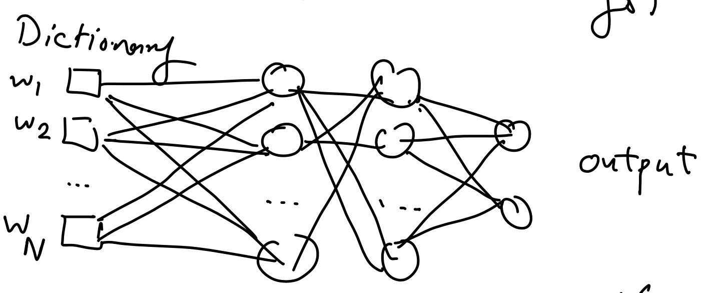
vector Representations (AKa Embeddings)

- Approach 1 X



X

- Sentence Structure X
- Semantic Sim. X
- Sentiment X
- ...
- You Cannot Piggyback on word Sim. for training the model

- approach 2:

allocate a vector of random numbers to each word

1	2	....		d
5	11		6	9

X does not Consider Semantic Similarities.

Vector Representations (Embeddings)

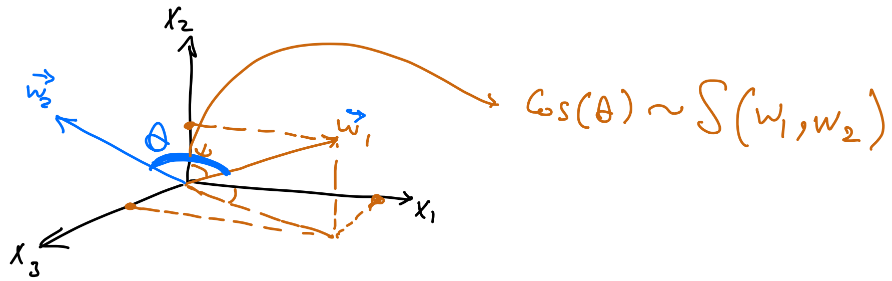
Given a Universe $U = \{u_1, \dots u_N\}$,

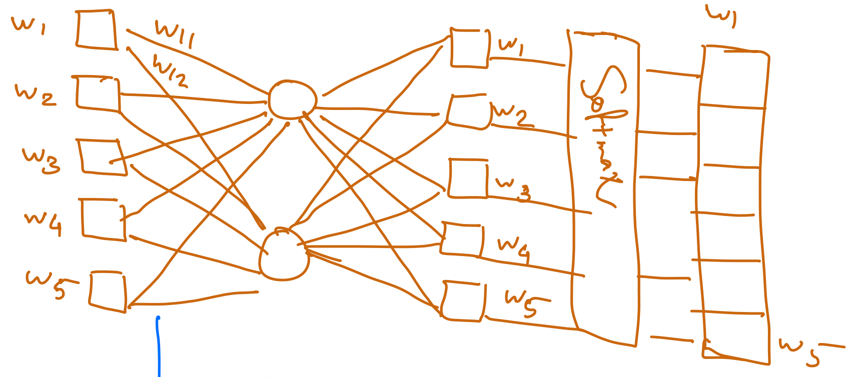
a function S, $S(u_i, u_i)$: Semantic Sim. b/ u_i, u_j

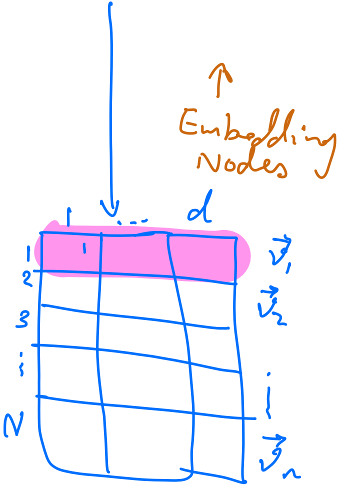
$\vec{v}_i = Vec(u_i)$ is a d-dimensional vector of numbers

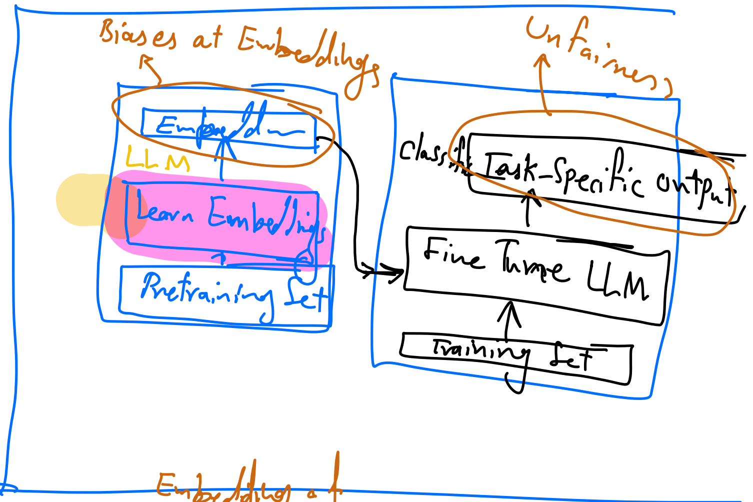
S.t.

