Operations on Word Vectors

Yilun Kuang

09/18/2023

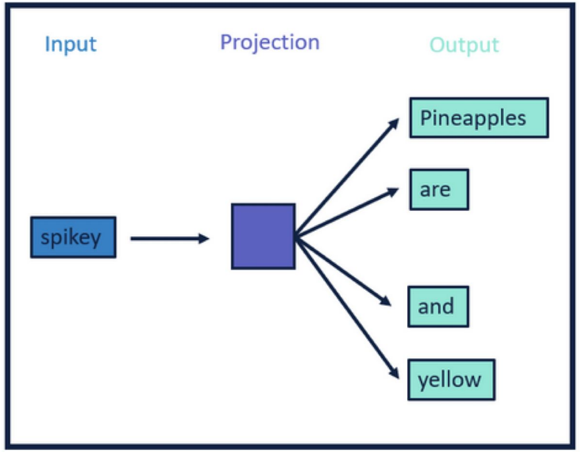
Recap

Goal: Map each word to a vector in $\mathbb{R}^d$ such that similar words have similar vectors.

Recap

Goal: Map each word to a vector in $\mathbb{R}^d$ such that similar words have similar vectors.

Skip-gram model: Given a word, predict its neighbouring words within a window.



Skip-gram

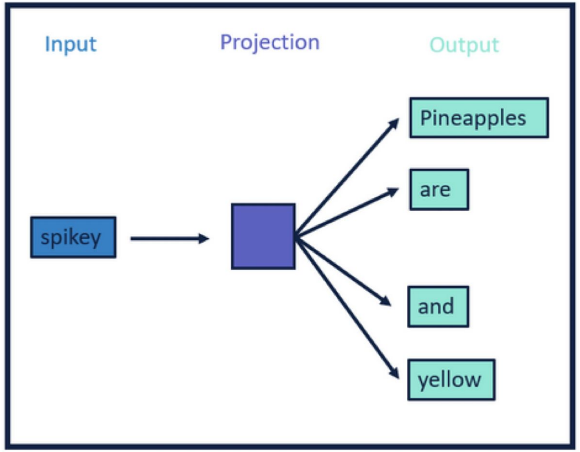
Source: https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285

Recap

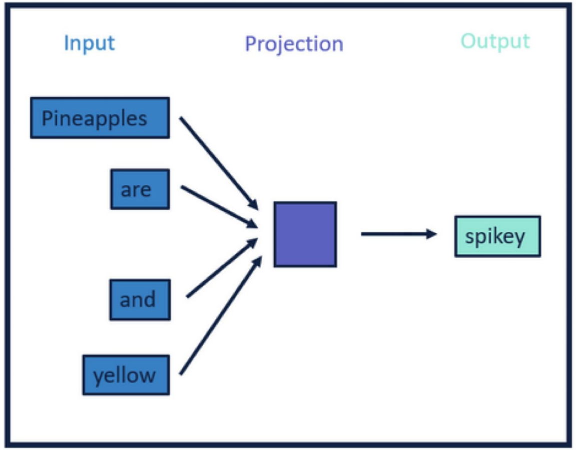
Goal: Map each word to a vector in $\mathbb{R}^d$ such that similar words have similar vectors.

Skip-gram model: Given a word, predict its neighbouring words within a window.

Continuous bag-of-words model: Given the context, predict the missing word.



Skip-gram



CBOW

Source: https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285

Recap

GloVe: Global Vectors (Pennington et al., 2014) — Use co-occurence matrix of each word pair. N(w, c) is the co-occurence count between word w and context c.

