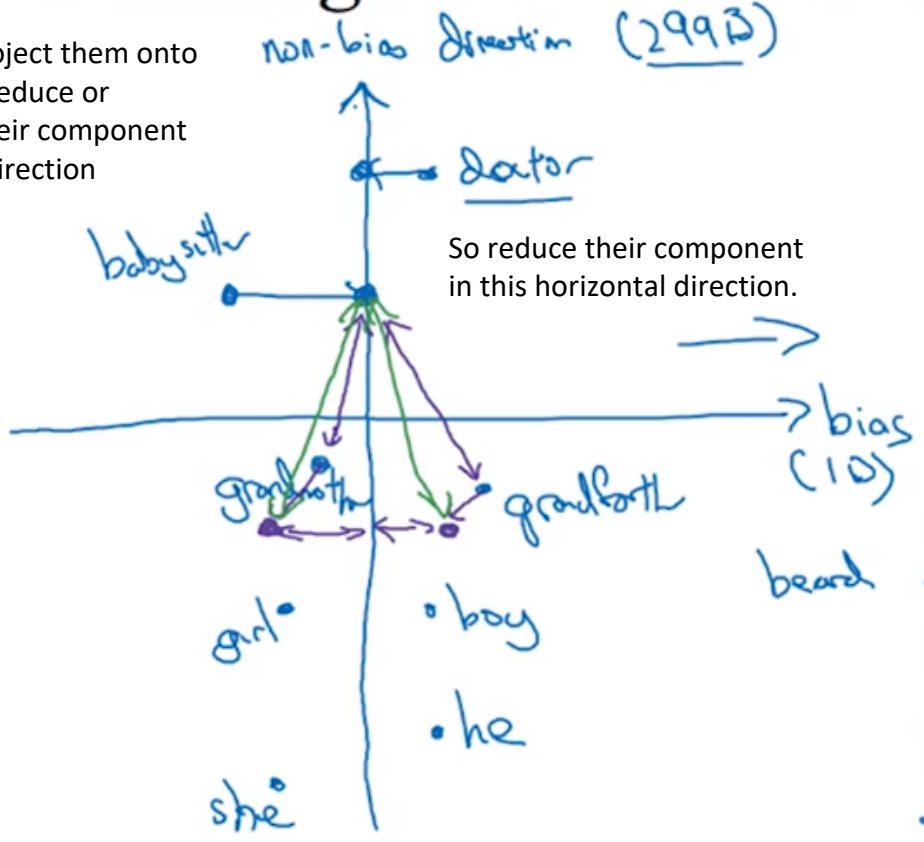
Andrew Ng

And so the first thing we are going to do is identify the direction corresponding to a particular bias we want to reduce or eliminate.

The bias direction can be higher than 1-dimensional, and rather than take an average as I describe it here, it's actually found using a more complicated algorithm called SVD(singular value decomposition), which is closely relate to, if you're familiar with principle component analysis, it uses ideas similar to the pca or the principle component analysis algorithm.

Addressing bias in word embeddings

Let's just project them onto this axis to reduce or eliminate their component in the bias direction



1. Identify bias direction.

$e_{he} - e_{she}$

$e_{male} - e_{female}$

⋮

→ average

2. Neutralize: For every word that is not definitional, project to get rid of bias.

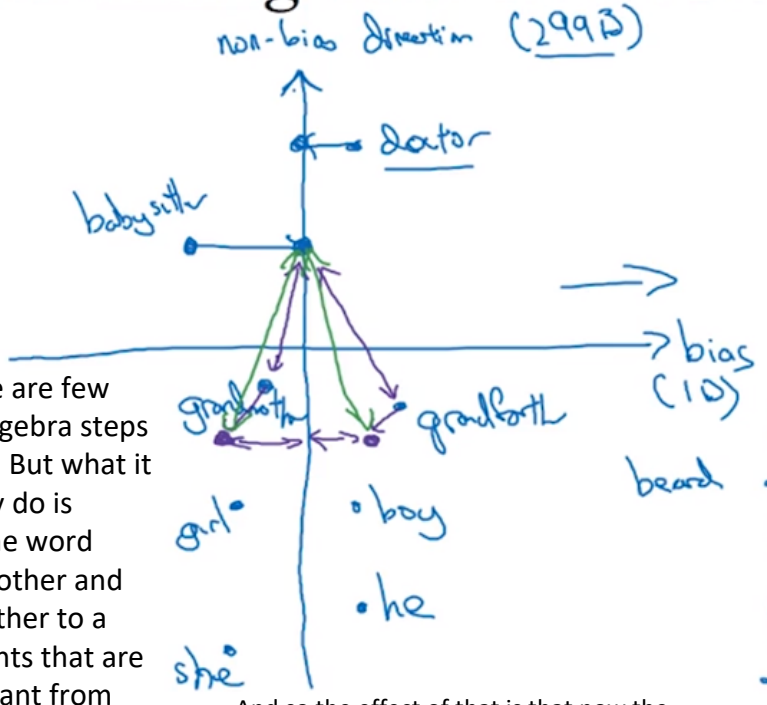
3. Equalize pairs.

→ grandmother − grandfather

girl boy

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Addressing bias in word embeddings



So there are few linear algebra steps for that. But what it basically do is move the word grandmother and grandfather to a pair points that are equidistant from this axis in the middle.

And so the effect of that is that now the distance between babysitter compare to these two word will be exactly the same.

1. Identify bias direction.

$e_{he} - e_{she}$

$e_{male} - e_{female}$

average

2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.

grandmother — grandfather

girl boy

So in the final equalization step, what we'd like to do is to make sure that words like grandmother and grandfather are both exactly the same similarity, or exactly the same distance from word that should be gender neutral.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Andrew Ng

How do you decide what word to neutralize? And so the authors did is train a classifier to try to figure out what words are definitional, what words should be gender-specific and what the word should not be. And it turn out that most words in the English language are not definitional, meaning that gender is not part of the definition. And it's just a relatively small subset of words like this,..., that should nor be neutralized.

And finally, the number of pairs you want to equalize that's actually also relatively small, and is at least for the gender example, it is quite feasible to hand-pick most of pairs you want to equalize.