Harvard John A. Paulson
School of Engineering
and Applied Sciences

WHERE
SCIENCE
AND
ENGINEERING
CONVERGE

Gender Bias in Text & Image Embeddings

Yuanbiao Wang
yuanbiaowang@fas.harvard.edu

Angel Hsu
angelyinhuahsu@g.harvard.edu

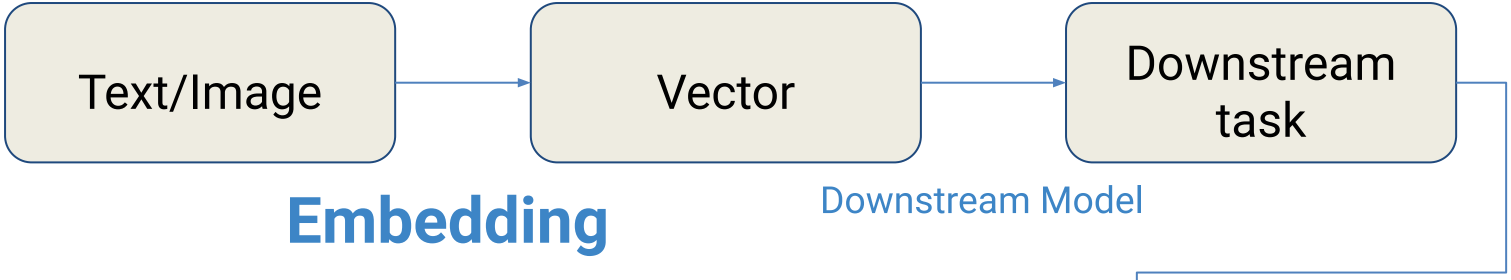
Morris Reeves
morrisreeves@g.harvard.edu

Xinyi Li
xinyi_li@g.harvard.edu

Agenda

- Motivation
- Definition of Bias and Metrics
- Bias in:
  - Training embeddings
  - Existing word embeddings
  - Downstream tasks
  - Image embeddings

Motivation

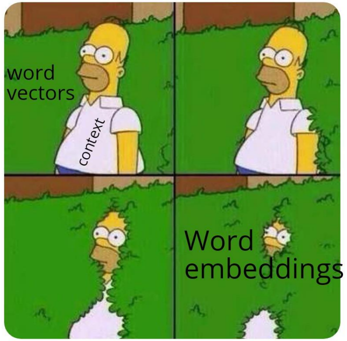


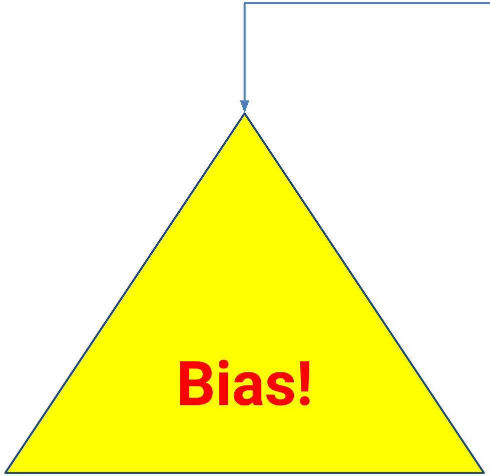
Text/Image → Vector → Downstream task

Embedding

Downstream Model

Word embeddings
in a nutshell





Bias!

How do we measure bias?

1. **Gender bias**
   - male_words = ['he', 'male', 'man', 'father', 'boy', 'husband']
   - female_words = ['she', 'female', 'woman', 'mother', 'girl', 'wife']
2. **Mean cosine similarity with offensive words**
   - If a gender-related word has higher cosine similarity with an offensive word compared to the corresponding word for the opposite gender, then we believe the embedding has a bias against that gender
   - Offensive/Profane Word List from CMU: https://www.cs.cmu.edu/~biglou/resources/

$$\frac{\sum_{i=1}^{n} cos(male\_word, bad\_word_i)}{n} - \frac{\sum_{i=1}^{n} cos(female\_word, bad\_word_i)}{n}$$

How do we measure bias?

3. **Bias in association: the WEAT score** (Caliskan et al. 2016)

- Inputs:
  - ○ 2 sets of target words: X, Y (e.g. {math, science}, {art, literature})
  - ○ 2 sets of attribute words: A, B (e.g. {male, man}, {female, woman})
- Intuitively, are X more similar to A than B, relative to Y?
- Weaknesses:
  - ○ Dependent on choice of target, attribute words

Avg. within-target difference in avg. cosine similarity (between each attribute word and the target word)

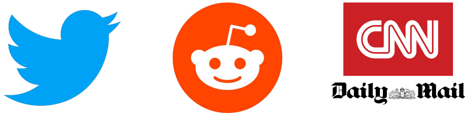
$$\frac{\text{mean}_{x\in X}\, s(x, A, B) - \text{mean}_{y\in Y}\, s(y, A, B)}{\text{std-dev}_{w\in X\cup Y}\, s(w, A, B)}$$

$$s(w, A, B) = \text{mean}_{a\in A}\cos(\vec{w}, \vec{a}) - \text{mean}_{b\in B}\cos(\vec{w}, \vec{b})$$

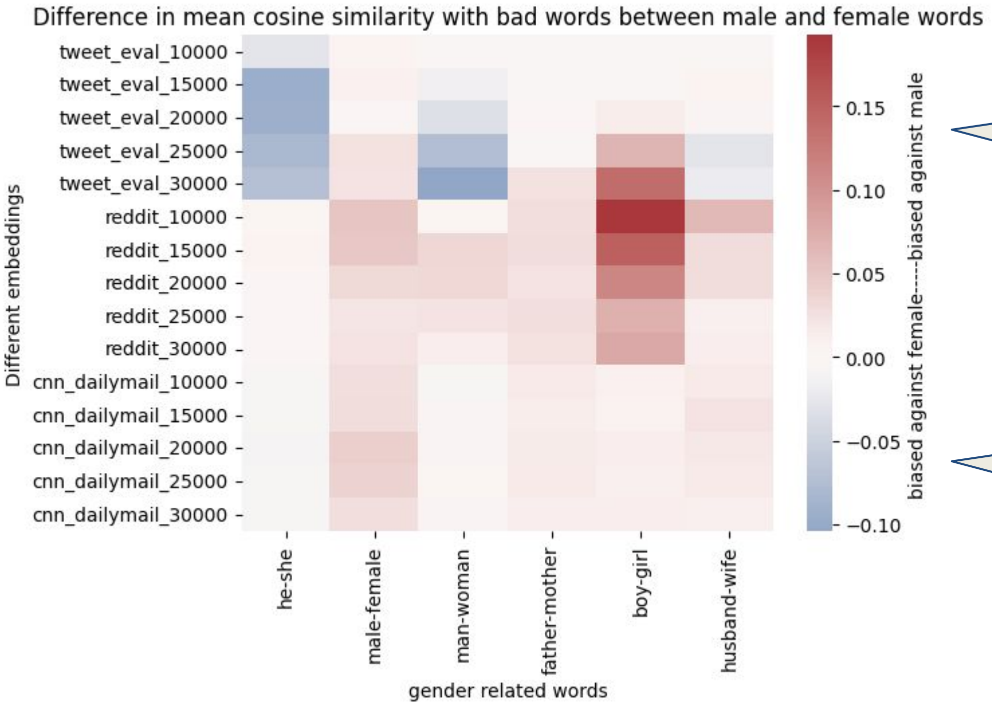
Where does bias come from?

Experimental setup

- **Parameters:**
  - *datasets* ∊ {twitter, reddit, cnn/dailymail}
  - *training size* ∊ {10000, 15000, 20000, 25000, 30000}
  - CBOW context window size = 5
  - Minimum word count = 5
- **Approach:**
  - For each dataset, train word2vec model on sample of size s
  - Compute WEAT score associated with each dataset/size pairing
- **Hypothesis:**
  - More data = more bias
  - Datasource bias ordering:



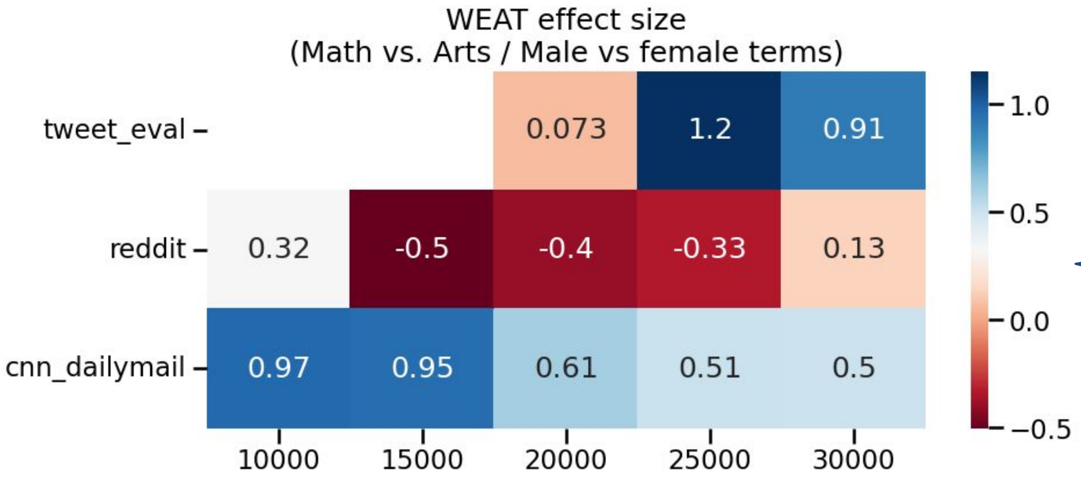
Bias (bad words vs. gender similarity)