Recap

GloVe: Global Vectors (Pennington et al., 2014) — Use co-occurence matrix of each word pair. N(w, c) is the co-occurence count between word w and context c.

context vector

word vector

bias terms (also learned)

$$J(\theta) = \sum_{w,c \in V} \underbrace{f(\mathrm{N}(w, c))} \cdot (u_c^T v_w + b_c + b_w - \log \mathrm{N}(w, c))^2$$

Source: https://lena-voita.github.io/nlp_course/word_embeddings.html

Recap

GloVe: Global Vectors (Pennington et al., 2014) — Use co-occurence matrix of each word pair. N(w, c) is the co-occurence count between word w and context c.

context vector

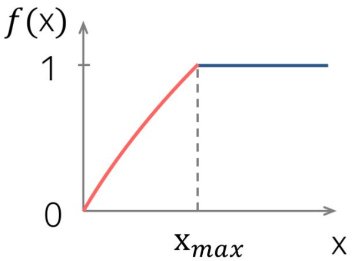
word vector

bias terms (also learned)

$$J(\theta) = \sum_{w,c \in V} f(\mathrm{N}(w, c)) \cdot (u_c^T v_w + b_c + \overline{b_w} - \log \mathrm{N}(w, c))^2$$

Weighting function to:

- penalize rare events
- not to over-weight frequent events



$$\begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max}, \\ 1 & \text{otherwise.} \end{cases}$$

$\alpha = 0.75$, $x_{max} = 100$

NYU

Source: https://lena-voita.github.io/nlp_course/word_embeddings.html

Similarity between Word Vectors

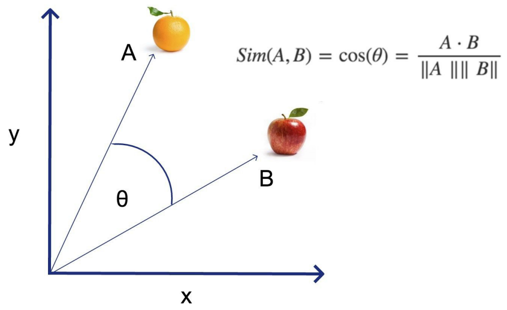
Question: Do the learnt word embeddings satisfy the desired property of similarity?

Similarity between Word Vectors

Question: Do the learnt word embeddings satisfy the desired property of similarity?

Use cosine similarity between any two word vectors.

Cosine Similarity



Source: https://www.linkedin.com/pulse/cosine-similarity-classification-michael-lin/

Word Analogy Task

In word analogy tasks, we ask questions like “a is to b as c is to ___”

Example: “London is to UK as Amsterdam is to Netherlands”

NYU

Word Analogy Task

For a —> b :: c —> ?, given word vectors v_a, v_b and v_c, we will find a word d such that $v_a - v_b \sim v_c - v_d$.

Word Analogy Task

For a —> b :: c —> ?, given word vectors v_a, v_b and v_c, we will find a word d such that $v_a - v_b \sim v_c - v_d$.

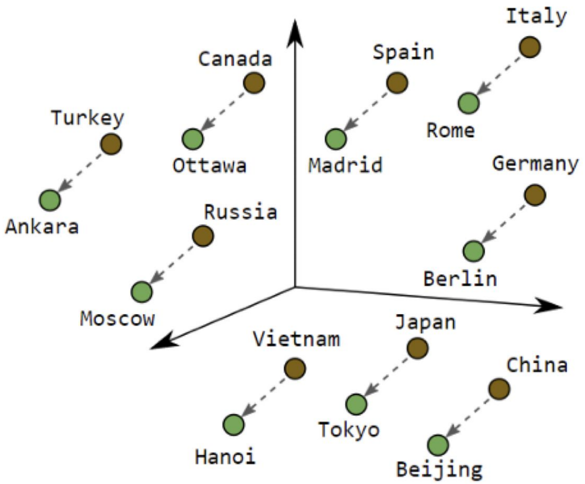
The difference $v_a - v_b$ represents the 'concept' (e.g. capital of country).

NYU

Word Analogy Task

For a —> b :: c —> ?, given word vectors v_a, v_b and v_c, we will find a word d such that $v_a - v_b \sim v_c - v_d$.

The difference $v_a - v_b$ represents the 'concept' (e.g. capital of country).



