Measuring Social Biases in Grounded Vision and Language Embeddings

Abstract

We generalize the notion of social biases from language embeddings to grounded vision and language embeddings. Biases are present in grounded embeddings, and indeed seem to be equally or more significant than for ungrounded embeddings. This is despite the fact that vision and language can suffer from different biases, which one might hope could attenuate the biases in both. Multiple ways exist to generalize metrics measuring bias in word embeddings to this new setting. We introduce the space of generalizations (Grounded-WEAT and Grounded-SEAT) and demonstrate that three generalizations answer different yet important questions about how biases, language, and vision interact. These metrics are used on a new dataset, the first for grounded bias, created by augmenting extending standard linguistic bias benchmarks with 10,228 images from COCO, Conceptual Captions, and Google Images. Dataset construction is challenging because vision datasets are themselves very biased. The presence of these biases in systems will begin to have real-world consequences as they are deployed, making carefully measuring bias and then mitigating it critical to building a fair society.

1 Introduction

Since the introduction of the Implicit Association Test (IAT) by Greenwald *et al.* [1998], we have had the ability to measure biases in humans related to certain social and cultural concepts, such as race. Caliskan *et al.* [2017] introduce an equivalent of the IAT for word embeddings, which are used throughout natural language processing, called the Word Embedding Association Test (WEAT). The results of testing bias in word embeddings using WEAT parallel those seen when testing humans: both reveal many of the same biases with similar significance. May *et al.* [2019] extend this work with a metric called the Sentence Encoder Association Test (SEAT), that probes biases in embeddings of sentences instead of just words. We take the next step and demonstrate how to test visually grounded embeddings, specifically embeddings from ViLBERT [Lu *et al.*, 2019] and VisualBERT [Li *et al.*, 2019], by extending prior work into what we term Grounded-WEAT and Grounded-SEAT.

Grounded embeddings are used for many consequential tasks in natural language processing, like visual dialog [Murahari *et al.*, 2019] and visual question answering [Hu *et al.*, 2019]. Many real-world tasks such as scanning documents and interpreting images in context employ joint embeddings as the performance gains are significant over using separate embeddings for each modality. It is therefore important to measure the biases of these grounded embeddings. Specifically, we seek to answer three questions:

Do joint embeddings have biases? Since visual biases can be different from those in text, we would expect to see a difference in the biases exhibited by grounded embeddings. Biases may cancel out between the two modalities or they might amplify one another. We find equal or larger biases for grounded embeddings compared to the ungrounded embeddings reported in May *et al.* [2019]. We hypothesize that this may be because visual datasets are much smaller and much less diverse than language datasets.

Can visual evidence that counters a stereotype alleviate biases? The advantage to having multiple modalities is that one modality can demonstrate that a bias is irrelevant to the particular task being carried out. For example, one might provide an image of a woman who is a doctor, and then measure the bias against woman doctors in the embeddings. We find that the bias is largely not impacted, i.e., direct visual evidence against a bias does not help. In general, grounded embeddings seem to be dominated by language, not vision.

To what degree do biases come from language vs. vision in joint embeddings? It may be that joint embeddings derive all of their biases from one modality, such as language. In this case, vision would be relevant to the embeddings, but would not impact the measured bias. We find that this is marginally not the case, and that vision is somewhat relevant to biases, but they are by and large dominated by language. Vision could have a more substantial impact on joint embeddings.

We generalize WEAT and SEAT to grounded embeddings to answer these questions. Several generalizations are possible, three of which correspond to the questions above, while the rest appear unintuitive or redundant. We constructed a dataset from COCO [Chen *et al.*, 2015] and Conceptual Captions [Sharma *et al.*, 2018] but discovered that their images and captions are heavily biased, making finding data for most

C3: EA/AA, (Un)Pleasant	1648	C6: M/W, Career/Family	780	C8: Science/Arts, M/W	718
C11: M/W, (Un)Pleasant	1680	+C12: EA/AA, Career/Family	748	+C13: EA/AA, Science/Arts	522
DB: M/W, Competent	560	DB: M/W, Likeable	480	M/W, Occupation	960
+DB: EA/AA, Competent	440	+DB: EA/AA, Likeable	360	EA/AA, Occupation	928
		Angry Black Woman	760		

(a) Number of images for all bias tests in the dataset collected from Google Images.