Increase in “bias” with training size may reflect better model training (given short tweet text)

No noticeable patterns in difference in mean cos. sim. for CNN / Dailymail

Bias: math vs. arts (by gender)



Math / Male

Arts-male association more prominent in reddit dataset vs. others

Math / Female

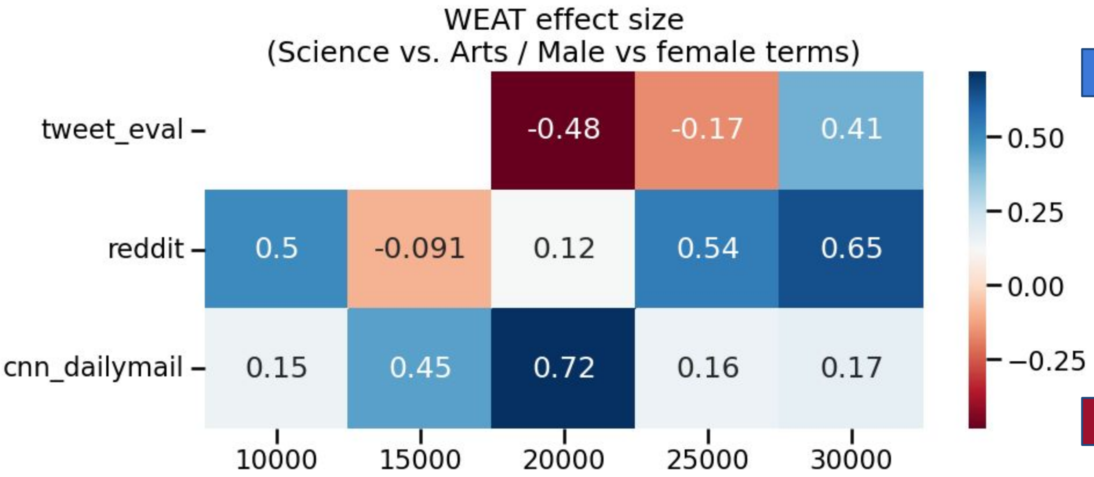
Math: *math, algebra, geometry, calculus, equations, computation, numbers, addition*

Arts: *poetry, art, dance, literature, novel, symphony, drama, sculpture*

Male terms: *male, man, boy, brother, he, him, his, son*

Female terms: *female, woman, girl, sister, she, her, hers, daughter*

Bias: science vs. arts (by gender)



Science/Male

Inconsistent WEAT score patterns

Science/Female

Science: *science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy*

Arts: *poetry, art, Shakespeare, dance, literature, novel, symphony, drama*

Male terms: *brother, father, uncle, grandfather, son, he, his, him*

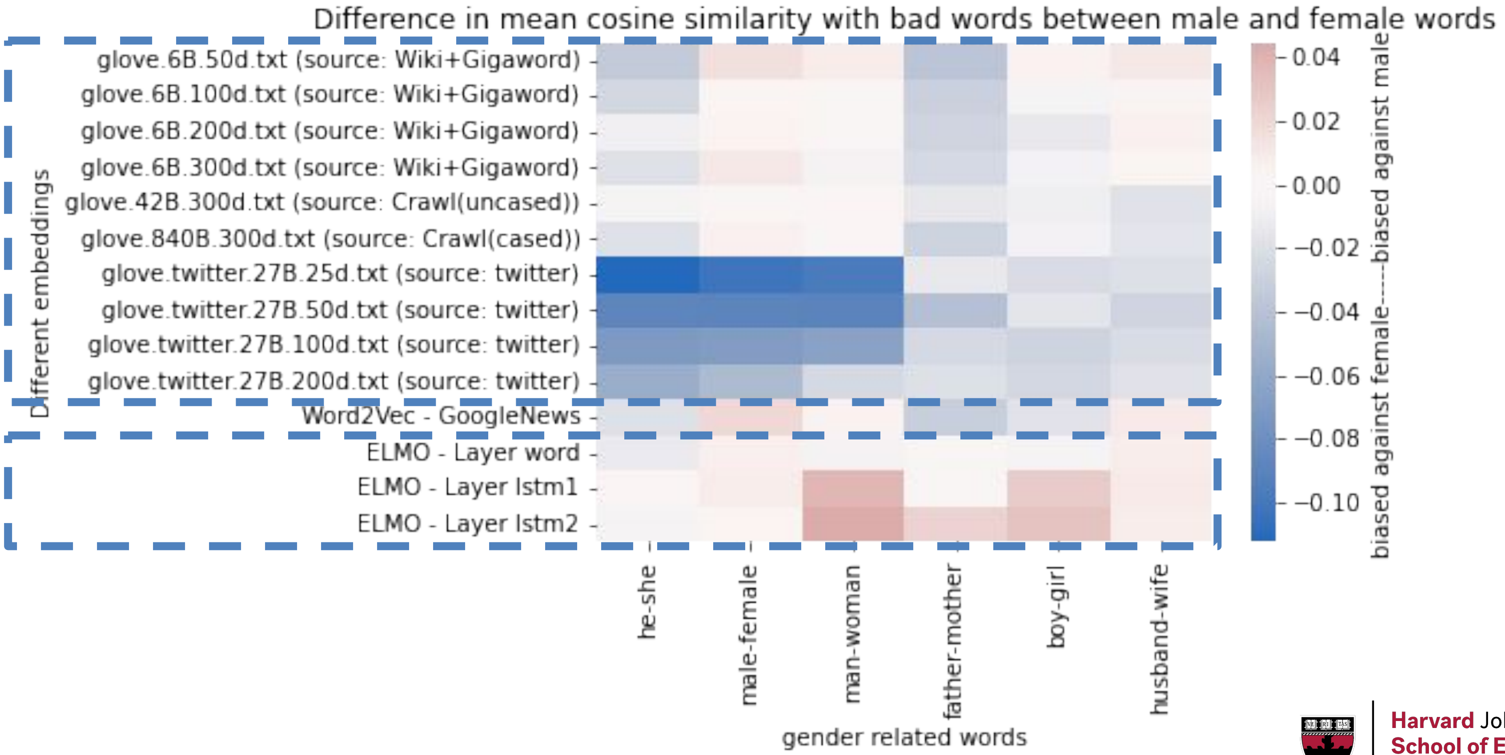
Female terms: *sister, mother, aunt, grandmother, daughter, she, hers, her*

Conclusions and limitations

- **Conclusions:**
  - **Training set source matters** (for embedding bias)
  - No obvious effect of training set size on embedding bias
- **Limitations:**
  - Training parameters (e.g. context size) set globally
  - WEAT results depend on word lists
  - Datasets are curated (e.g. Twitter sentiment, reddit TLDR)
- **Questions raised:**
  - Embedding bias vs. poor embedding / training quality?

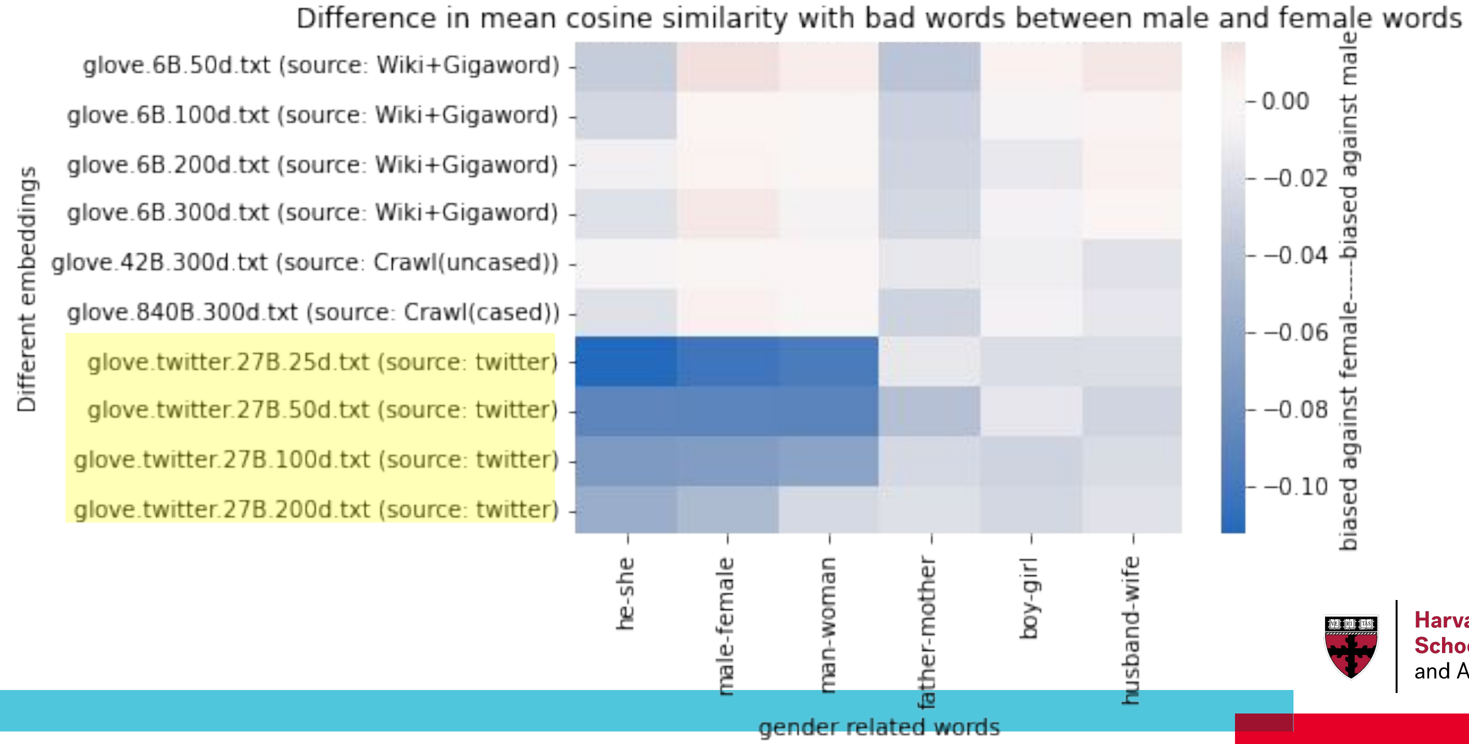
Are existing word embeddings biased?