Country-Capital

NYU

Source: https://developers.google.com/machine-learning/crash-course/images/linear-relationships.svg

Word Similarity Tasks

- WordSim353
- SimLex-999 (similarity rather than relatedness)

Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

Source: https://developers.google.com/machine-learning/crash-course/images/linear-relationships.svg

Bias in Word Vectors

The difference $v_a - v_b$ represents the ‘concept’ — if a is woman and b is man, then it represents ‘gender’.

NYU

Bias in Word Vectors

The difference v_a - v_b represents the 'concept' — if a is woman and b is man, then it represents 'gender'.

Compute projections of occupations on this difference v_a - v_b

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

NYU

Bolukbasi et al. 2016

Bias in Word Vectors

Similarly, we can obtain vectors for other concepts like race and religion.

Compute projections of occupations on this difference $v_a - v_b$.

Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Note: The vectors were obtained from training on reddit data from USA users

NYU

Manzini et al., 2019

Debiasing Word Vectors

For a concept vector g and word vector e, obtain the biased component:

$$e_{\text{biased}} = \frac{e \, . \, g}{||g||^2} g$$

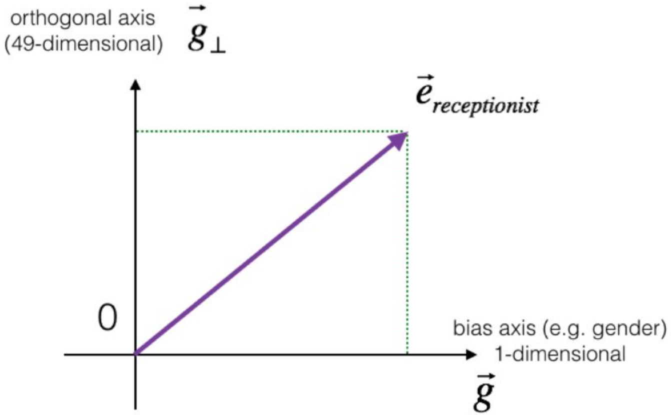
Subtract from the original vector to obtain the debiased vector

$$e_{\text{debiased}} = e - e_{\text{biased}}$$

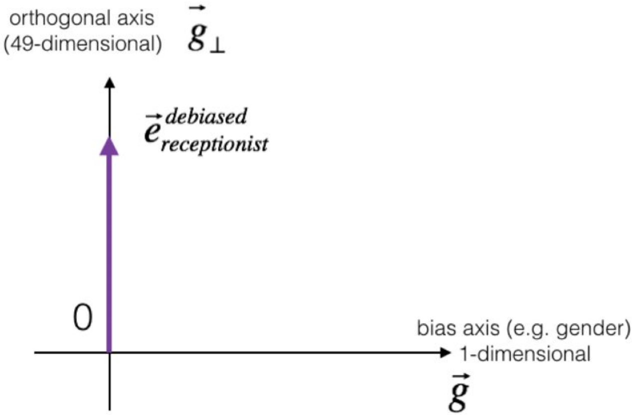
NYU

Manzini et al., 2019

Debiasing Word Vectors



before neutralizing,
"receptionist" is positively correlated with the bias axis



after neutralizing,
debased version, with the component
in the direction of the bias axis (g) zeroed out

