Measuring social bias in knowledge graph embeddings

Link: <https://assets.amazon.science/96/d6/d6555711405cbc1463360700dad4/measuring-social-bias-in-knowledge-graph-embeddings.pdf>

**Motivation**

What is the problem being solved?

* Word embeddings encode gender and racial biases
* harmful effects in downstream tasks
* e.g. fact checking task
* e.g. knowledge graph embeddings get utilized as input to language model -> impact all downstream NLP tasks

Bias in embeddings

* prejudice in favor or against a person, group, or thing that is considered to be unfair

Knowledge Graph Embeddings

a score function g(.) which takes as input the embeddings of a fact in triple form and outputs a score, denoting how likely this triple is to be correct.

s = g(e1, r1, e2)

where

* e1/2 = the dimension d embeddings of entities ½
* r1 = the dimension d embedding of relation 1

Previous work

* “Word Embedding Association Test”: cosine distance between entity embeddings and the average entity embeddings of two sets of attribute words
* direction between vectors

Why is the previous work insufficient to solve the problem?

* Distance-based metrics are not suitable for measuring bias in knowledge graph embeddings
  + knowledge graph embedding models do not use distance between two entity embeddings when making predictions
  + it uses the distance between some transformation of one entity embedding with the relation embedding
* Associations in the distance space become less correlated with associations in the score function space

**Approach**

Score function: sj,p =g(ej,rp,ep)

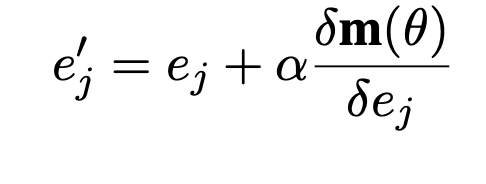
Bias: bp =g(ea male person,rp,ep)−g(ea female person,rp,ep)

* average such comparisons across multiple pairs of male/female entities
* the distribution of human entities for each profession in real-world knowledge graphs with sensitive attributes is not balanced
* only represent whether the model is able to give higher scores to people’s correct professions

Impose a gender on a person (gender = a/b)

m(θ) = g(ej , rs, ea) − g(ej , rs, eb)

Get updated embeddings

****

where

* e′j = the updated embedding for person j
* rg the embedding of the sensitive relation i (gender)
* ea and eb the embeddings of attributes a and b (male and female)

Check whether imposing a gender creates bias

∇p =g(e′j,rp,ep)−g(ej,rp,ep)

* Repeat the process for all humans in the knowledge graph -> calculate the average changes
* Does not involve complementing/altering the training data or training procedure
* Does not depend on the particular entity/set of entities chosen

**Results**

Wikidata

Data: more males entities than female entities

* Bias exists in supposedly-neutral professions
* There is a strong correlation between the counts and bias
  + Even in cases where counts are balanced between genders, bias still exists

FB3M

* The method is equally applicable to any transformation function
* Embeddings of professions encode potentially harmful social biases
* Proposed method of measuring this bias is effective at exposing the most gendered professions with the more sophisticated transformation function

**Contributions**

* ​​Demonstrated that differences in the distributions of entities in real-world knowledge graphs
* Show that the differences translate into harmful biases related to professions being encoded in embeddings
* knowledge graphs are formed of real-world entities -> cannot simply equalize the counts/ correct history

⇒ care is needed when applying knowledge graph embed- dings in NLP pipelines

⇒ need to develop robust methods to debias such embeddings

**Limitations**

* Comparison between two values of a sensitive attribute
* Assign people a fixed label for some attributes
* “sensitive attributes” have more than two potential attributes
* only present the model’s representation of the relationship between any two of these alternative values at a time

**Future work**

* Compare the differences in how bias is encoded for different transformation functions