Are existing embeddings biased? Overview



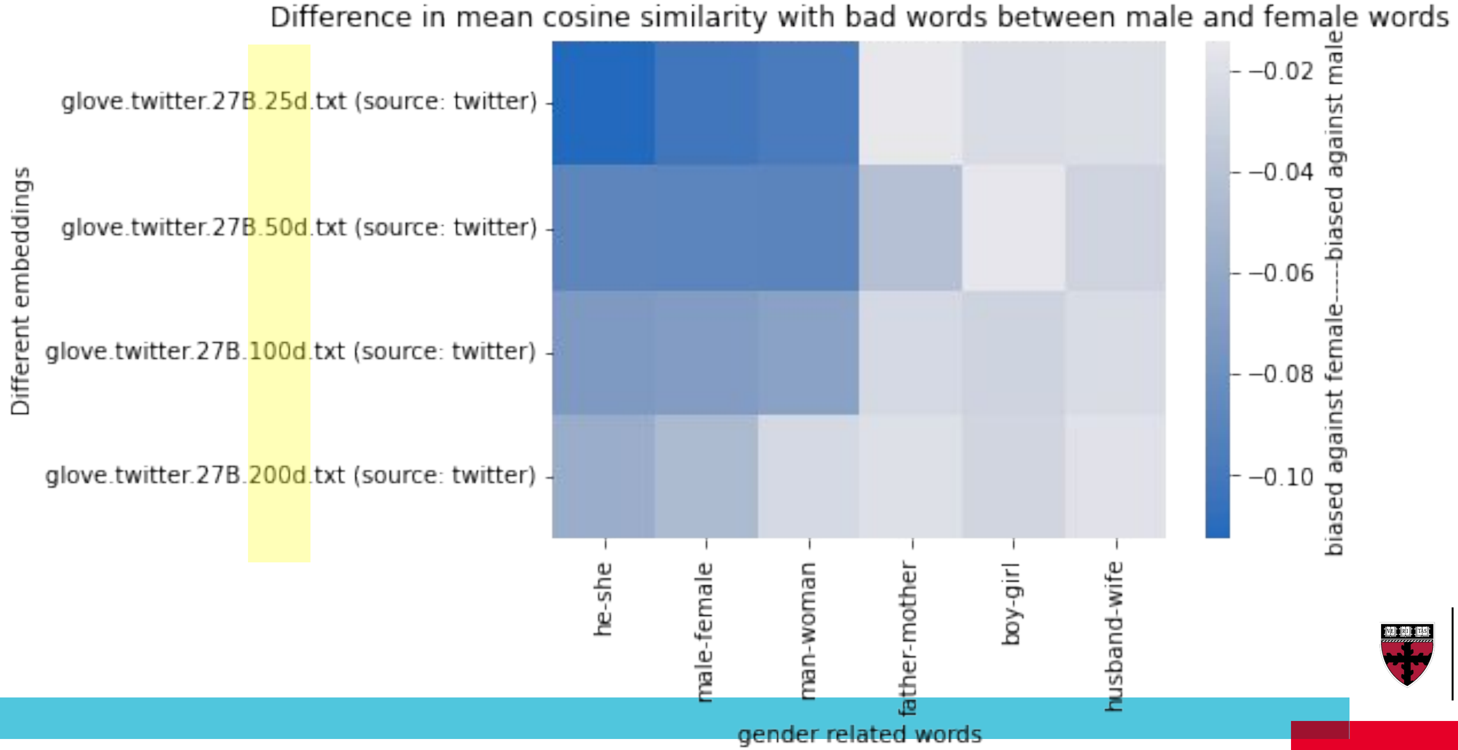
Are existing embeddings biased?

Informal corpus → more bias!



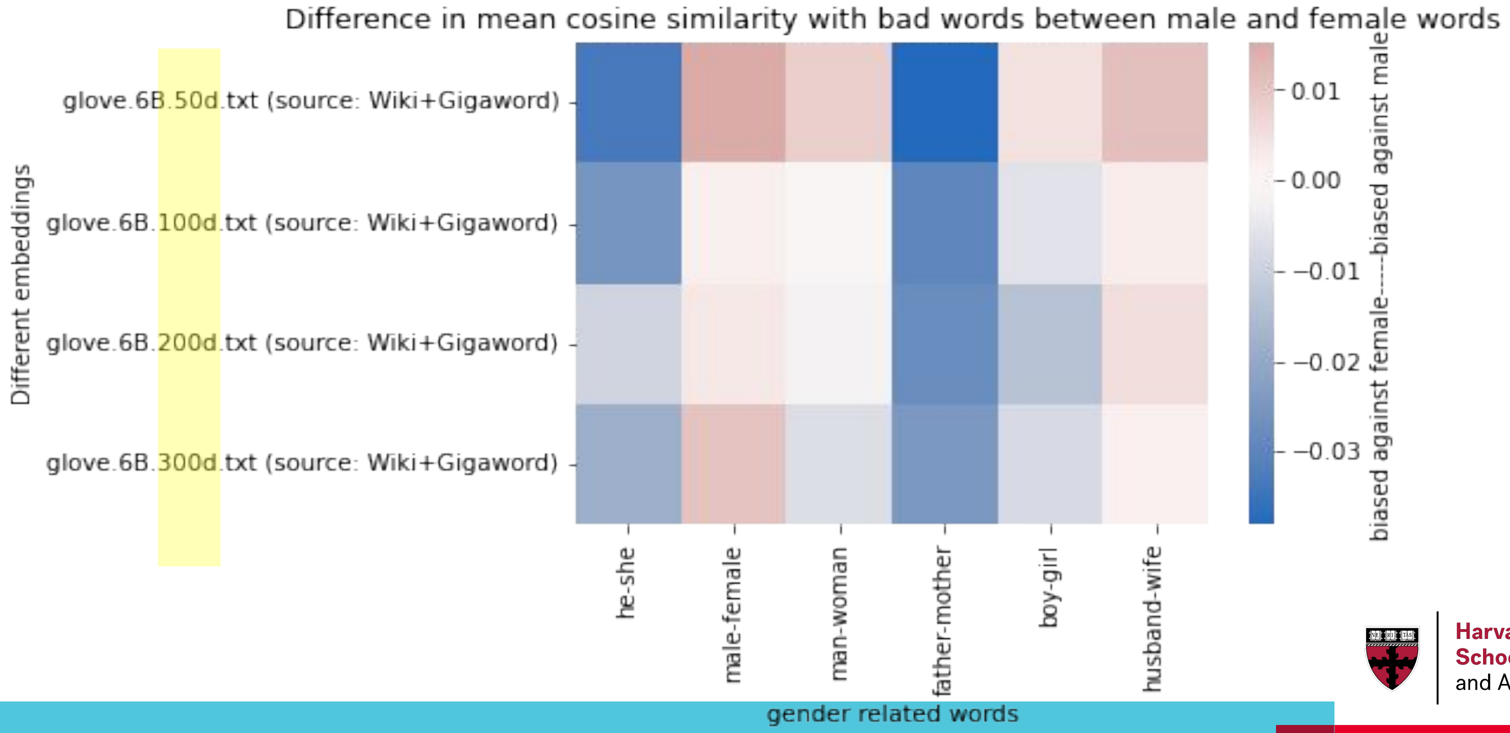
Are existing embeddings biased?

Shorter embedding → more bias!