$$\cos\angle(\vec{v}_i, \vec{v}_j) \sim S(u_i, u_j)$$



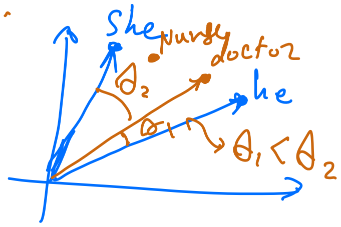


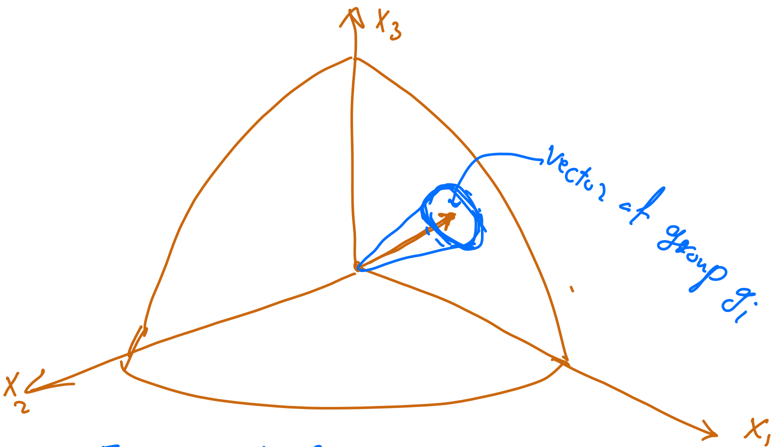




word Embeddings are biased if the Embedding of words that should be nuetral are more similar to one of the groups.

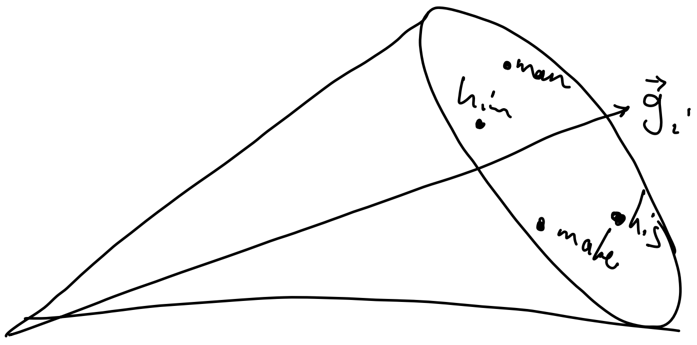
e.g.,





- Every word Semantically representing g_i (male, man, him, his...) should have a small angle $\vec{g_i}$, ⇒ Falling inside its cone.

Assumption → - Every Representative word, is an iid Sample from the Surface of the Cone



$$E[\vec{g}_i] = 1/K \sum \vec{v}_i$$

Given a Set of words $\{\underset{\downarrow \{\text{doctor, Eng, } \ldots\}}{A}, \underset{\downarrow \{\text{nurse, } \ldots\}}{B}\}$

$$\mathrm{WEAT} = \frac{1}{l}\left[\sum_{a_i \in A} \mathrm{Sim}(\vec{a_i}, \vec{g}_1) - \mathrm{Sim}(\vec{a_i}, \vec{g}_2) + \sum_{b_i \in B} \mathrm{Sim}(\vec{b_i}, \vec{g}_2) - \mathrm{Sim}(\vec{b_i}, g_1)\right]$$

$$\mathrm{Sim}(\vec{a}, \vec{g}) = \mathrm{Cos}\angle(\vec{a}, \vec{g})$$

$\Downarrow$ Extension To Sentences

Embedding $\mathcal{E}$: Sentece $\longrightarrow \vec{v}$

↳ Instructor.

measuring bias of a Sentence Embedding

Step 1: Finding Embeddings for groups

g_1 : male g_2 : Female

- Use a Predefined Set of Sentences:
  A) have a high association with g_i
  B) " a minimal (to Zero) Extra Information.

