Measuring Social Biases in Grounded Vision and Language Embeddings

Candace Ross, Boris Katz & Andrei Barbu
CSAIL, Massachusetts Institute of Technology
{ccross,boris,abarbu}@mit.edu

Abstract

We generalize the notion of measuring social biases in word embeddings to visually grounded word 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 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. Many IAT tests focus on social biases, such as inherent beliefs about someone based on their racial or gender identity. Social biases have negative implications for the most marginalized people, e.g., applicants perceived to be Black based on their names are less likely to receive job interview callbacks than their white counterparts (Bertrand and Mullainathan, 2004).

Caliskan et al. (2017) introduce an equivalent of the IAT for word embeddings, called the Word Embedding Association Test (WEAT), to measure word associations between concepts. 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 visually-grounded BERT-based models by extending prior work into what we term Grounded-WEAT and Grounded-SEAT. The models we evaluate are ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019) , LXMert (Tan and Bansal, 2019) and VL-BERT (Su et al., 2019).

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 encode social biases? Since visual biases can be different from those in language, we would expect to see a difference in the biases exhibited by grounded embeddings. Biases in one modality might dampen or amplify the other. 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 used to train multimodal models are much smaller and much less diverse than language datasets.

Can grounded evidence that counters a stereotype alleviate biases? The advantage to having multiple modalities is that one modality can demon-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 998–1008
June 6–11, 2021.

strate that a learned bias is irrelevant to the particular task being carried out. For example, one might provide an image of a woman who is a doctor alongside a sentence about a doctor, and then measure the bias against women doctors in the embeddings. We find that the bias is largely not impacted, i.e., direct visual evidence against a bias helps little.

To what degree are biases encoded in grounded word embeddings from language or vision? It may be that grounded word 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, in general, both modalities contribute to encoded bias, but some model architectures are more dominated by language. Vision could have a more substantial impact on grounded word 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 first extracted images from COCO (Chen et al., 2015) and Conceptual Captions (Sharma et al., 2018); the images and English captions in these datasets lack diversity, making finding data for most existing bias tests nearly impossible. To address this, we created an additional dataset from Google Images that depicts the targets and attributes required for all bias tests considered. This work does not attempt to reduce bias in grounded models. We believe that the first critical step to doing so, is having metrics and a dataset to understand grounded biases, which we introduce here.

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

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 grounded word embeddings have social biases,
4. showing that grounded evidence has little impact on social biases, and
5. showing that biases come from a mixture of language and 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 presence of harmful social biases in such models is critical. Caliskan et al. (2017) introduce the Word Embedding Association 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; (Zhao et al., 2019) also evaluated gender bias in contexutal embeddings from ELMo. 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.

Blodgett et al. (2020) provide an in-depth analysis of NLP papers exploring bias in datasets and models and also highlight key areas for improvement in approaches. We point the reader to this paper and aim to draw from key suggestions from this work throughout.

3 The Grounded WEAT/SEAT Dataset

Existing WEAT/SEAT bias tests (Caliskan et al. (2017), May et al. (2019) and Tan and Celis (2019)) contain sentences for categories and attributes; we augment these tests to a grounded domain by pairing each word/sentence with an image. VisualBERT and ViLBERT were trained on COCO and Conceptual Captions respectively, so we use the images in these datasets' validation splits by querying the captions for the keywords. To compensate for their lack of diversity, we collected another version of the dataset where the images are top-ranked hits on Google Images. Results on COCO and Conceptual Captions are still important for the bias tests that can be collected, for two reasons. First, it gives us an indication of where datasets are lack-

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 (ABW)	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 (ABW)* 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.

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.

ing: the fact that images cannot be sourced for so many tests means these datasets particularly lack representation for these identities. Second, since COCO and Conceptual Captions form part of the training sets for VisualBERT and ViLBERT, 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. We provide original links to all collected images and scripts to download them.

