Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models

Tejas Srinivasan  
University of Southern California  
tejas.srinivasan@usc.edu

Yonatan Bisk  
Carnegie Mellon University  
ybisk@cs.cmu.edu

Abstract
Numerous works have analyzed biases in vision and pre-trained language models individually - however, less attention has been paid to how these biases interact in multimodal settings. This work extends text-based bias analysis methods to investigate multimodal language models, and analyzes intra- and inter-modality associations and biases learned by these models. Specifically, we demonstrate that VL-BERT (Su et al., 2020) exhibits gender biases, often preferring to reinforce a stereotype over faithfully describing the visual scene. We demonstrate these findings on a controlled case-study and extend them for a larger set of stereotypically gendered entities.

1 Introduction
Pre-trained contextualized word representations (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; Lan et al., 2020; Raffel et al., 2020) have been known to amplify unwanted (e.g. stereotypical) correlations from their training data (Zhao et al., 2019; Kurita et al., 2019; Webster et al., 2020; Vig et al., 2020). By learning these correlations from the data, models may perpetuate harmful racial and gender stereotypes.

The success and generality of pre-trained transformers has led to several multimodal representation models (Su et al., 2020; Tan and Bansal, 2019; Chen et al., 2019) which utilize visual-linguistic pre-training. These models also condition on the visual modality, and have shown strong performance on downstream visual-linguistic tasks. This additional input modality allows the model to learn both intra- and inter-modality associations from the training data - and in turn, gives rise to unexplored new sources of knowledge and bias. For instance, we find (see Figure 1) the word purse’s female association can override the visual evidence. While there are entire bodies of work surrounding bias in vision (Buolamwini and Gebru, 2018) and language (Blodgett et al., 2020), there are relatively few works at the intersection of the two. As we build models that include multiple input modalities, each containing their own biases and artefacts, we must be cognizant about how each of them are influencing model decisions.

In this work, we extend existing work for measuring gender biases in text-only language models to the multimodal setting. Specifically, we study how within- and cross-modality biases are expressed for stereotypically gendered entities in VL-BERT (Su et al., 2020), a popular visual-linguistic transformer. Through a controlled case study ($\S$3), we find that visual-linguistic pre-training leads to VL-BERT viewing the majority of entities as “more masculine” than BERT (Devlin et al., 2019) does. Additionally, we observe that model predictions rely heavily on the gender of the agent in both the language and visual contexts. These findings are corroborated by analysis over a larger set of gendered entities ($\S$4).

2 Methodology
2.1 Sources of Gender Bias
We define gender bias as undesirable variations in how the model associates an entity with different genders, particularly when they reinforce harmful
We identify three sources of learned bias when visual-linguistic models are making masked word predictions - visual-linguistic pre-training, the visual context, and the language context. Visual-linguistic pre-training refers to biases the model has learned while being exposed to image-text pairs during pre-training, whereas the other two are biases expressed during inference.

### 2.2 Measuring Gender Bias

We measure associations between entities and gender in visual-linguistic models using template-based masked language modeling, inspired by methodology from Kurita et al. (2019). We provide template captions involving the entity \( E \) as language inputs to the model, and extract the probability of the [MASK]-ed entity. We denote extracted probabilities as:

\[
P_{L/VL}(E|g) = P([\text{MASK}] = E|g \text{ in input})
\]

where \( g \) is a gendered agent in one of the input modalities. \( L \) and \( VL \) are the text-only BERT (Devlin et al., 2019) and VL-BERT (Su et al., 2020) models respectively. Our method for computing association scores is simply:

\[
S(E, g) = \ln \frac{P(E|g)}{P(E|g_N)}
\]

where probabilities in the numerator and denominator vary depending on the bias source we want to analyze. We summarize how we vary our normalizing term and compute association scores for each bias source in Table 1.