C6: M/W, Career/Family	254	M/W, Occupation	229

(b) Number of images for bias tests in the dataset collected from COCO.

C6: M/W, Career/Family	203	M/W, Occupation	171

(c) Number of images for bias tests in the dataset collected from Conceptual Captions.

Table 1: The number of images per bias test in our dataset (EA/AA=European American/African American names; M/W=names of men/women, renamed from M/F to reflect gender rather than sex). Tests prefixed by "C" are from Caliskan *et al.* [2017]; *Angry Black Woman* and "DB" prefixes are from May *et al.* [2019]; prefixes "+C" and "+DB" are from Tan and Celis [2019]. Each class contains an equal number of images per target-attribute pair. The dataset sourced from Google Images is complete, shown in (a). Datasets sourced from COCO and Conceptual Captions, shown in (b) and (c) respectively, contain a subset of the tests because the lack of gender and racial diversity in these datasets makes creating balanced data for grounded bias tests impractical.

existing bias tests nearly impossible. To address this, we created an alternate dataset from Google Images that depicts the targets and attributes required for all bias tests considered.

The dataset introduced along with the metrics presented can serve as a foundation for future work to eliminate biases in joint embeddings. In addition, they can be used as a sanity check before deploying systems to understand what kinds of biases are present. It is unclear what relationship between language biases and visual biases exists in humans, as the IAT has not been used in this way before.

Our contributions are:

1. Grounded-WEAT and Grounded-SEAT answering three questions about biases in grounded embeddings,
2. a new dataset for testing biases in grounded systems,
3. demonstrating that joint embeddings have social biases,
4. showing that vision generally does not mitigate biases, and
5. showing that biases largely come from language, rather than being introduced by vision.

2 Related Work

Models that compute word embeddings are widespread [Mikolov *et al.*, 2013; Devlin *et al.*, 2018; Peters *et al.*, 2018; Radford *et al.*, 2018]. Given their importance, measuring the biases present in such models is critical. Caliskan *et al.* [2017] introduce the Word Embedding Assocation Test, WEAT, based on the Implicit Association Test, IAT, to measure biases in word embeddings. WEAT measures social biases using multiple tests that pair target concepts, e.g., gender, with attributes, e.g., careers and families.

May *et al.* [2019] generalize WEAT to biases in sentence embeddings, introducing the Sentence Encoder Association Test (SEAT). Tan and Celis [2019] generalize SEAT to contextualized word representations, e.g., the encoding of a word in context in the sentence. These advances are incorporated into the grounded metrics developed here, by measuring the bias of word embeddings, sentence embeddings, as well as contextualized word embeddings.

3 Dataset Statistics

We build on the bias tests from Caliskan *et al.* [2017], May *et al.* [2019] and Tan and Celis [2019] by augmenting them with images. Our new dataset contains 10,228 images; see table 1 for a breakdown of the number of images per bias test. To compensate for the lack of diversity in COCO and Conceptual Captions, we collected another version of the dataset where the images are top-ranked hits on Google Images. Results on the existing datasets are still important for the bias tests that can be collected, for two reasons. First, it gives us an indication of where COCO and Conceptual Captions are lacking: the fact that images cannot be collected for all identities in the tests means these datasets are particularly biased in those ways. Second, since COCO and Conceptual Captions form part of the training sets for VisualBERT and ViLBERT respectively, this ensures that biases are not a property of poor out-of-domain generalization. The differences in bias in-domain and out-of-domain appear to be small. Images were collected prior to the implementation of the experiment.