{"He is a boy", "His gender is Male",...}

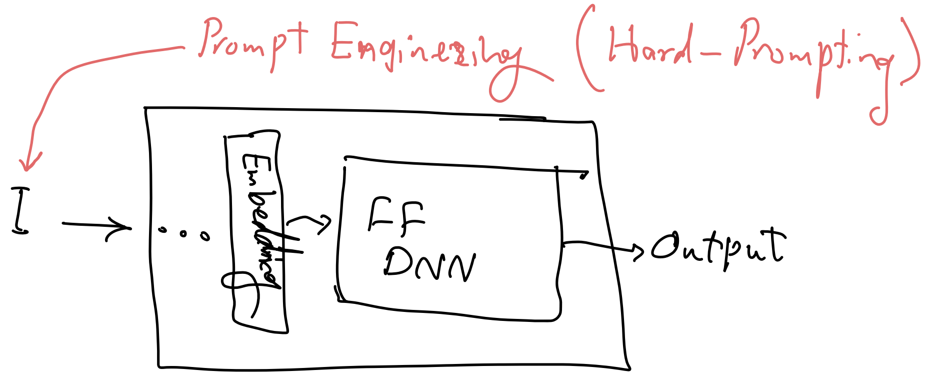
- Assume each Sentence is an IID Sample in the "Cone" of g_i

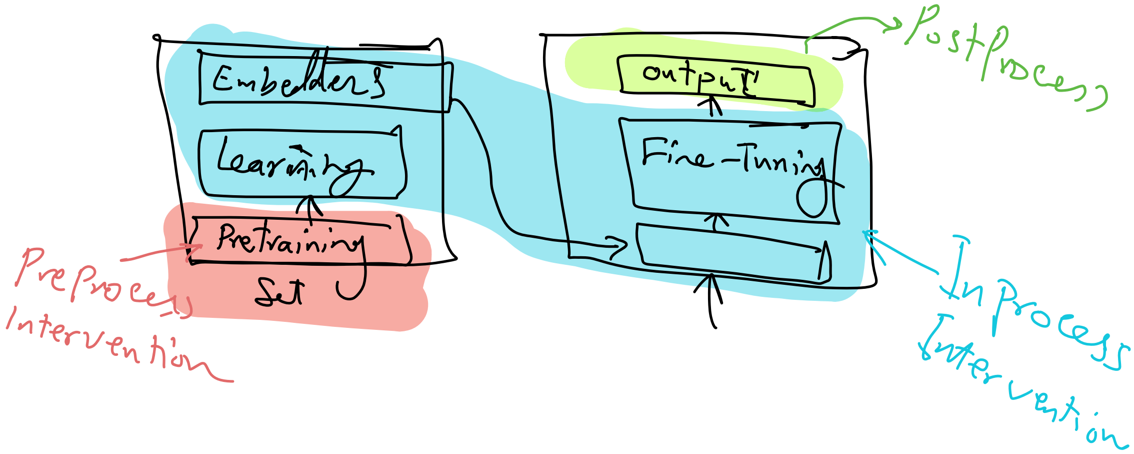
- Take the average as the Expected value of $\vec{g}_i$

$$\vec{g}_i = 1/K \sum_{i=1}^{K} \vec{v}_i$$

$$\text{SEAT} = 1/l \left(\sum_{a \in A} \text{Sim}(a, \vec{g}_1) - \text{Sim}(a, \vec{g}_2) + \sum_{b \in B} \text{Sim}(b, \vec{g}_2) - \text{Sim}(b, \vec{g}_1) \right)$$

Sets of Sentences





Preprocess interventions

- Adding Sentences (CDA: Counterfactual Data Aug.)

  [He] is a doctor

  ↳ [She] is a doctor...

- Rewriting Sentences

  " [Name] follows his dream of being a doctor "

  " " " the " " " " " " "

- Removing Sentences: Removes the Problematic Sentences

Issues:

1) The Size of Pretraining Set is HUGE!

↳ Debiasing the input for Fine Tuning ✗ ✓

2) Relying on Existing Tools is questionable!

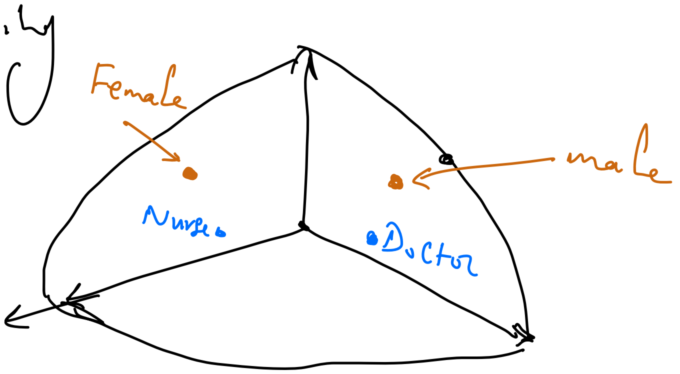
3) Problematic for more than two groups

↳ Significantly change the data!