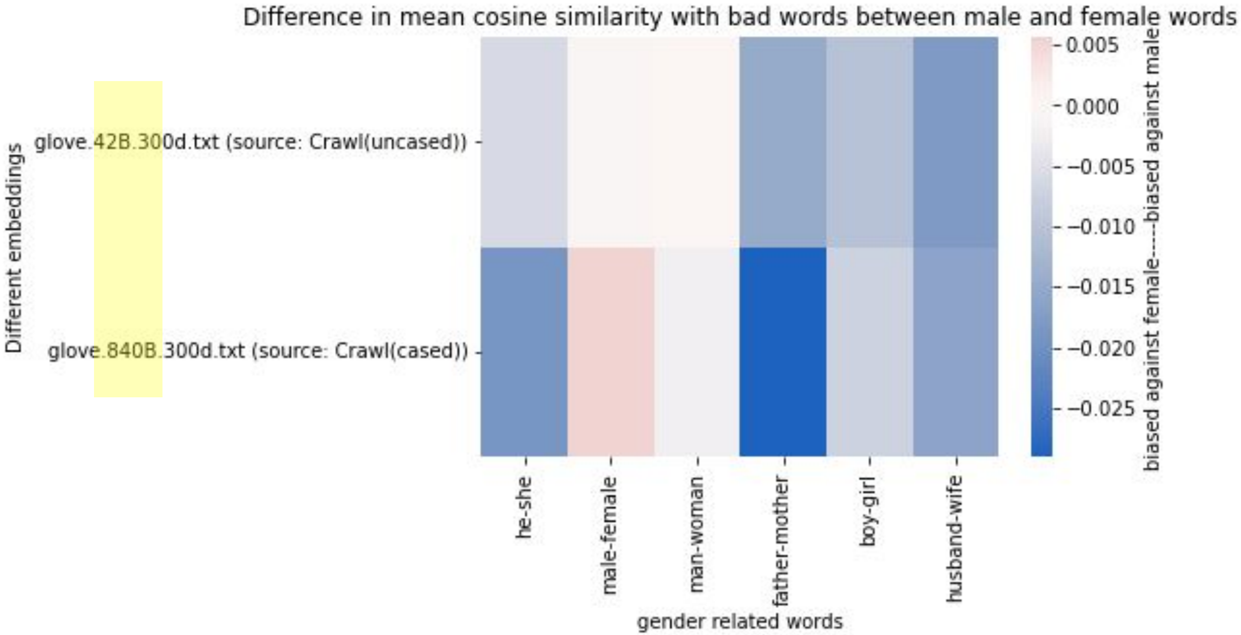
Are existing embeddings biased?

Shorter embedding → more bias!



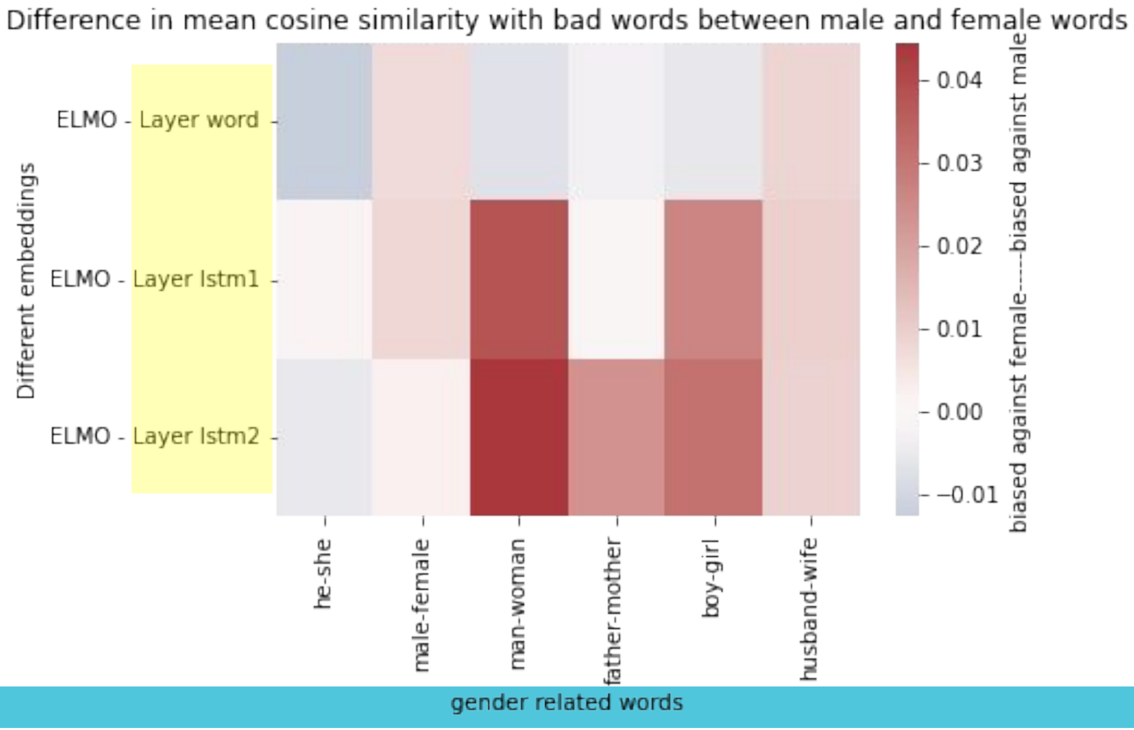
Are existing embeddings biased?

Larger dataset → more bias!



Are existing embeddings biased?

Deeper in ELMo network→ more bias!



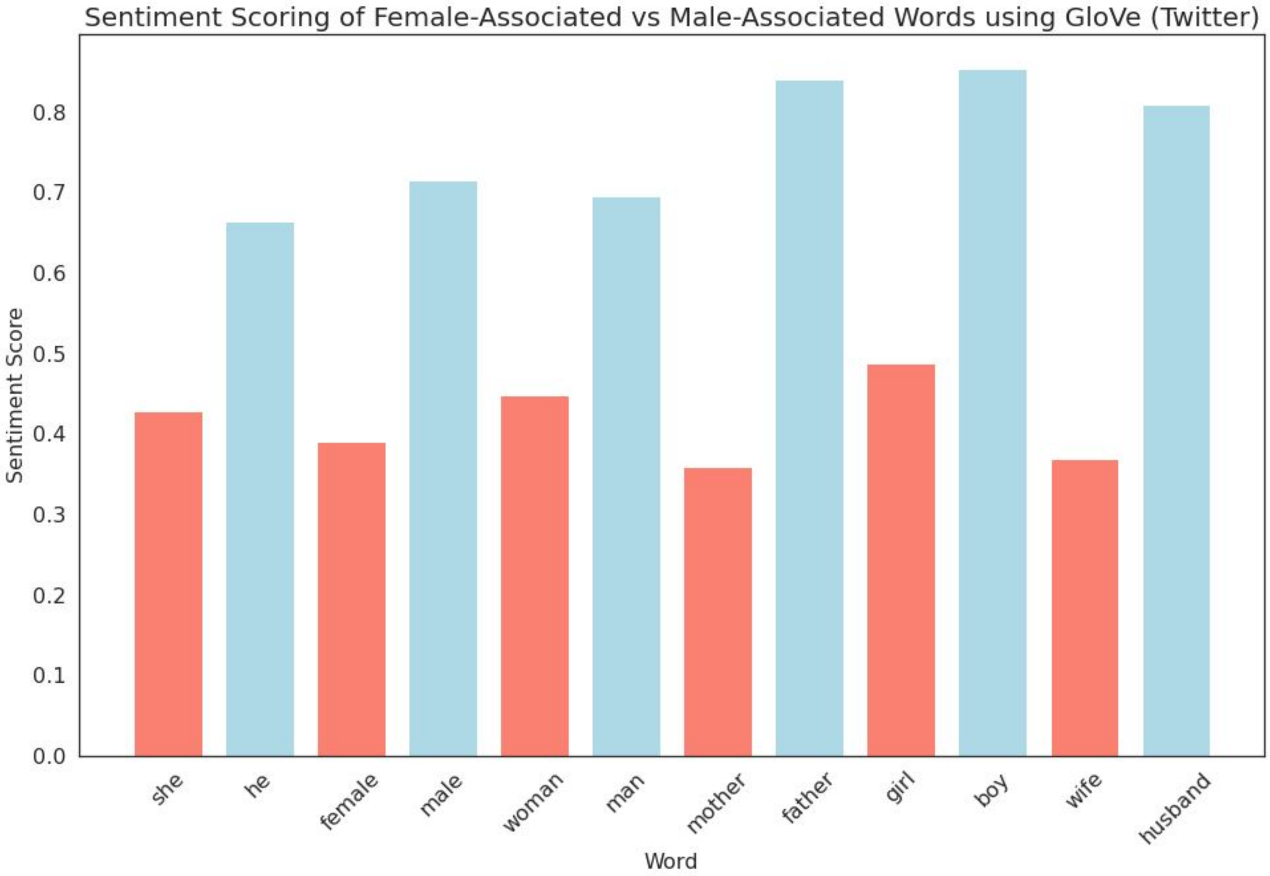
Are existing embeddings biased?

Yes!

They contain the bias in the context!