4 Methods

Existing WEAT/SEAT bias tests (Caliskan et al., 2017) base the Word Embedding Association 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 systems. The average cosine similarity between pairs of word embeddings is used as the basis of an indicator of bias, as in:

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

where s measures how close on average the embedding for word w is compared to the words in attribute set A and attribute set B. Such relative distances between word vectors indicate how related two concepts are and are directly used in many natural language processing 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 may encode social biases by making some targets, e.g., men's names or women's names, closer to one profession than another. In this case, bias is defined as one of the two targets being significantly closer to one set of

Embedding index	Word
1	Man
2	Woman
3	Lawyer
4	Teacher

(a) Possible embeddings for an ungrounded model

Embedding index	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 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.

socially stereotypical 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:

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

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:

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

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 with the category of target x, and A_y and B_y, with the category of target y. Example images for the bias test for the intersectional racial and gender stereotype that black women are inherently angry, 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 grounded 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 grounded word embeddings can indicate if there is an over-reliance on one modality over another.

We evaluate bias tests on embeddings produced by Transformer-based vision and language models which take as input an image and a caption. Models are used to produce three kinds of embeddings (of

[1] An alternate way to construct such a dataset might have ambiguity about which of two agents a sentence is referring to, more closely mirroring how language is used. This would require images that simultaneously depict both targets, e.g., both a man and woman who are teachers. 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.

single-word captions, full sentence captions, and word embeddings in the context of a sentence) that are each tested for biases. These embeddings correspond to the hidden states of the language output of each model. For single-stream models like VisualBERT and VL-BERT, these are the hidden states corresponding to the language token inputs. For two-stream models like ViLBERT and LXMERT, these are the outputs of the language Transformer. When computing word and sentence embeddings, we follow May et al. (2019) and take the hidden state corresponding to the `[CLS]` token (shown in blue in fig. 2). When computing contextual embeddings, we follow Tan and Celis (2019) and take the embedding in the sequence corresponding to the token for the relevant contextual word, e.g., for the sentence "The *man* is there", we take the embedding for the token "man" (shown in green in fig. 2). Note there can be multiple contextual tokens when a contextual word is subword tokenized; we take the sequence corresponding to the first token. To mask the language, every contextual token in the input is set to `[MASK]`. To mask the image, every region of interest or bounding box with a person label is masked. Models which did not use bounding boxes during training could not be included in image masking tests.

4.1 Experiment 1: Do joint embeddings encode social 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 grounded evidence that counters a stereotype alleviate biases?

An advantage of grounded embeddings is that we can readily show scenarios that clearly counter social stereotypes. For example, the model may have a strong prior that men are more likely to have some professions, but are the embeddings different when the visual input provided shows women in those professions? 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 grounded evidence to the contrary would overcome them.

4.3 Experiment 3: To what degree are biases encoded in grounded word embeddings from language or vision?

Even if biases exist, one might wonder how much of the bias comes from language and how much comes from vision? Perhaps all of the biases come from language and vision only plays a small auxiliary role, or vice versa. We can probe this question in at least two ways. First, one could use images that are both congruent and incongruent with the stereotype. We would in that case check if the model changes its embeddings in response to the congruent or incongruent images. Similarly to eq. (3), in this case 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

VisualBERT	[CLS] TOK0 ...	TOK_CONTEXTUAL ...	TOKN [SEP]	[IMG] IMG0 ...	IMGN
VL-BERT	[CLS] TOK0 ...	TOK_CONTEXTUAL ...	TOKN [SEP]	IMG0 IMG1 ...	IMGN [END]
ViLBERT	[CLS] TOK0 ...	TOK_CONTEXTUAL ...	TOKN [SEP]		
LXMert	[CLS] TOK0 ...	TOK_CONTEXTUAL ...	TOKN [SEP]	[CROSS_MODAL]	

Figure 2: Each row shows the output sequence corresponding to a given model's output. For ViLBERT and LXMERT, we only show the output of the language Transformer. For word and sentence embeddings, we take the encoding corresponding to the [CLS] token; for contextual embeddings, we take the encpding corresponding to the word in context, [TOK_CONTEXTUAL].

depict the target concept:

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

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.