4 Methods

Caliskan *et al.* [2017] base the Word Embedding Assocation Test (WEAT) on an IAT test administered to humans. Two sets of target words, X and Y, and two sets of attribute words, A and B, are used to probe the system. The average cosine similarity between pairs of word embeddings is used as the basis of an indicator of bias, as in:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b) \quad (1)$$

where s measures how close on average the embedding for word w is compared to the words in attribute A and attribute B. Being systematically closer to A as opposed to B, or vice versa, is an indication that the concepts are more closely related. Such relative distances between word vectors are used in many tasks, e.g., analogy completion [Drozd *et al.*, 2016].

By incorporating both target word classes X and Y, this distance can be used to measure bias. The space of embeddings has structure that may make all targets, e.g., both men's names

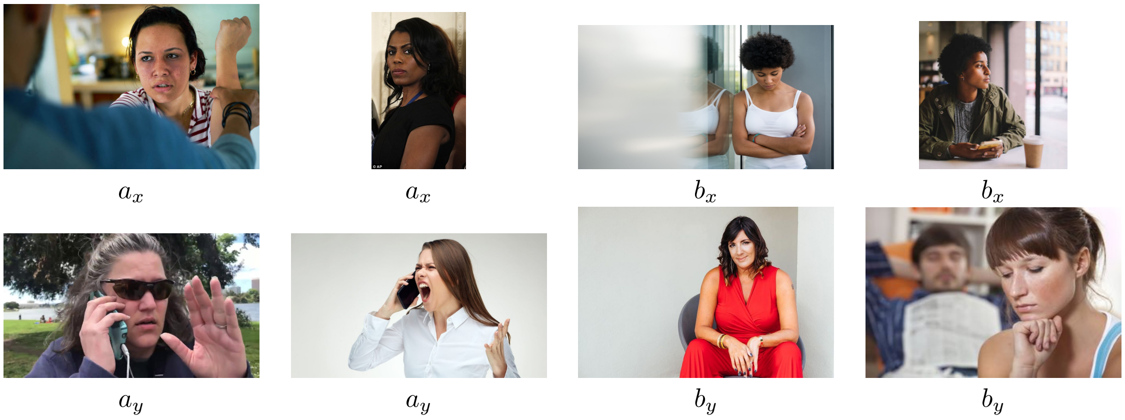

Figure 1: One example set of images for the bias class *Angry black women stereotype* [Collins, 2004], where the targets, X and Y, are typical names of *black women* and *white women*, and the linguistic attributes are *angry* or *relaxed*. The top row depicts black women; the bottom row depicts white women. The two left columns depict aggressive stances while the two right columns depict more passive stances. The attributes for the grounded experiment, A_x, B_x, A_y, and B_y, are images that depict a target and in the context of an attribute.

Embedding #	Word
1	Man
2	Woman
3	Lawyer
4	Teacher

(a) Possible embeddings for an ungrounded model

Embedding #	Word	What the image shows
1	Man	*Any Man*
2	Man	*Any Woman*
3	Woman	*Any Man*
4	Woman	*Any Woman*
5	Lawyer	*Man Lawyer*
6	Lawyer	*Man Teacher*
7	Lawyer	*Woman Lawyer*
8	Lawyer	*Woman Teacher*
9	Teacher	*Man Lawyer*
10	Teacher	*Man Teacher*
11	Teacher	*Woman Lawyer*
12	Teacher	*Woman Teacher*

(b) Possible embeddings for a visually grounded model

Table 2: The content of a trivial hypothetical grounded dataset to demonstrate the intuition behind the three experiments. The dataset could be used to answer questions about biases in association between gender and occupation. Each entry is an embedding that can be computed with an ungrounded model, (a), and with a grounded model, (b), for this hypothetical dataset. This demonstrates the additional degrees of freedom when evaluating bias in grounded datasets. In the subsections that correspond to each of the experiments, sections 4.1 to 4.3, we explain which parts of this hypothetical dataset are used in each experiment. Our experiments only use a subset of the possible embeddings, leaving room for new metrics that answer other questions.