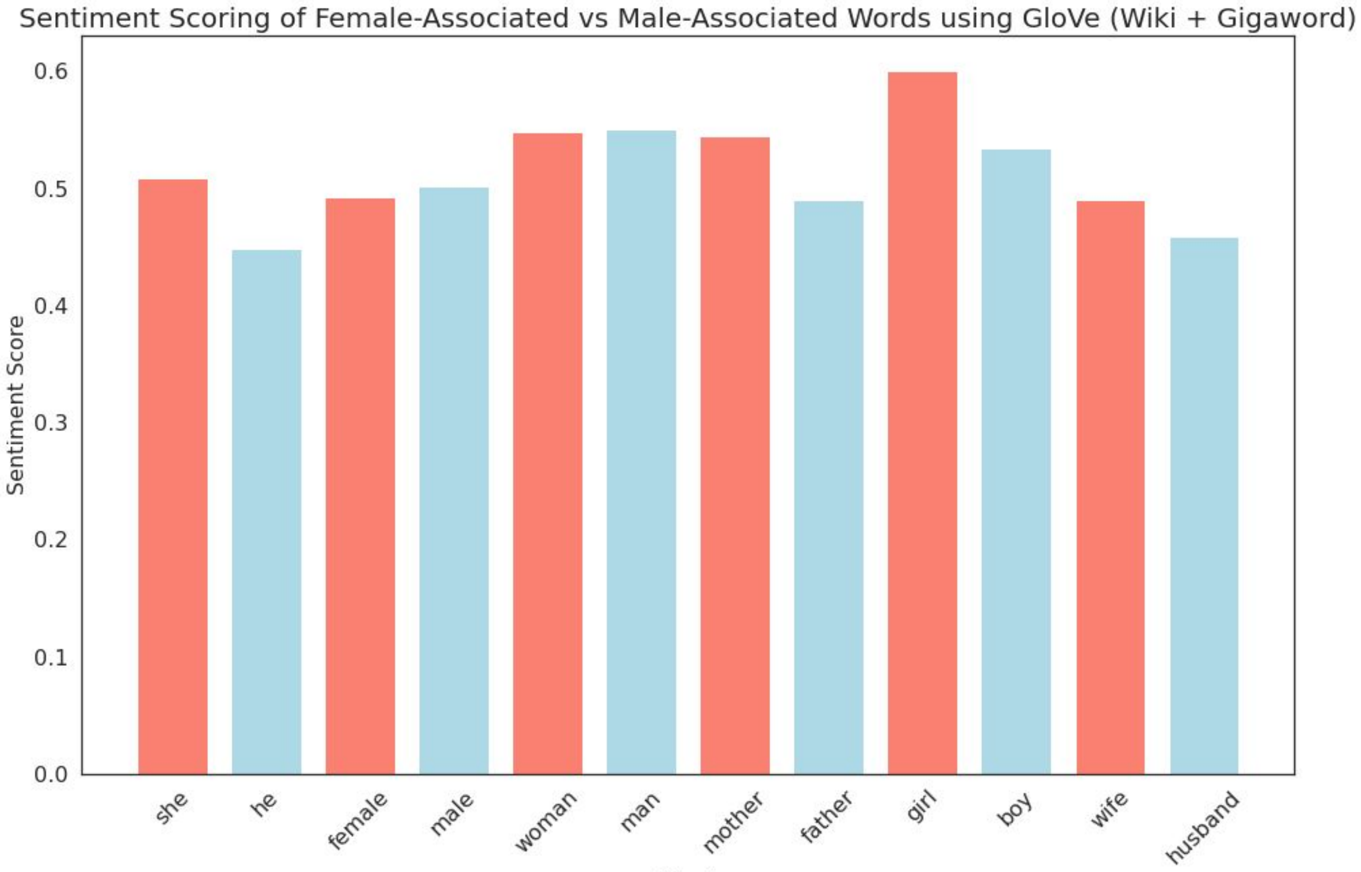
Does bias in embeddings diffuse to downstream tasks?

GloVe (Twitter): Side-by-Side Comparison



Harvard John A. Paulson School of Engineering and Applied Sciences

GloVe (Wikipedia 2014 + Gigaword 5): Side-by-Side Comparison



Harvard John A. Paulson School of Engineering and Applied Sciences

GloVe (Twitter) Sentiment Prediction

Female-Associated Words

Word	Sentiment Score	Sentiment
she	0.43	Negative
female	0.39	Negative
woman	0.45	Negative
mother	0.36	Negative
girl	0.49	Negative
wife	0.37	Negative

Male-Associated Words

Word	Sentiment Score	Sentiment
he	0.66	Positive
male	0.71	Positive
man	0.70	Positive
father	0.84	Positive
boy	0.85	Positive
husband	0.81	Positive

GloVe (Wikipedia 2014 + Gigaword 5) Sentiment Prediction

Female-Associated Words

Word	Sentiment Score	Sentiment
she	0.51	Positive
female	0.49	Negative
woman	0.55	Positive
mother	0.54	Positive
girl	0.60	Positive
wife	0.49	Negative

Male-Associated Words

Word	Sentiment Score	Sentiment
he	0.45	Negative
male	0.50	Positive
man	0.55	Positive
father	0.49	Negative
boy	0.53	Positive
husband	0.46	Negative

Does bias in embeddings diffuse to downstream tasks?

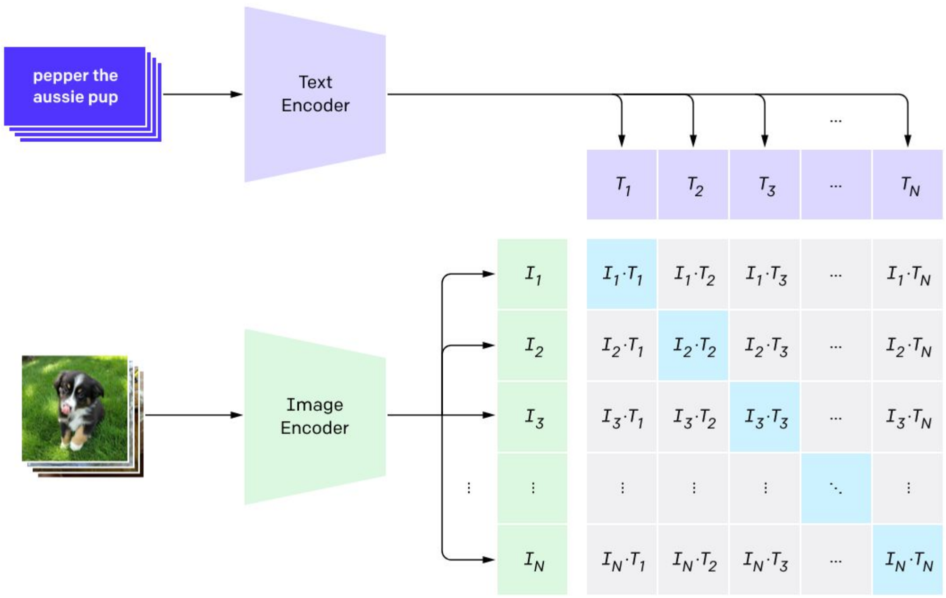
Yes!

Bias in embeddings diffuses to downstream tasks, supporting our conclusions above.

Bias In
Image embeddings

CLIP: Contrastive Language Image Pretraining

1. Contrastive pre-training



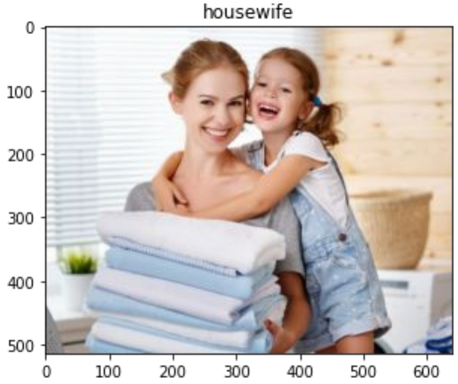
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

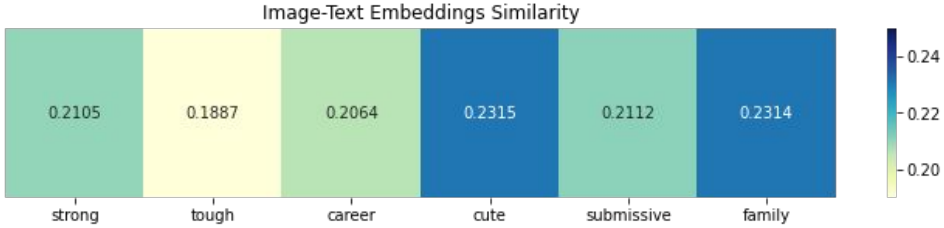
Co-occurent images and text to bring two modalities together

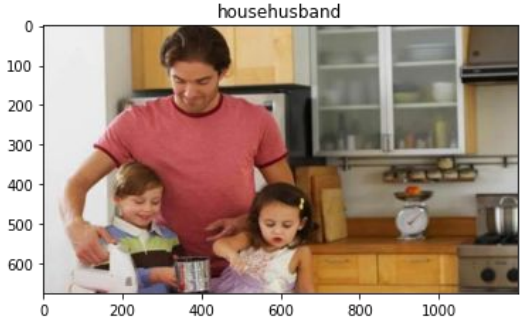
Credits: https://openai.com/blog/clip/

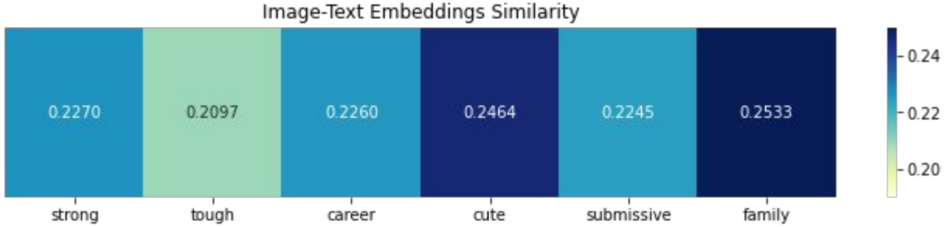
CLIP: Visualization

Demo:
https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQIGwFQmz0?usp=sharing