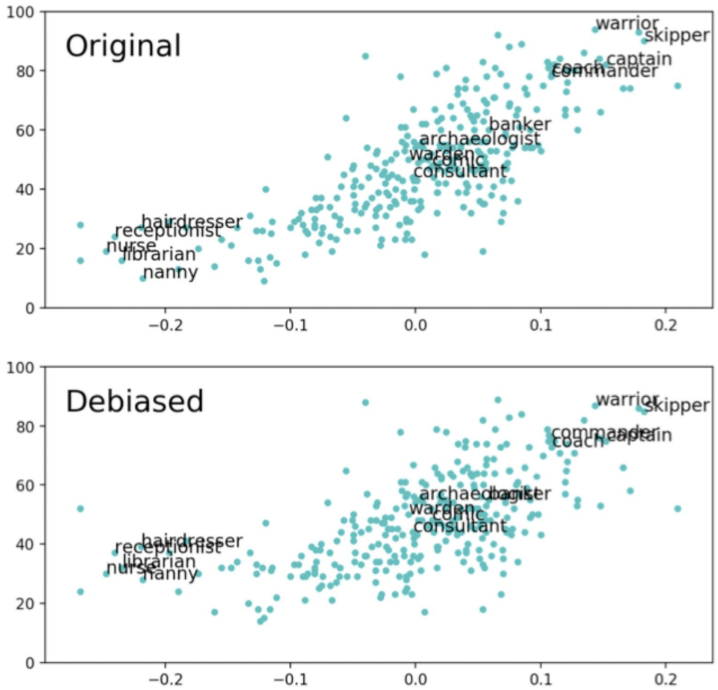
NYU

Source: https://zhangruochi.com/Operations-on-word-vectors-Debiasing/2019/03/28/

Debiasing Word Vectors

Previous method ensures that vector is orthogonal to the concept vector.

Not always effective in debiasing —- the word vectors corresponding to occupations are still clustered according to gender.



Gonen and Goldberg, 2019

Other Debiasing Methods

Ravfogel et al 2020:

- There is no single direction corresponding to concepts — it can span in multiple directions.
- Propose Iterative Null-space Projection (INLP) — iteratively neutralise/debias the vectors.

Other Debiasing Methods

- **Algorithm 1** Iterative Nullspace Projection (INLP)

  Input: (X, Z): a training set of vectors and protected attributes
  n: Number of rounds

  Result: A projection matrix P

  Function `GetProjectionMatrix`(X, Z):

  $X_{projected} \leftarrow X$
  $P \leftarrow I$
  for $i \leftarrow 1$ **to** n **do**
  $W_i \leftarrow$ TrainClassifier$(X_{projected}, Z)$
  $B_i \leftarrow$ GetNullSpaceBasis(W_i)
  $P_{N(W_i)} \leftarrow B_i Bi^T$
  $P \leftarrow P_{N(W_i)} P$
  $X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
  end
  return P

e.g. Dataset of (occupation, gender) where we have word vectors for each occupation along with the biased gender.

NYU

Other Debiasing Methods

-

Algorithm 1 Iterative Nullspace Projection (INLP)

Input : (X, Z): a training set of vectors and protected attributes
n: Number of rounds

Result: A projection matrix P

Function `GetProjectionMatrix`(X, Z)**:**

$X_{projected} \leftarrow X$
$P \leftarrow I$
for $i \leftarrow 1$ **to** n **do**
$W_i \leftarrow$ TrainClassifier$(X_{projected}, Z)$
$B_i \leftarrow$ GetNullSpaceBasis(W_i)
$P_{N(W_i)} \leftarrow B_i Bi^T$
$P \leftarrow P_{N(W_i)} P$
$X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
end
return P

→ e.g. Train a linear classifier to predict gender from occupation.

Other Debiasing Methods

-

Algorithm 1 Iterative Nullspace Projection (INLP)

Input : (X, Z): a training set of vectors and protected attributes
n: Number of rounds