and women's names, closer to one profession than the other. Bias is defined as one of the two targets being significantly closer to one set of attribute words compared to the other. The test in eq. (1) is computed for each set of targets, determining their relative distance to the attributes. The difference between the target distances reveals which target sets are more associated with which attribute sets:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \qquad (2)$$

The effect size, i.e., the number of standard deviations in which the peaks of the distributions of embedding distances differ, of this metric is computed as

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)} \qquad (3)$$

May *et al.* [2019] extend this test to measure sentence embeddings, by using sentences in the target and attribute sets. Tan and Celis [2019] extend the test to measure contextual effects, by extracting the embedding of single target and attribute tokens in the context of a sentence rather than the encoding of the entire sentence. We demonstrate how to extend these notions to a grounded setting, which naturally adapts these two extensions to the data, but requires new metrics because vision adds new degrees of freedom to what we can measure.

To explain the intuition behind why multiple grounded tests are possible, consider a trivial hypothetical dataset that measures only a single property; see table 2. This dataset is complete: it contains the cross product of every target category, i.e., gender, and attribute category, i.e., occupation, that can happen in its minimal world. In the ungrounded setting, only 4 embeddings can be computed because the attributes are independent of the target category. In the grounded setting, by definition, the attributes are words and images that correspond to one of the target categories. This leads to 12 possible grounded embeddings[1]; see table 2. We subdivide the attributes A and B into two categories, A_x and B_x, which depict the attributes

[1] An alternative way to construct such a dataset might have ambi-

with the category of target x, and A_y and B_y, with the category of target y. Example images for bias test related to a racial and gender stereotype that anger is more prevalent in black women, are shown in fig. 1. These images depict the target's category and attributes; they are the equivalent of the attributes in the ungrounded experiments.

With these additional degrees of freedom, we can formulate many different grounded tests in the spirit of eq. (2). We find that three such tests, described next, have intuitive explanations and measure different but complementary aspects of bias in word embeddings. These questions are relevant to both bias and to the quality of word embeddings. For example, attempting to measure the impact of vision separately from language on joint embeddings can indicate if there is an over-reliance on one modality over another.

4.1 Experiment 1: Do joint embeddings contain biases?

This experiment measures biases by integrating out vision and looking at the resulting associations. For example, regardless of what the visual input is, are men deemed more likely to be in some professions compared to women? Similarly to eq. (2), we compute the association between target concepts and attributes, except that we include all of the images:

$$s(X,Y,A,B) = \sum_{x \in X} s(x, A_x \cup A_y, B_x \cup B_y) - \sum_{y \in Y} s(y, A_x \cup A_y, B_x \cup B_y)$$

To be concrete, for the trivial hypothetical dataset in table 2, this corresponds to $S(1,\{5,7\},\{10,12\}) - S(4,\{5,7\},\{10,12\})$, which compares the bias relative to *man* and *woman* against *lawyer* or *teacher* across all target images. If no bias is present, we would expect the effect size to be zero. Our hope would be that the presence of vision at training time would help alleviate biases even if at test time any images are possible.

4.2 Experiment 2: Can visual evidence that counters a stereotype alleviate biases?

An advantage of grounded over ungrounded embeddings is that we can show scenarios that clearly counter social stereotypes. For example, the model may think that men are more likely to have some professions, but are the embeddings different when visual input to the contrary is provided? Similarly to eq. (3), we compute the association between target concept and attributes, except that we include only images that correspond to the target concept's category:

$$s(X,Y,A,B) = \sum_{x \in X} s(x, A_x, B_x) - \sum_{y \in Y} s(y, A_y, B_y)$$

To be concrete, for the trivial hypothetical dataset in table 2, this corresponds to $S(1,\{5\},\{10\}) - S(4,\{7\},\{12\})$, which computes the bias of *man* and *woman* against *lawyer* and *teacher* relative to only images that actually depict lawyers and teachers who are men when comparing to target *man* and lawyers and teachers who are women when comparing to target *woman*. If no bias was present, we would expect the effect size to be zero. Our hope would be that even if biases exist, clear visual evidence to the contrary would overcome them.