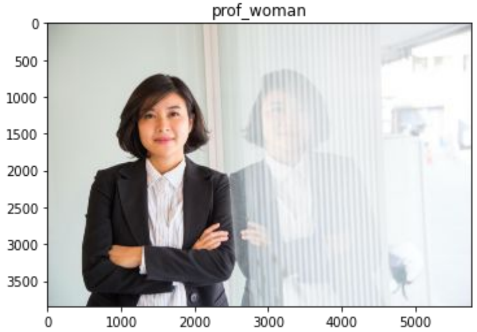




Harvard John A. Paulson School of Engineering and Applied Sciences

CLIP: Visualization

Demo:
https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQIGwFQmz0?usp=sharing


prof_woman

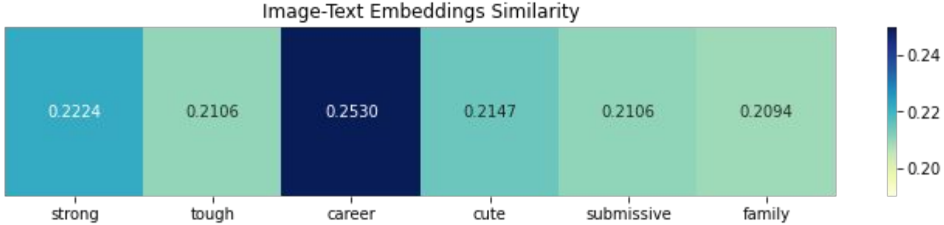


Image-Text Embeddings Similarity

career_man

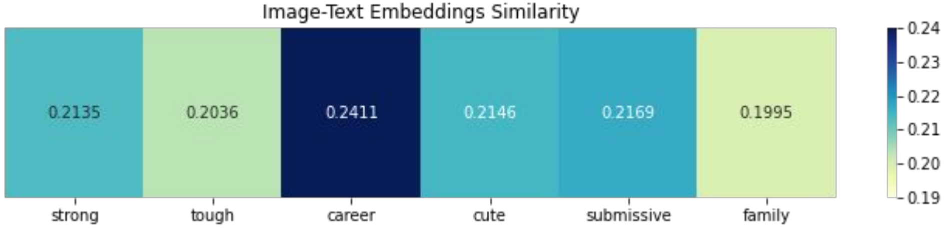


Image-Text Embeddings Similarity

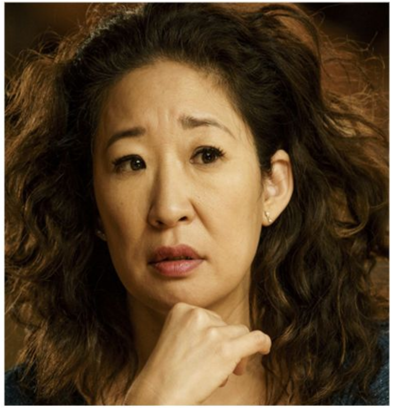
CLIP Visualization: Extensions

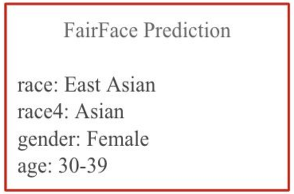
More results about racial, age-related, and stereotype-related disparities in the appendix slides

Demo:
https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQlGwFQmz0?usp=sharing

CLIP: Statistics

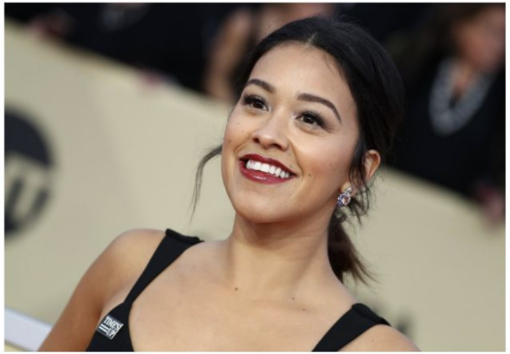
Dataset: FairFace

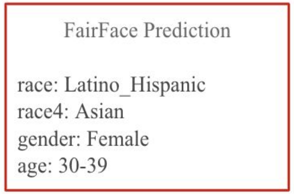




FairFace Prediction

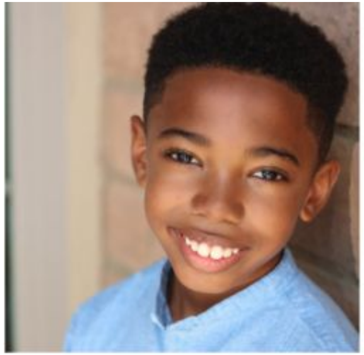
race: East Asian
race4: Asian
gender: Female
age: 30-39

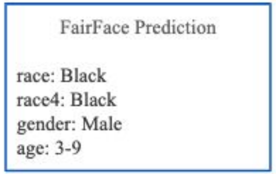




FairFace Prediction

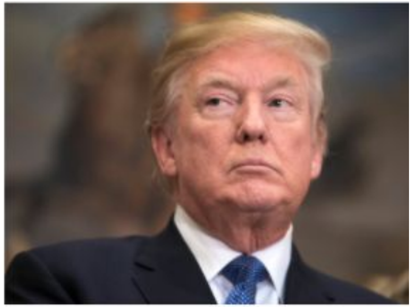
race: Latino_Hispanic
race4: Asian
gender: Female
age: 30-39

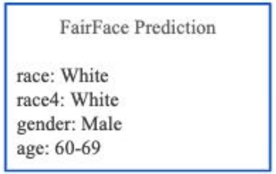




FairFace Prediction

race: Black
race4: Black
gender: Male
age: 3-9

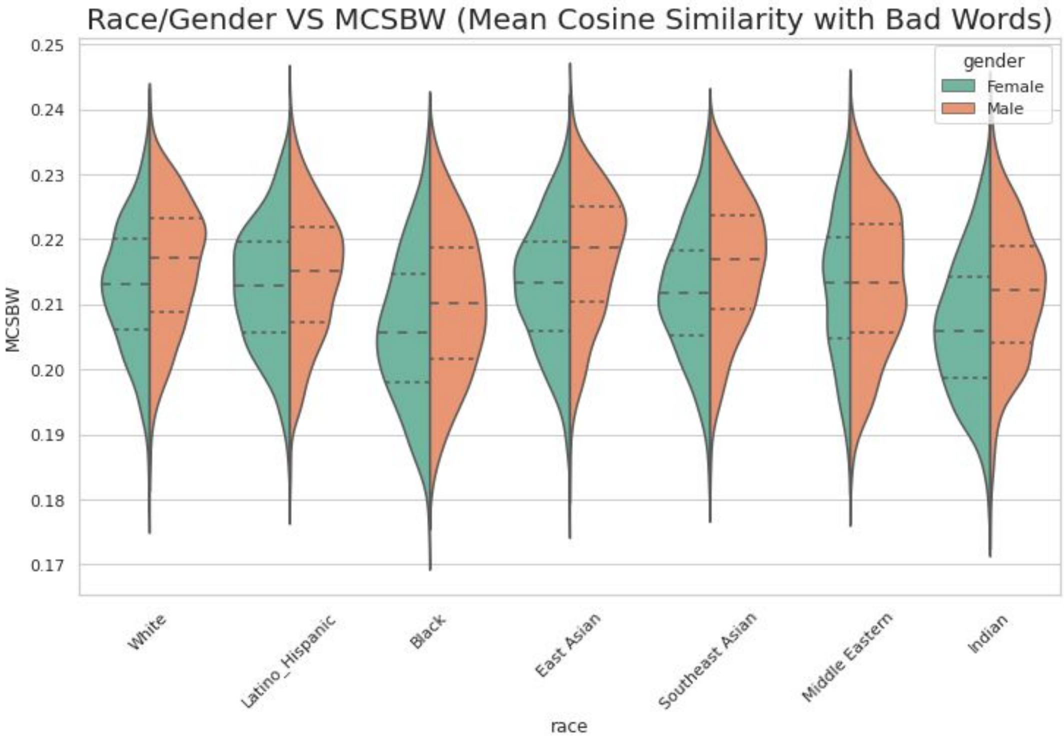




FairFace Prediction

race: White
race4: White
gender: Male
age: 60-69

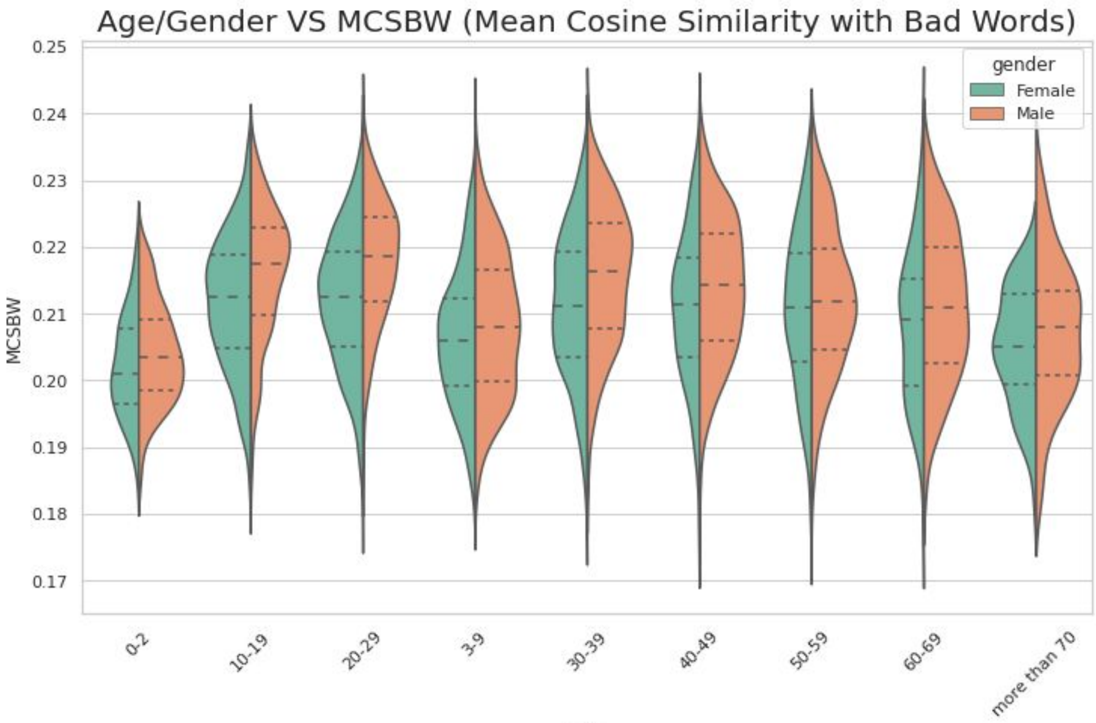
CLIP: Statistics



Males / East Asians are more likely to be associated with negative phrases

Demo: https://colab.research.google.com/drive/16ftNqde0os-jl_sq0UBLaTiPEviON1hW?usp=sharing