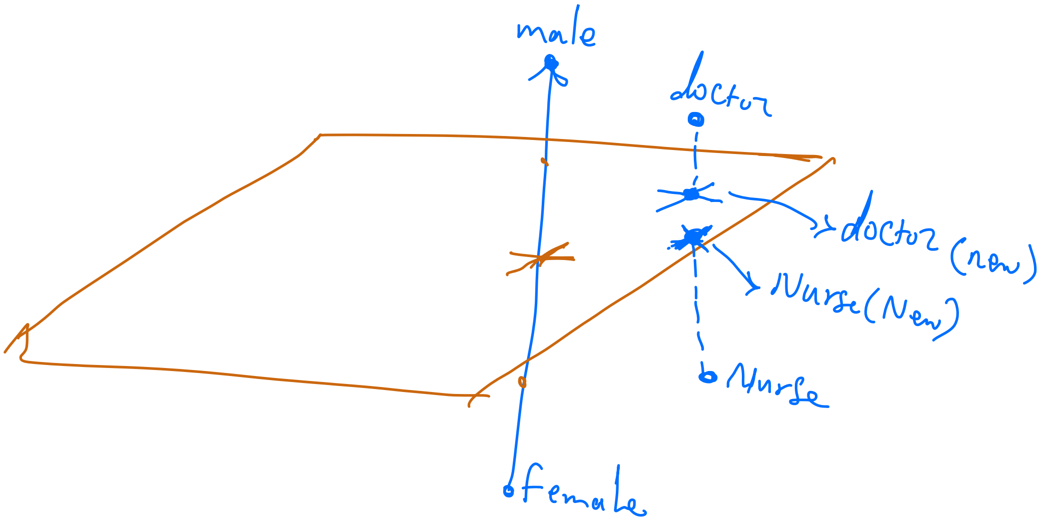
4) For unconsidered groups, the Entire learning Process Should repeat!

Debiasing the Embeddings (Inprocess Interventions)

↳ word Embedding



Goal: minimally change the Embeddings to remove the Bias

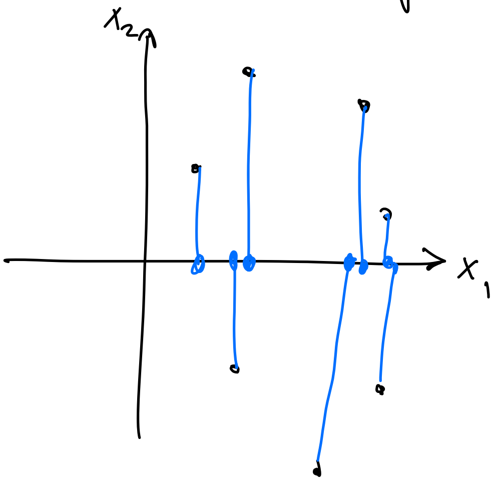


Approach 1: Project ALL the words.

↳ Reducing the dimension by 1

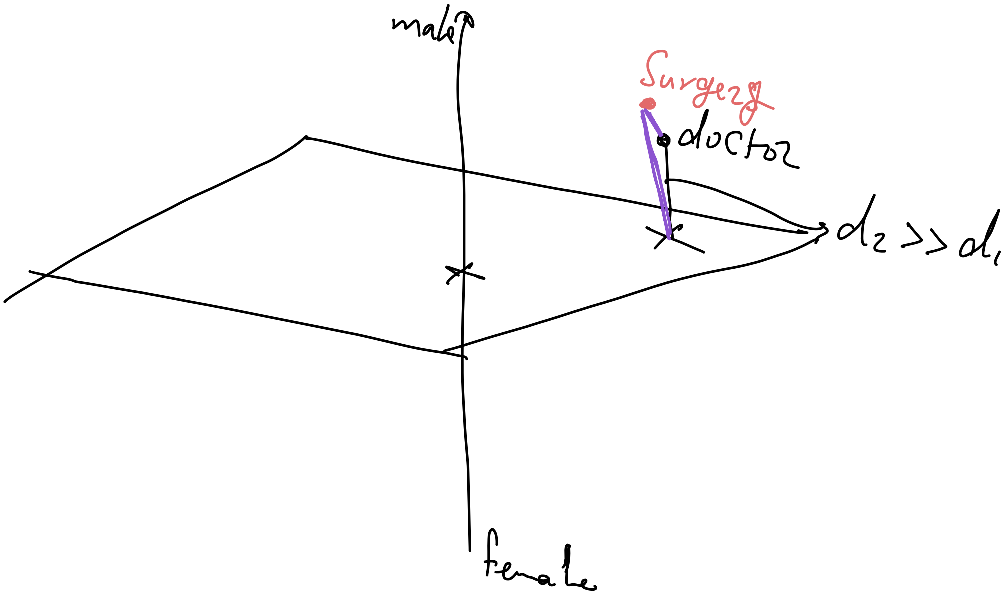
↳ missing some information

We lose useful associations with groups



Approach 2: Only Project the "Problematic" words and don't change the others.

↓ doctor, nurse, ...



- we may mess up semantic similarities with projected words.