4.3 Experiment 3: To what degree do biases come from language vs. vision in joint embeddings?

Even if biases exist, one might wonder if vision is relevant to them? Perhaps all of the biases come from language and vision only plays a small auxiliary role. To probe this, we use images that both support and counter the stereotype. In other words, if the model does not change its biases regardless of the images being shown, then vision does not play a role in encoding biases. Note that we are not saying that the embeddings do not consider vision, but merely that it may or may not have an effect on biases specifically. Similarly to eq. (3), we compute the association between target concepts and attributes, except that we compare cases when images support stereotypes to cases where images counter stereotypes and do not depict the target concept:

$$s(X,Y,A,B) = \frac{1}{2}\Big(\big|\sum_{x \in X} s(x, A_x, B_x) - \sum_{x \in X} s(x, A_y, B_y)\big| + \big|\sum_{y \in Y} s(y, A_y, B_y) - \sum_{y \in Y} s(y, A_x, B_x)\big|\Big)$$

To be concrete, for the trivial hypothetical dataset in table 2, this corresponds to $\frac{1}{2}(|S(1,\{5\},\{10\}) - S(1,\{7\},\{12\})| + |S(2,\{7\},\{12\}) - S(2,\{5\},\{10\})|)$, which compares the bias relative to *man* against *lawyer* or *teacher* and *woman* against *lawyer* or *teacher* relative to images that are either evidence for these occupations as men and women. We take the absolute value of the two, since they may be biased in different ways. If no bias was present, we would expect the effect size to be zero. This provides a finer-grained metric for the relevance of vision to embeddings.

5 Results

We evaluate VisualBERT [Li *et al.*, 2019] on images from COCO, ViLBERT [Lu *et al.*, 2019] on images Conceptual Captions, and both models on images we collected from Google Images. Images features are computed in the same manner as in the original publications for both VisualBERT and ViLBERT. We compute p-values using the updated permutation test described in May *et al.* [2019] In each case, we evaluate the task-agnostic base model without task-specific fine tuning. The effect of task-specific training on biases is an interesting open question for future work.

Overall, the results are consistent with prior work on biases in both humans and models such as BERT. The experiments were run on VisualBERT COCO, VisualBERT Google Images, ViLBERT Conceptual Captions and ViLBERT Google Images.

guity in the attributes about what the target is, more closely mirroring the language setting. This would require images that simultaneously depict both targets, e.g., both a man and woman who are teachers, so as not to provide evidence for either case. Finding such data is difficult and may be impossible in many cases, but it would also be a less realistic measure of bias. In practice, systems built on top of grounded embeddings will not be used with balanced images, and so while in a sense more elegant, this construction may completely misstate the biases one would see in the real world.

Gender	Level	VisualBERT Google	ViLBERT Google
C6: M/W, Career/Family	W	*0.86*	*1.05*
	S	*1.05*	*1.14*
	C	-0.08	*0.71*
C8: Science/Arts, M/W	W	*0.54*	*0.51*
	S	*0.62*	*0.14*
	C	*0.25*	-0.11
C11: M/W, Pleasant/Unpleasant	W	-0.66	-0.31
	S	-0.74	-0.84
	C	*0.23*	-0.50
Double Bind: M/W, Competent	W	-0.23	*0.30*
	S	-0.10	-0.04
	C	*0.50*	-0.04
Double Bind: M/W, Likeable	W	-0.60	0.09
	S	-0.11	-1.12
	C	*0.39*	0.14
Occupations: M/W, Occupation	W	*0.91*	*1.80*
	S	*0.98*	*1.82*
	C	*1.12*	*1.55*