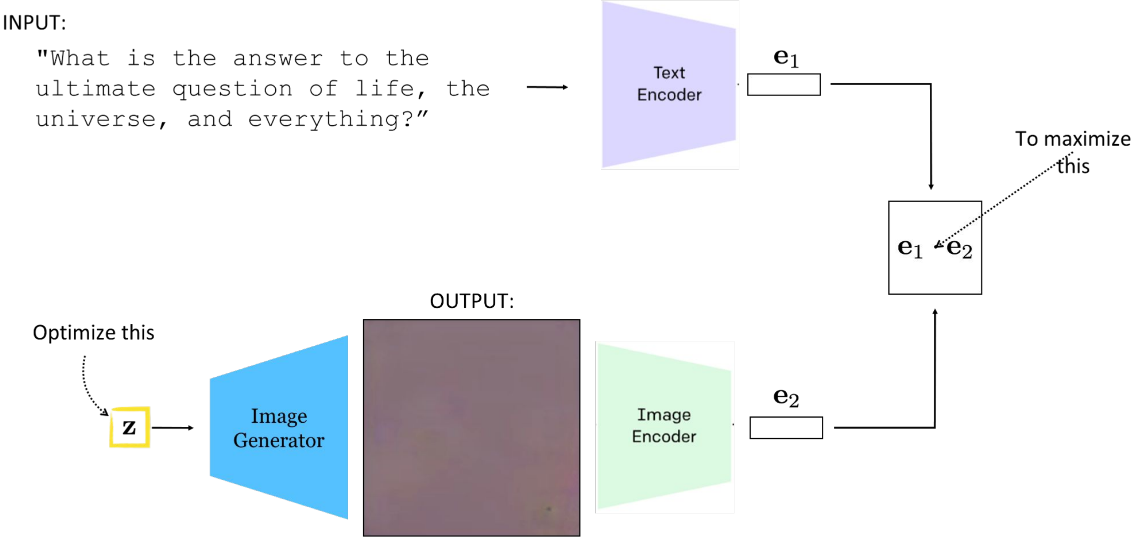
CLIP: Statistics



Teenagers / Middle-aged people are more likely to be associated with bad words

CLIP: Image Generation

New capabilities by plugging pretrained models together: CLIP+GAN



Code: https://colab.research.google.com/drive/1_4PQqzM_0KKytCzWtn-ZPi4cCa5bwK2F?usp=sharing

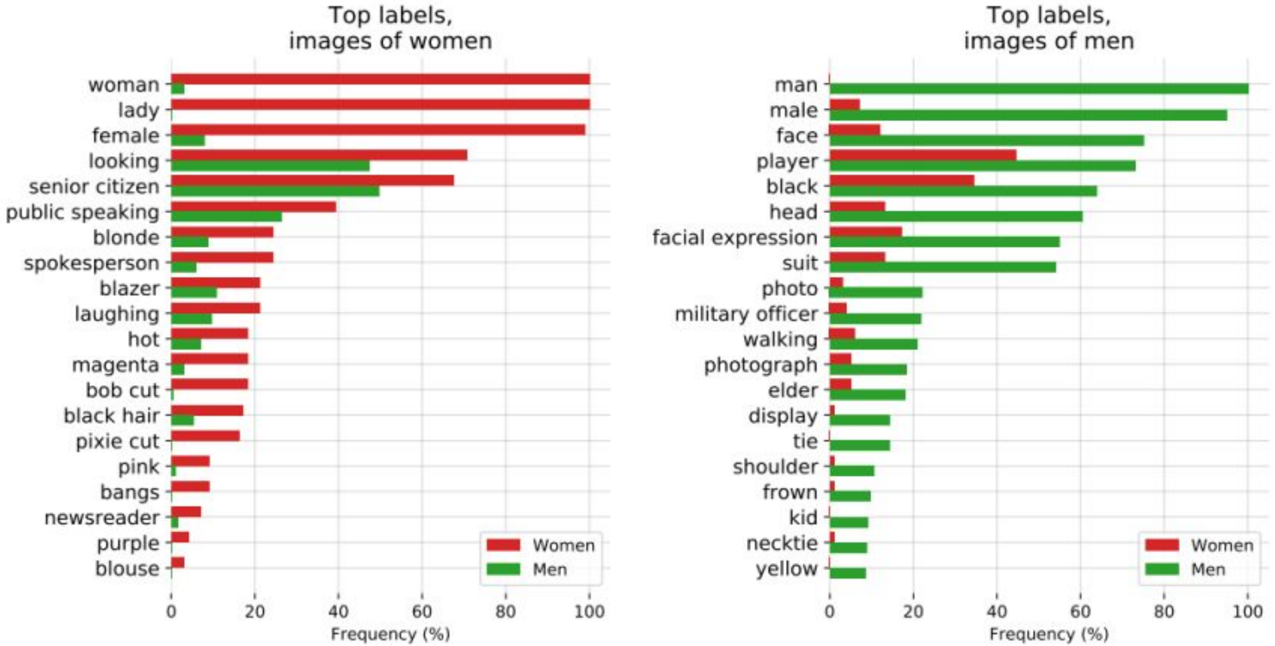
Demo: https://colab.research.google.com/drive/1_4PQqzM_0KKytCzWtn-ZPi4cCa5bwK2F?usp=sharing

Credits: MIT 6.869 Prof. Philip Isola
Source: Katherine Crowson

CLIP: Image Generation

A street gang member is accused of gun violence and grand theft

CLIP: Downstream tasks



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford et. al.

Harvard John A. Paulson School of Engineering and Applied Sciences

Key Takeaways

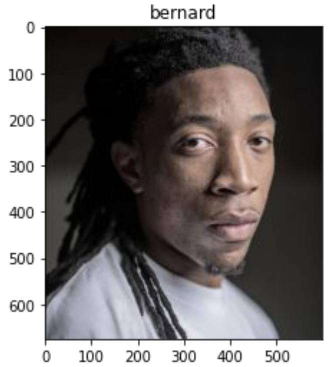
1. **What <u>matters</u>?**
   a. Training data
   b. Bias evaluation metric (WEAT? Bad words?)
   c. Embedding source and embedding size
   d. Downstream tasks (e.g. sentiment prediction and image analysis)
2. **Image embeddings are <u>not immune to bias</u>**
3. **Can leverage the above <u>understanding and awareness of bias</u> in computational models to facilitate <u>bias mitigation</u>, reducing bias embedded in the models we train and develop and <u>enabling more fairness</u> in our world of computation**

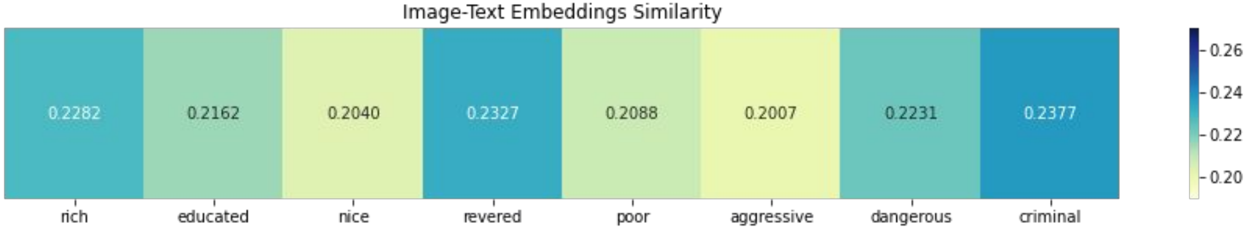
Thank you!

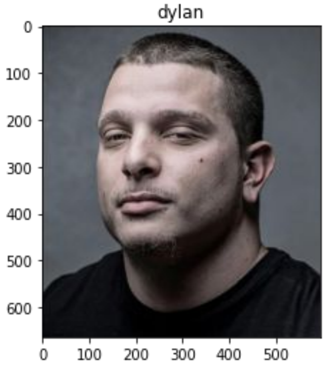
Appendix

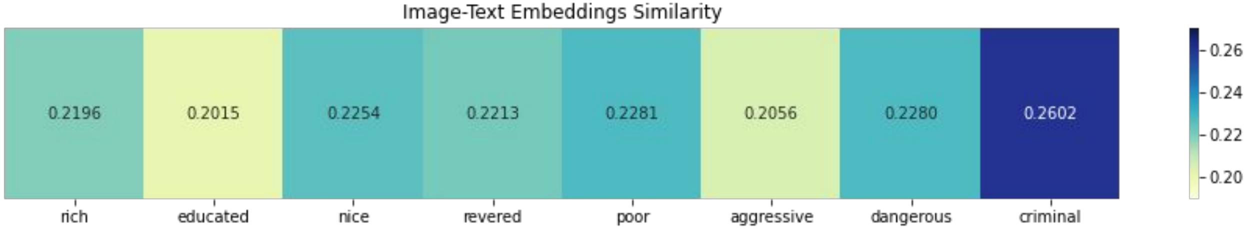
CLIP: Visualization

Demo:
https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQlGwFQmz0?usp=sharing