An alternate way to probe this bias makes use of the same test as in Experiment 2 with the addition of masking by taking advantage of how these models are pretrained with masked language tokens and masked image regions. VisualBERT only uses masked language modeling and never masks image regions during training; it therefore cannot be probed using this method. For each test, we alternatively mask either language tokens or image regions relevant to that specific test and measure the encoded bias. When masking image regions we mask regions that contain people. For example, in test C3, we mask every name and every pleasant or unpleasant term while token masking and every person while image masking. This ablates the potential bias in one modality, allowing us to probe the other.

5 Results

We evaluate each model on images from the dataset used for pretraining and our collected images from Google Image search. Pretraining datasets are MS-COCO for VisualBERT (Li et al., 2019) and LXMert (Tan and Bansal, 2019) and Conceptual Captions for ViLBERT (Lu et al., 2019) and VL-BERT (Su et al., 2019)[2]. Image features are computed in the same manner as in the original publications. We compute p-values using the updated permutation test described in May et al. (2019). In each case, we evaluate the task-agnostic, pretrained 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 with ungrounded models such as BERT. 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 encode social biases? See Experiment 1, section 4.1. The results presented in table 3 and table 6 clearly indicate that 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 that more pronounced social biases exist in grounded compared to ungrounded embeddings.

Can grounded evidence that counters a stereotype alleviate biases? See Experiment 2, section 4.2. The results presented in table 4 and table 6 indicate that 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

[2]Some pretraining images for VL-BERT are from the Visual Genome.

Gender	Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C6: M/W, Career/Fam	W	*0.57*	*1.04*	*0.55*	*1.61*
	S	-0.18	*0.98*	*0.69*	-0.02
	C	-0.61	*0.76*	*0.17*	*0.46*
C8: Science/Arts, M/W	W	*0.77*	*0.59*	*0.43*	-0.29
	S	*0.62*	*0.26*	–	*0.19*
	C	*0.30*	-0.32	*0.13*	*0.26*
C11: M/W, Pleasant	W	-0.66	-0.91	-0.08	-1.20
	S	-0.74	-1.08	-0.20	0.01
	C	*0.42*	-0.62	*0.25*	-0.18
Competent: M/W, Competent	W	-0.23	-0.57	-1.18	-1.28
	S	-0.28	-0.29	-0.55	-1.35
	C	-0.67	0.20	-0.48	*0.31*
Likeable: M/W, Likeable	W	-1.24	-1.26	-1.10	-0.91
	S	0.10	-0.12	*0.60*	-0.03
	C	-0.42	*1.25*	-0.83	-0.19
Occupation: M/W, Occupation	W	0.02	*0.86*	*1.56*	*1*
	S	*0.77*	*0.95*	*1.32*	-0
	C	*0.98*	*1.53*	*0.52*	0.11

Race	Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C3: EA/AA, Pleasant	W	*0.23*	*0.31*	-0.16	*1.37*
	S	*0.31*	*0.25*	*0.19*	*0.93*
	C	-0.01	-0.29	*0.44*	*0.68*
C12: EA/AA, Career/Family	W	-0.29	0.04	-0.04	-1.45
	S	-0.54	0.05	-0.32	-0.96
	C	*0.36*	*0.92*	*0.88*	0.08
C13: EA/AA, Science/Arts	W	0.04	*0.61*	*0.58*	-1.44
	S	*0.12*	*0.35*	*0.16*	*0.98*
	C	*0.58*	*1.09*	*0.92*	*0.90*
Double Bind: EA/AA, Competent	W	*0.75*	*1.28*	*0.98*	*1.44*
	S	*1*	*1.14*	*1.30*	*1.48*
	C	*1.10*	*1.19*	*1.46*	*1.54*
Double Bind: EA/AA, Likeable	W	-0.25	*0.41*	*0.93*	*0.87*
	S	-0.09	*0.73*	-0.04	*1.01*
	C	*0.97*	*1.09*	*1.40*	0.12
Occupation: EA/AA, Occupation	W	-0.15	-0.41	-0.71	1.38
	S	-0.26	-0.26	-0.40	-0.06
	C	-0.70	-0.37	-1.11	0.12
Angry Black Woman Stereotype	W	-0.07	0.41	-1.31	*1.59*
	S	-0.50	*0.46*	-0.12	-0.48
	C	*0.71*	*0.66*	*1.27*	-0.13

Table 3: The results for all bias classes on Experiment 1 using Google Images that asks *Do joint embeddings encode social biases?* Numbers represent effect sizes and p-values for the permutation test described in section 4. They are highlighted in blue 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. All models are biased. Note that out of domain, biases appear to be amplified.