1. **Visual-Linguistic Pre-Training (\( S_{PT} \))**: We compute the association \( \text{shift} \) due to VL pre-training, by comparing the extracted probability \( P_{VL} \) from VL-BERT with the text-only BERT - thus \( P_L \) is the normalizing term.

\[
S_{L}(E, g) = E_{I \sim I_E} [S_L(E, g|I)]
\]

2. **Language Context (\( S_{L} \))**: For an image \( I \), we replace the gendered agent \( g \) with the gender-neutral term person (\( p \)) in the caption, and compute the average association score over a set of images \( I_E \) which contain the entity \( E \).

\[
S_{L}(E, g) = E_{I \sim I_E} [S_L(E, g|I)]
\]

3. **Visual Context (\( S_{V} \))**: We collect a set of images \( I_g \) which contain the entity \( E \) and gendered agent \( g \), and compute the average extracted probability by providing language input with gender-neutral agent:

\[
\hat{P}_{VL}(E|I_g) = E_{I \sim I_g}[P_{VL}(E|I)]
\]

We normalize by comparing to the output when no image is provided (\( P_{VL}(E) \)).

For each bias source, we can compute the bias score for that entity by taking the difference of its female and male association scores:

\[
B(E) = S(E, f) - S(E, m)
\]

The sign of \( B(E) \) indicates the direction of gender bias - positive for “female,” negative for “male.”

### 3 Case Study

In this section, we present a case study of our methodology by examining how gender bias is expressed in each bias source for several entities.
3.1 Entities

We perform an in-depth analysis of three pairs of entities, each representing a different type of entity: clothes (apron, suit), bags (briefcase, purse), and drinks (wine, beer). The entities are selected to show how unequal gender associations perpetuate undesirable gender stereotypes (e.g., aprons are for women, while suits are for men). For each entity, we also collect a balanced set $I_E = I_f \cup I_m$ of 12 images - 6 images each with men ($I_m$) and women ($I_f$) (images in Appendix A).

For each entity pair, we created a different template caption (Table 2). We use these template captions to compute association scores $S(E, g)$, where $g \in G = \{\text{male}, \text{female}\}$.

In the following sections, we analyze how VL-BERT exhibits gender bias for these entities, for each of the bias sources identified in Section 2.1.

3.2 Visual-Linguistic Pre-Training Bias

In Figure 2, we plot each entity’s pre-training association shift score, $S_{PT}(E, m/f)$, where positive scores indicate that visual-linguistic pre-training amplified the gender association, while negative shift scores indicate weakened associations. The difference between female and male association shift scores represents the entity’s gender bias caused by visual-linguistic pre-training, $B_{PT}(E)$.

It is immediately evident that visual-linguistic pre-training affects all objects differently. Some objects have increased association scores for both genders (briefcase), while others have decreased associations (suit and apron). Moreover, even when the associations shift in the same direction for both genders, they rarely move together - for briefcase, the association scores for both genders shift positively, but to a much larger degree for male. On the other hand, for apron, wine and beer, the association shifts are negative for both genders, but the dampening is more pronounced for female. The exception is suit, for which both association shift scores are approximately the same.

We see a third type of behavior with the entity purse, where association shifts positively for male but negatively for female. Combining these trends, we conclude that VL-BERT generally appears to have more male-associated entities than BERT.

3.3 Language Context Bias

Figure 3 plots language association scores, which look at the masked probability of $E$ when the agent in the caption is man/woman, compared to the gender-neutral person.

For the entity purse, we see that when the agent in the language context is female the model is much more likely to predict that the masked word is purse, but when the agent is male the probability becomes much lower. We similarly observe that some of the entities show considerably higher confidence when the agent is either male or female (briefcase, apron, beer), indicating that the model has a language gender bias for these enti-
Figure 4: Visual association scores $S_V(E, m/f)$. Positive association scores indicate that the model becomes more confident in the presence of a visual context.

For some entities (suit and wine), VL-BERT does not exhibit considerable bias as association scores with both genders are similar.