Race	Level	VisualBERT Google	ViLBERT Google
C3: EA/AA, Pleasant/Unpleasant	W	*0.23*	*0.14*
	S	*0.31*	-0.14
	C	*0.23*	-0.19
C12: EA/AA, Career/Family	W	-0.29	*0.43*
	S	-0.54	*0.34*
	C	*0.10*	*1.02*
C13: EA/AA, Science/Arts	W	0.04	*0.21*
	S	*0.12*	*0.68*
	C	*0.45*	*0.64*
Double Bind: EA/AA, Competent	W	*0.61*	*0.87*
	S	0.24	0.25
	C	*1.50*	-1.21
Double Bind: EA/AA, Likeable	W	0.21	-0.23
	S	*0.27*	-0.74
	C	-0.56	-1.08
Occupations: EA/AA, Occupation	W	-0.40	0.02
	S	-0.41	*0.46*
	C	-0.20	-0.80
Angry Black Woman Stereotype	W	-0.07	0.26
	S	-0.50	*0.47*
	C	*0.46*	*1.13*

Table 3: The results for all bias classes on Experiment 1 using Google Images that asks *Do joint embeddings contain biases?* Numbers represent effect sizes and p-values for the permutation test described in section 4. They are highlighted in red when p-values are below 0.05. Each bias type and model are tested three times against (W) word embeddings, (S) sentence embeddings, and (C) contextualized word embeddings. The answer to the question clearly appears to be yes. Both ViLBERT and VisualBERT are biased. This is in line with prior work on bias. Note that out of domain, biases appear to be amplified.

Gender	Level	VisualBERT Google	ViLBERT Google
C6: M/W, Career/Family	W	*0.93*	*0.89*
	S	*1.38*	*0.87*
	C	-0.13	*0.31*
C8: Science/Arts, M/W	W	*0.54*	*0.51*
	S	*0.62*	*0.14*
	C	*0.25*	-0.11
C11: M/W, Pleasant/Unpleasant	W	-1.48	-0.21
	S	-1.13	-0.79
	C	-0.11	-0.57
Double Bind: M/W, Competent	W	0.23	*0.78*
	S	-1.68	-0.79
	C	-0.18	-0.68
Double Bind: M/W, Likeable	W	-1.19	*0.36*
	S	*0.82*	-0.93
	C	-0.04	-0.61
Occupations: M/W, Occupation	W	*0.79*	*1.77*
	S	*1.07*	*1.81*
	C	*1.14*	*1.64*

Race	Level	VisualBERT Google	ViLBERT Google
C3: EA/AA, Pleasant/Unpleasant	W	*1.55*	*0.24*
	S	*1.54*	*0.12*
	C	*0.30*	*0.11*
C12: EA/AA, Career/Family	W	-0.04	*0.57*
	S	*0.36*	*0.27*
	C	*0.17*	*1.02*
C13: EA/AA, Science/Arts	W	-1.74	*0.30*
	S	-0.08	*1.14*
	C	*0.85*	*0.89*
Double Bind: EA/AA, Competent	W	*0.88*	*1.13*
	S	*0.45*	0.26
	C	*1.53*	-1.25
Double Bind: EA/AA, Likeable	W	*1.18*	0.16
	S	0.26	*0.81*
	C	-0.69	-0.44
Occupations: EA/AA, Occupation	W	-0.12	-0.19
	S	-0.34	*0.42*
	C	0.23	-0.74
Angry Black Woman Stereotype	W	0.34	*1.61*
	S	*0.49*	*1.55*
	C	*0.91*	*1.65*

Table 4: The results for all bias classes on Experiment 2 using Google Images that asks *Can joint embeddings be shown visual evidence that a bias does not apply?* Numbers represent effect sizes and p-values for the permutation test described in section 4. They are highlighted in red when p-values are below 0.05. Each bias type and model are tested three times against (W) word embeddings, (S) sentence embeddings, and (C) contextualized word embeddings. The answer to the question appears to be no, although fewer tests are statistically significant compared to table 3 showing that visual evidence is helpful.