Gender	Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C6: M/W, Career/Fam	W	*1.05*	*1.09*	-0.20	*1.97*
	S	-0.57	*1.34*	*0.78*	*1.57*
	C	-0.86	*0.65*	*0.21*	*0.44*
C8: Science/Arts, M/W	W	*0.77*	*0.59*	*0.43*	-0.29
	S	*0.62*	*0.26*	–	*0.19*
	C	*0.30*	-0.32	*0.13*	*0.26*
C11: M/W, Pleasant	W	-1.48	-1.33	-0.13	-0.77
	S	-1.13	-1.17	-0.55	-0.21
	C	-0.15	-0.46	*0.38*	-0.17
Competent: M/W, Competent	W	0.23	0.23	-1.37	*1.50*
	S	-0.12	-0.35	-0.98	-1.14
	C	-0.60	-0.08	-1.11	*0.44*
Likeable: M/W, Likeable	W	-1.31	-0.61	-0.93	-1.98
	S	*1.76*	-0.16	-0.81	*1.99*
	C	-0.11	*1.31*	-1	-0.12
Occupation: M/W, Occupation	W	-0.77	0.05	*1.33*	-1.74
	S	*0.33*	0.22	*0.58*	-0.20
	C	*0.90*	*1.46*	*0.34*	0.16

Race	Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C3: EA/AA, Pleasant	W	*1.55*	*1.03*	*0.60*	*1.34*
	S	*1.54*	*0.85*	*0.84*	-0.08
	C	*0.26*	-0.14	*0.58*	*0.76*
C12: EA/AA, Career/Family	W	-0.04	*0.88*	*0.93*	-1.49
	S	*0.36*	*0.81*	*0.33*	-1.27
	C	*0.84*	*1.02*	*0.98*	*0.18*
C13: EA/AA, Science/Arts	W	-1.74	*1.27*	-0.38	-1.51
	S	-0.08	*1.04*	-0.13	*0.95*
	C	*1*	*1.39*	*0.97*	*0.96*
Double Bind: EA/AA, Competent	W	*1.13*	*1.56*	*1.06*	*1.41*
	S	*1.25*	*1.45*	*1.25*	*1.45*
	C	*1.11*	*1.20*	*1.46*	*1.57*
Double Bind: EA/AA, Likeable	W	*0.29*	*1.13*	*1.29*	*0.90*
	S	*0.42*	*1.04*	*0.43*	*1.29*
	C	*0.93*	*1.12*	*1.40*	0.06
Occupation: EA/AA, Occupation	W	-0.04	-0.48	-0.33	-1.40
	S	0.15	-0.18	0.22	-0.03
	C	-0.57	-0.19	-1.10	0.10
Angry Black Woman Stereotype	W	0.34	-0.28	-0.27	*1.67*
	S	*0.49*	-0.53	*0.31*	0.03
	C	*1.71*	*1.44*	*1.34*	-0.21

Table 4: The results for all bias classes on Experiment 2 using Google Images that asks *Can joint embeddings be shown grounded 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 blue 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.