### 3.4 Visual Context Bias

For each of our entities, we also plot the visual association score $S_V(E, u)$ with male and female in Figure 4. We again observe that the degree of association varies depending on whether the image contains a man or woman. For purse and apron, the model becomes considerably more confident in its belief of the correct entity when the agent is female rather than male. Similarly, if the agent is male, the model becomes more confident about the entity in the case of briefcase and beer. For suit and wine, the differences are not as pronounced. In Table 3, we can see some examples of the model’s probability outputs not aligning with the object in the image. In both cases, the model’s gender bias overrides the visual evidence (the entity).

### 4 Comparing Model Bias with Human Annotations of Stereotypes

To test if the trends in the case study match human intuitions, we curate a list of 40 entities, which are considered to be stereotypically masculine or feminine in society.\(^3\) We analyze how the gendered-ness of these entities is mirrored in their VL-BERT language bias scores. To evaluate the effect of multimodal training on the underlying language model, we remove the visual input when extracting language model probabilities and compare how the language bias varies between text-only VL-BERT and the text-only BERT model.

\(^3\)We surveyed 10 people and retained 40/50 entities where majority of surveyors agreed with a stereotyped label.

| Visual Context, $I$ | $P_{VL}(purse|I)$ | $P_{VL}(briefcase|I)$ |
|--------------------|-----------------|-------------------|
| purse              | 0.0018 ✓        | 0.084 x           |
| briefcase          | 0.4944 x        | 0.067 ✓           |

Table 3: Examples of images where the probability outputs do not align with the visual information.

For the language input, we create template captions similar to those described in Table 2. For every entity $E$, we compute the language bias score $B_L(E)$ by extracting probabilities from the visual-linguistic model, $P_{VL}(E|f/m/p)$.

$$S_L(E, m/f) = \ln \frac{P_{VL}(E|m/f)}{P_{VL}(E|p)}$$

$$B_VLBert_L(E) = S_L(E, f) - S_L(E, m) = \ln \frac{P_{VL}(E|f)}{P_{VL}(E|m)}$$

Positive values of $B_VL(E)$ correspond to a female bias for the entity, while negative values correspond to a male bias. We plot the bias scores in Table 5a. We see that the language bias scores in VL-BERT largely reflect the stereotypical genders of these entities - indicating that the results of Section 3.3 generalize to a larger group of entities.

We can also investigate the effect of visual-linguistic pretraining by comparing these entities’ VL-BERT gender bias scores with their gender bias scores under BERT. We compute the language bias score for BERT, $B_{Bert}^L(E)$, by using the text-only language model probability $P_L(E|g)$ instead. We plot the difference between entities’ VL-BERT and BERT bias scores in Table 5b. Similar to trends observed in Section 3.2, we see that the majority of objects have increased masculine association after pre-training ($B_VL^{VLBert} < B_VL^{Bert}$).

### 5 Related Work

**Vision-and-Language Pre-Training** Similar to BERT (Devlin et al., 2019), vision-and-language transformers (Su et al., 2020; Tan and Bansal, 2019; Chen et al., 2019) are trained with masked language modeling and region modeling with multiple input modalities. These models yield state-of-the-art results on many multimodal tasks: e.g. VQA (Antol et al., 2015), Visual Dialog (Das et al., 2017), and VCR (Zellers et al., 2019).
Bert for 40 entities which are stereotypically considered masculine or feminine. For the majority of entities, the direction of the gender bias score aligns with the stereotypical gender label, indicating that VL-BERT reflects these gender stereotypes.

(a) $B_{VL-BERT}^{L}(E)$ for 40 gendered entities. The distribution of entities is skewed towards increased masculine/decreased feminine association for VL-BERT, indicating VL pre-training shifts the association distribution for most entities towards men. Note that VL-BERT still associates *cat* with women and *cigar* with men (see 5a), but less strongly than BERT.