Gender	Level	VisualBERT Google	ViLBERT Google
C6: M/W, Career/Family	W	0.15	-0.11
	S	0.64	-0.38
	C	0.05	-0.43
C8: Science/Arts, M/W	W	–	–
	S	–	–
	C	–	–
C11: M/W, Pleasant/Unpleasant	W	*1.19*	0.12
	S	0.54	*0.06*
	C	0.34	*-0.09*
Double Bind: M/W, Competent	W	0.40	0.52
	S	1.62	-0.74
	C	0.95	-0.59
Double Bind: M/W, Likeable	W	0.87	0.28
	S	0.95	0.14
	C	*0.99*	-0.70
Occupations: M/W, Occupation	W	0.15	*-0.25*
	S	0.18	0.01
	C	0.29	*0.40*

Race	Level	VisualBERT Google	ViLBERT Google
C3: EA/AA, Pleasant/Unpleasant	W	*1.46*	0.10
	S	*1.43*	*0.27*
	C	0.07	*0.30*
C12: EA/AA, Career/Family	W	0.26	0.16
	S	0.84	-0.07
	C	0.07	0
C13: EA/AA, Science/Arts	W	-1.74	0.10
	S	-0.20	0.67
	C	0.45	0.32
Double Bind: EA/AA, Competent	W	0.42	0.49
	S	0.27	0.07
	C	0.23	-0.31
Double Bind: EA/AA, Likeable	W	1.04	0.41
	S	0.09	1.30
	C	*0.16*	0.74
Occupations: EA/AA, Occupation	W	0.29	*-0.22*
	S	0.08	*-0.05*
	C	0.42	*0.09*
Angry Black Woman Stereotype	W	0.44	1.52
	S	0.87	1.44
	C	0.52	1.14

Table 5: The results for all bias classes on Experiment 3 using Google Images that asks *To what degree do biases come from language vs. vision in joint embeddings?* Numbers represent effect sizes and p-values for the permutation test described in section 4. They are highlighted in red when p-values are below 0.05. Each bias type and model are tested three times against (W) word embeddings, (S) sentence embeddings, and (C) contextualized word embeddings. This answer appears to be yes, vision does play a significant role in the structure of biases.

		Experiment 1		Experiment 2		Experiment 3	
Gender	Level	VisualBERT COCO	ViLBERT ConCap	VisualBERT COCO	ViLBERT ConCap	VisualBERT COCO	ViLBERT ConCap
C6: M/W, Career/Family	W	0.13	*1.01*	0.15	*1.12*	0.08	0.27
	S	*0.48*	*1.16*	*0.52*	*1.21*	0.13	0.16
	C	-0.24	*0.75*	-0.50	*0.64*	0.28	-0.13
Occupations: M/W, Occupation	W	-0.07	*1.44*	-0.64	*1.94*	0.59	*1.89*
	S	-0.23	*1.12*	0.09	*1.89*	*0.28*	1.80
	C	0.19	*0.67*	-0.04	*1.95*	0.22	1.94

Table 6: The results for two classes of bias on all three experiments using COCO for VisualBERT and Conceptual Captions for ViLBERT. Images for other bias classes could not be found in these datasets. These results are generally consistent with those found on Google Images.

Following Tan and Celis [2019], each experiment examines the bias in three types of embeddings: word embeddings, sentence embeddings, and contextualized word embeddings. While there is broad agreement between these different ways of using embeddings, they are not identical in terms of which biases are discovered. It is unclear which of these methods is more sensitive, and which finds biases that are more consequential in predicting the results of a larger system constructed from these models. Methods to mitigate biases will hopefully address all three embedding types and all of the three questions we restate below.