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 are encoded biases in joint embeddings from language or vision? See Experiment 3, section 4.3. The results for the second variant of Experiment 3 which is performed by masking the input text or image are presented in table 5 and table 6 are generally significant, more so for language than vision. We report results for the sentence-level encoding and observed similar results for the word-level encoding. We did not measure contextual encodings as they would include the encoding for the `[MASK]` token. This indicates that biases arise from both modalities, but this does differ by model architecture. For VLBERT language appears to dominate. The results for the first variant of Experiment 3 congruent with

Gender	Mask	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C6	T	*0.14*	*1*	*1.18*	-0
	I	–	*0.87*	*0.69*	-0.03
C8	T	*0.46*	*0.41*	*0.11*	*0.27*
	I	–	*0.39*	0.04	*0.18*
C11	T	-0.47	-1.21	-1.33	0.03
	I	–	-1.11	-0.22	0.02
Competent	T	-0.06	-0.40	-0.21	-1.99
	I	–	-0.35	-0.55	-1.05
Likeable	T	-0.07	-0.18	*0.28*	-1.99
	I	–	-0.11	*0.72*	*0.64*
Occupation	T	0.05	*1.08*	*0.92*	-0.17
	I	–	*0.91*	*1.32*	0

Race	Mask	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
C3	T	*0.33*	*0.34*	*0.33*	-0.01
	I	–	*0.31*	*0.21*	*0.95*
C12	T	-0.52	0.05	-0.39	0
	I	–	0.08	-0.36	-1.06
C13	T	-0	*0.33*	-0.10	-0
	I	–	*0.33*	*0.17*	*0.95*
Competent	T	-0.44	*1.10*	*1.33*	-1.99
	I	–	*1.15*	*1.29*	*1.45*
Likeable	T	-0.68	*0.58*	0.11	-1.99
	I	–	*0.73*	-0.14	*1.06*
Occupation	T	-0.27	-0.24	-0.65	-0.17
	I	–	-0.30	-0.38	-0.25
ABW	T	*0.76*	*0.54*	-0.01	-0.42
	I	–	*0.43*	-0.13	-0.08

Table 5: The results for all bias classes on Experiment 3, using the second masking variant of the experiment, with Google Images asking the question *To what degree are biases encoded in grounded word embeddings from language or vision?* Numbers represent effect sizes and p-values for the permutation test described in section 4. All numbers were measured over sentence-level encodings. They are highlighted in blue when p-values are below 0.05. Biases are measured for masked tokens (T) and masked image regions (I). This answer appears to be that both vision and language play a significant role, but this differs across model architectures.

		Experiment 1				Experiment 2				Experiment 3				
Gender	Level	VisualBERT COCO	ViLBERT ConCap	LXMert COCO	VLBERT ConCap	VisualBERT COCO	ViLBERT ConCap	LXMert COCO	VLBERT ConCap	Mask	VisualBERT COCO	ViLBERT ConCap	LXMert COCO	VLBERT ConCap
C6	W	0.13	*0.94*	*0.92*	-0.14	0.15	*0.95*	*0.61*	*1.98*	T	–	*1.15*	0.01	0
	S	*0.28*	*1.11*	*1.32*	0	*0.41*	*0.83*	*1.16*	-1.17	I	–	*1.09*	*1.32*	-0
	C	-0.20	*0.80*	*1.53*	*0.61*	-0.99	*0.58*	*1.46*	0					
Occupation	W	-0.07	*0.75*	*0.39*	-0.31	-0.64	-0.52	-0.66	*1.99*	T	–	*0.74*	-0.07	0
	S	-0.23	*0.73*	-0.18	-0.01	0.09	-0.30	-1.14	*0.69*	I	–	*0.71*	-0.17	-0
	C	-0.32	*0.58*	-0.14	0.01	-0.35	*1.96*	-0.70	*0.90*					

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

Number of statistically significant tests out of 6 total gender bias tests

	Experiment 1				Experiment 2				Experiment 3				
Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google	Mask	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
W	2	2	3	3	2	2	2	2	T	-	1	3	4
S	2	1	3	3	3	3	2	2	I	-	2	3	3
C	3	3	3	4	2	3	3	4					

Number of statistically significant tests out of 7 total race bias tests

	Experiment 1				Experiment 2				Experiment 3				
Level	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google	VisualBERT Google	ViLBERT Google	LXMert Google	VLBERT Google	Mask	ViLBERT Google	ViLBERT Google	LXMert Google	VLBERT Google
W	2	4	4	3	3	4	5	4	T	-	0	5	2
S	3	4	5	3	4	3	5	5	I	-	4	5	3
C	5	7	5	6	6	4	5	6					