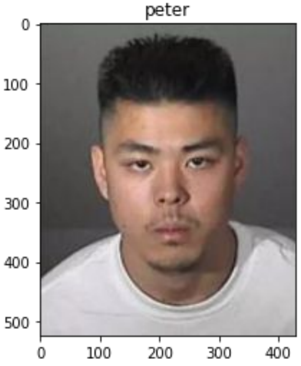


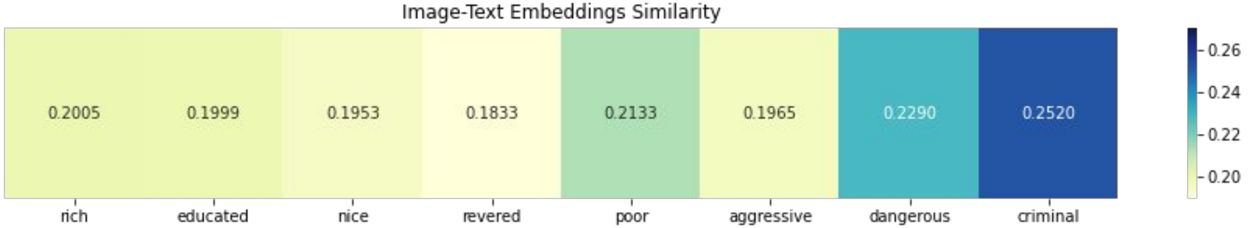


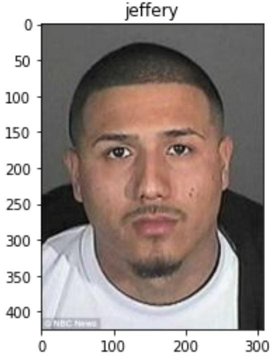


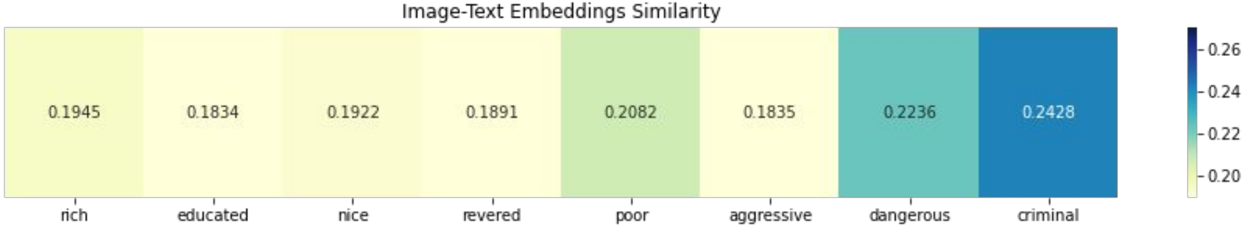
CLIP: Visualization

Demo: https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQIGwFQmz0?usp=sharing



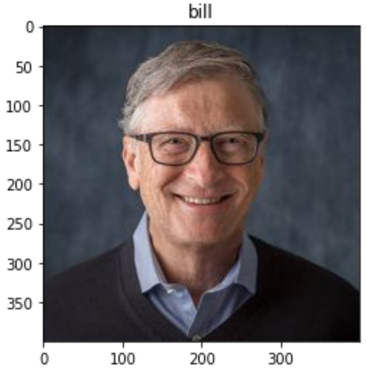


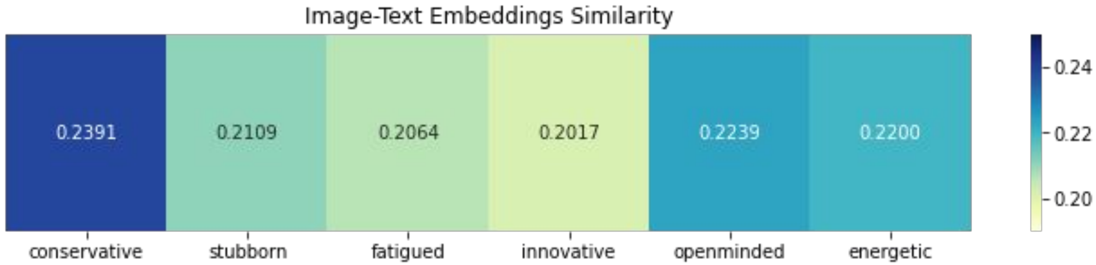


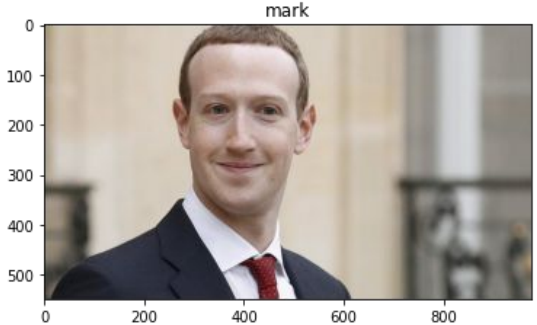


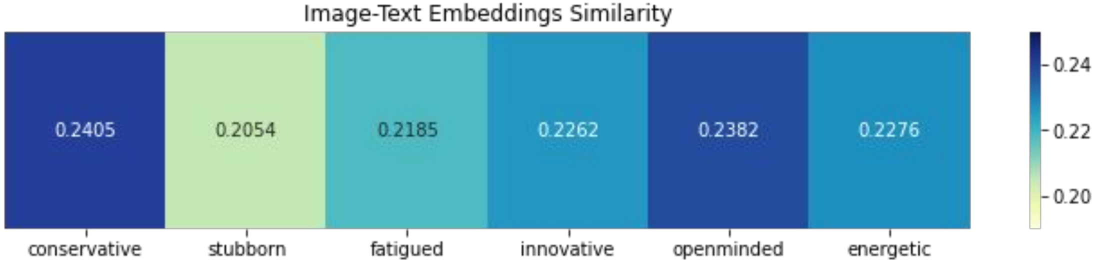
CLIP: Visualization

Demo:
https://colab.research.google.com/drive/12_-2T-jm1NlmaVxQKpLDMIGQIGwFQmz0?usp=sharing



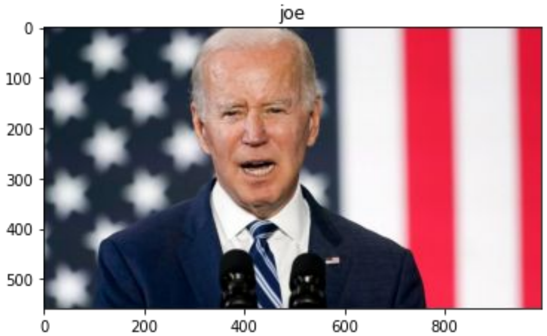


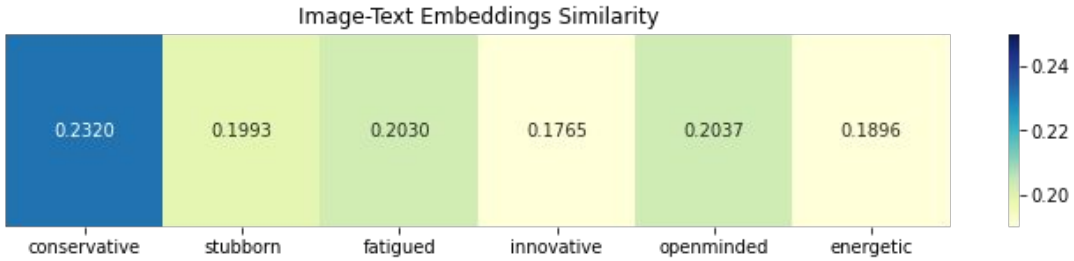


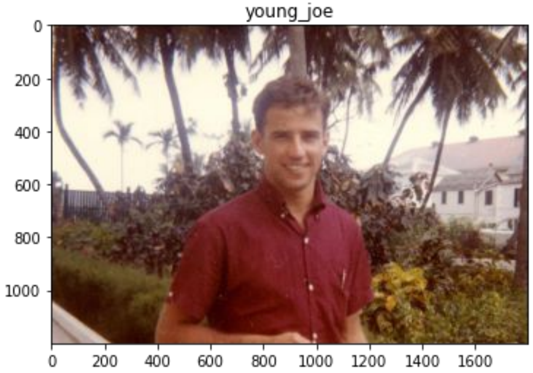


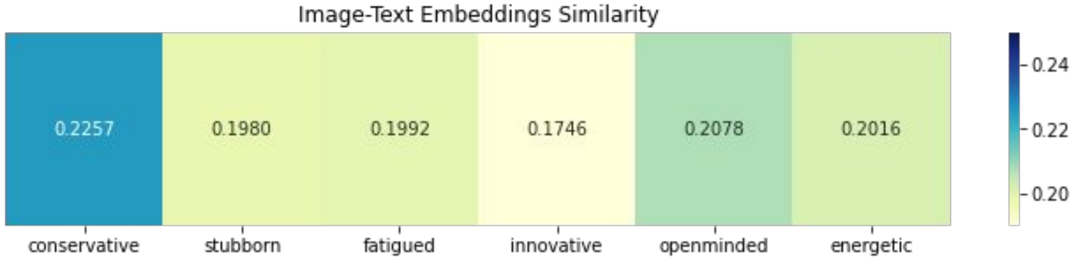
CLIP: Visualization

CONTEXT > PERSONAL CHARACTERISTICS!









CLIP: Image Generation

Who is conservative, stubborn and does not like technology?

CLIP: Image Generation

Harvard University student

CLIP: Image Generation

A beautiful person who does housework