Do joint embeddings contain biases? See Experiment 1, section 4.1. The results presented in table 3 and table 6 clearly indicate the answer is yes. Numerous biases are uncovered with results that are broadly compatible with May *et al.* [2019] and Tan and Celis [2019]. It appears as if more biases exist in the grounded embeddings compared to the ungrounded ones.

Can visual evidence that counters a stereotype alleviate biases? See Experiment 2, section 4.2. The results presented in table 4 and table 6 indicate the answer is no. Biases are somewhat attenuated when models are shown evidence against them, but overall, preconceptions about biases tend to overrule direct visual evidence to the contrary. This is worrisome for the applications of such models. In particular, using such models to search or filter data in the service of creating new datasets may well introduce new biases.

To what degree do biases come from language vs. vision in joint embeddings? See Experiment 3, section 4.3. The results presented in table 5 and table 6 are generally not significant; this indicates that biases mostly arise from language and that vision contributes relatively little. It could be that joint embeddings largely ignore vision, or that the biases in language are so powerful that vision does not contribute to them given that on any one example it appears unable to override the existing biases (experiment 2). It is encouraging that models clearly do consider vision, but unfortunately the fact that biases in vision and text are not always the same, does not appear to help here.

6 Discussion

Grounded embeddings have biases and vision does not appear to help eliminate them. At test time, vision has difficulty overcoming biases, even when presented counter-stereotypical evidence. This is worrisome for deployed systems that use such embeddings, as it indicates that they ignore visual evidence that a bias does not hold for a particular interaction. Overall, vision has only a mild effect on biases, with language dominating. We enumerated all information available in the grounded setting and selected three interpretable questions that we answered above. Other questions could potentially be asked using the dataset we developed, although we did not find any others that were intuitive or non-redundant.

While we discuss joint vision and language embeddings, the methods introduced here apply to any grounded embeddings, such as joint audio and language embeddings [Kiela and Clark, 2015; Torabi *et al.*, 2016]. Measuring bias in such data would require collecting a new dataset, but could use our metrics, Grounded-WEAT and Grounded-SEAT, to answer the same three questions.

Many joint models are transferred to a new dataset without fine-tuning. We demonstrate that going out-of-domain into a new dataset amplifies biases. This need not be so: out-of-domain models have worse performance which might result in fewer biases. We did not test task-specific fine-tuned models, although intend to do so in the future.

Humans clearly have biases, not just machines. Although, initial evidence indicates that when faced with examples that goes against prejudices, i.e., counter-stereotyping, there is a significant reduction in human biases [Peck *et al.*, 2013; Columb and Plant, 2016]. Straightforward applications of this idea are far from trivial, as Wang *et al.* [2019] show that merely balancing a dataset by a certain attribute is not enough to eliminate bias. Perhaps artificially debiasing visual datasets can serve as a mechanism to debias shared embeddings. We hope that these datasets and metrics will be useful for understanding human biases in grounded settings as well as further the development of new methods to debias representations.

References

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

Patricia Hill Collins. *Black sexual politics: African Americans, gender, and the new racism*. Routledge, 2004.

Corey Columb and E Ashby Plant. The obama effect six years later: The effect of exposure to obama on implicit anti-black evaluative bias and implicit racial stereotyping. *Social Cognition*, 34(6):523–543, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530, 2016.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. *arXiv:1911.06258*, 2019.

Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, 2015.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *arXiv:1903.10561*, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv:1912.02379*, 2019.

Tabitha C Peck, Sofia Seinfeld, Salvatore M Aglioti, and Mel Slater. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and cognition*, 22(3):779–787, 2013.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv:1802.05365*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI preprint*, 2018.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13209–13220, 2019.

Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv:1609.08124*, 2016.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv:1804.06876*, 2018.