Table 7: A summary of all previous results on the new image dataset derived from Google searches showing the number of significant bias test partitioned by the type of test. There are a total of 6 gender bias tests and 7 race bias test. Experiments 1 and 2 show no strong differences between models while in Experiment 3 ViLBERT stands out.

these results, with, large effect sizes (s=0.42 for ViLBERT and s=0.467 for VisualBERT with 12% of tests being statistically significant) demonstrating that language contributes more than vision. It could be that the biases in language are so powerful that vision does not contribute to them given that in any one example it appears unable to override the existing biases (experiment 2). It is encouraging that models do consider vision, but the differing biases in vision and text do not appear to help.

6 Discussion

Visually grounded embeddings have biases similar to ungrounded embeddings 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, language and vision each contribute to encoded bias, yet the means of using vision to mitigate is not immediately clear. We enumerated the combinations of inputs possible 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, but intend to do so in the future.

Humans clearly have biases, not just machines. Although, initial evidence indicates that when faced with examples that go 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 manipulating visual datasets can debias shared embeddings. We hope that these datasets and metrics will lead to understanding human biases in grounded settings as well as the development of new methods to debias representations.

Acknowledgments

This work was supported by the Center for Brains, Minds and Machines, NSF STC award 1231216, the Toyota Research Institute, the MIT CSAIL Systems that Learn Initiative, the NSF Graduate Research Fellowship, the DARPA GAILA program, the United States Air Force Research Laboratory under Cooperative Agreement Number FA8750-19-2-1000, and the Office of Naval Research under Award Number N00014-20-1-2589 and Award Number N00014-20-1-2643. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Ethical Considerations

We would like to urge subsequent work to avoid a common ethical problem we have noticed while reviewing the literature on bias in NLP. Much prior work refers to gender as "male" and "female", thereby conflating gender and sex. Recent work in psychology has disentangled these two concepts, and conflating them both blinds us to a type of bias while actively causing harm.

Our approach studies societal biases in models. These biases are inherently unjust, predisposing models toward judging people by skin color, age, etc. They are also practically damaging; they can result in real-world consequences. As part of large systems these biases may not be apparent as the source of discrimination, and it may not even be apparent that systems are treating individuals differently. People may even acclimatize to being treated differently or may interpret a machine discriminating based on race or gender as an inevitable but fair consequence of using a particular algorithm. We vehemently disagree. All systems and algorithm choices are made by humans, all data is curated by humans, and ultimately humans decide what to do with and when to use models. All unequal outcomes are a deliberate choice; engineers should not be able to hide behind the excuse of a black-box or a complex algorithm. We believe that by revealing biases, by providing tests for biases that are as focused as possible on the smallest units of systems, we can both assist the development of better models and allow the auditing of models to ascertain their fairness.

Data was collected in an ethical manner approved by the institution IRB board. No crowd-

sourced workers were employed. Instead we used a *top k* keyword search on Google Images. Because we collected images from the web, there is no straightforward way to use self-identified characteristics for gender and race. We expect biases and preconceived notions of identity to have some bearing on label accuracy. The dataset includes images available for free on the web and simple captions, e.g., Here is a man.

The biases we evaluate in this paper are based on various theories and works in psychology, such as the trope of the angry Black woman. Of course, that literature itself is limited; there are many biases which affect billions of people but do not appear in any available test, e.g., for almost any ethnic group there are those who will believe they do not work hard, but there are virtually no ethnic-group-specific tests. There are also likely biases which we have not yet articulated. Unfortunately, at present there is no coherent theory of biases to generate an exhaustive list and test them.

References

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hann Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of ACL*.

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

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

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

Corey Columb and E Ashby Plant. 2016. 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.

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

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. 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.

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

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

Douwe Kiela and Stephen Clark. 2015. 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.

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

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

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

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

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

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

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

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

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

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

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

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv:1904.03310*.

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