(b) $B_{VL-BERT}^{L}(E) - B_{BERT}^{L}(E)$ for the 40 gendered entities. The distribution of entities is skewed towards increased masculine/decreased feminine association for VL-BERT, indicating VL pre-training shifts the association distribution for most entities towards men. Note that VL-BERT still associates *cat* with women and *cigar* with men (see 5a), but less strongly than BERT.

Figure 5

Bias Measurement in Language Models
Bolukbasi et al. (2016) and Caliskan et al. (2017) showed that static word embeddings like Word2Vec and GloVe encode biases about gender roles. Biases negatively effect downstream tasks (e.g., coreference (Zhao et al., 2018; Rudinger et al., 2018)) and exist in large pretrained models (Zhao et al., 2019; Kurita et al., 2019; Webster et al., 2020). Our methodology is inspired by Kurita et al. (2019), who utilized templates and the Masked Language Modeling head of BERT to show how different probabilities are extracted for different genders. We extend their text-only methodology to vision-and-language models.

Bias in Language + Vision Several papers have investigated how dataset biases can override visual evidence in model decisions. Zhao et al. (2017) showed that multimodal models can amplify gender biases in training data. In VQA, models make decisions by exploiting language priors rather than utilizing the visual context (Goyal et al., 2017; Ramakrishnan et al., 2018). Visual biases can also affect language, where gendered artefacts in the visual context influence generated captions (Hendricks et al., 2018; Bhargava and Forsyth, 2019).

6 Future Work and Ethical Considerations
This work extends the bias measuring methodology of Kurita et al. (2019) to multimodal language models. Our case study shows that these language models are influenced by gender information from both language and visual contexts - often ignoring visual evidence in favor of stereotypes.

Gender is not binary, but this work performs bias analysis for the terms “male” and “female” – which are traditionally proxies for cis-male and cis-female. In particular, when images are used of male and female presenting individuals we use images that self-identify as male and female. We avoid guessing at gender presentation and note that the biases studied here in this unrealistically simplistic treatment of gender pose even more serious concerns for gender non-conforming, non-binary, and trans-sexual individuals. A critical next step is designing more inclusive probes, and training (multi-modal) language models on more inclusive data. We welcome criticism and guidance on how to expand this research. Our image based data suffers from a second, similar, limitation on the dimension of race. All individuals self-identified as “white” or “black”, but a larger scale inclusive data-collection should be performed across cultural boundaries and skin-tones with the self-identification and if appropriate prompts can be constructed for LLMs.

References
Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In 2015 IEEE International Conference on
Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Aylin Caliskan, Joanna J Bryson, and Arvind Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. arXiv preprint arXiv:1912.00578.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 4349–4357.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1080–1089. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 6325–6334. IEEE Computer Society.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In European Conference on Computer Vision, pages 793–811. Springer.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21:1–67.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 1548–1558.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. **VL-BERT: pre-training of generic visual-linguistic representations.** In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hao Tan and Mohit Bansal. 2019. **LXMERT: Learning cross-modality encoder representations from transformers.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. **Causal mediation analysis for interpreting neural nlp: The case of gender bias.** *arXiv preprint arXiv:2004.12265*.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. **Measuring and reducing gendered correlations in pre-trained models.** *arXiv preprint arXiv:2010.06032*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. **From recognition to cognition: Visual commonsense reasoning.** In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. **Gender bias in contextualized word embeddings.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints.** In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. **Gender bias in coreference resolution: Evaluation and debiasing methods.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
### A Images Collected for Case Study

In Table 4, we show the different images collected for our Case Study in Section 3.

| Entity | Gender of Agent | Images Used ($I_{m/f}$) |
|--------|-----------------|-------------------------|
| Purse  | Male            | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |
| Briefcase | Male   | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |
| Apron  | Male            | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |
| Suit   | Male            | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |
| Wine   | Male            | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |
| Beer   | Male            | ![Images](image-url)     |
|        | Female          | ![Images](image-url)     |

Table 4: Images collected for case study in Section 4