Result: A projection matrix P

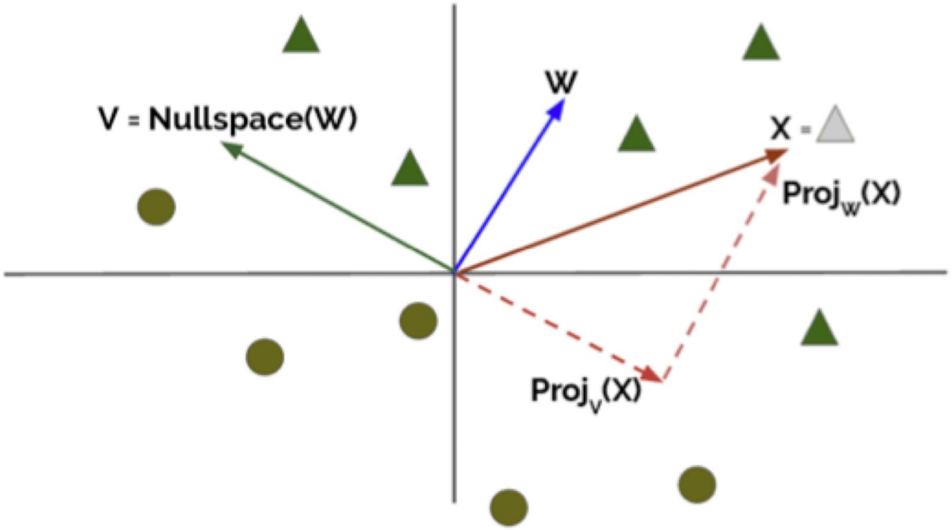
Function `GetProjectionMatrix`(X, Z):

$X_{projected} \leftarrow X$
$P \leftarrow I$
for $i \leftarrow 1$ **to** n **do**
$W_i \leftarrow \text{TrainClassifier}(X_{projected}, Z)$
$B_i \leftarrow \text{GetNullSpaceBasis}(W_i)$
$P_{N(W_i)} \leftarrow B_i B_i^T$
$P \leftarrow P_{N(W_i)} P$
$X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
end
return P

Project X onto nullspace of W —> predicting Z (e.g. gender) from new X will not work.

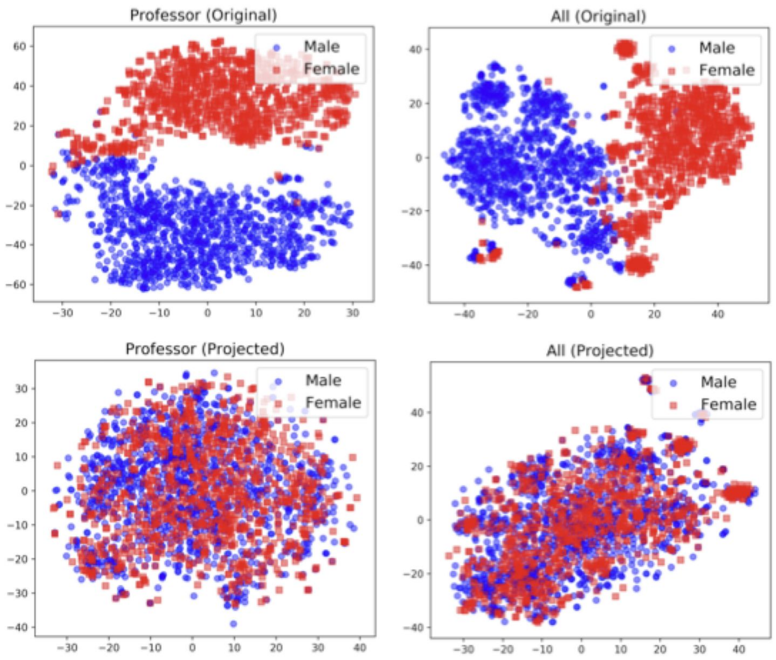
Other Debiasing Methods

- W: weight of a linear classifier trained to predict Z from X
- Project on null-space
- Iterate



Other Debiasing Methods

- Does not suffer from the issue we saw with earlier debiasing method.
- Representations are now not clustered according to protected attribute (e.g. gender).



NYU

Ravfogel et al., 2020

Summary

- Word vectors encode a notion of similarity, which can be helpful for retrieval, word analogy tasks etc.
- Word vectors can encode biases from the data —> Need to evaluate and use appropriate debiasing methods.

Acknowledgement

This presentation is adapted from Nitish Joshi, 31st January 2022