Evaluating Biases in Text Embeddings of Multi-modal Models: A CLIP Case Study

Luis Henrique Simplicio Ribeiro, Isha Vaish, Alison Yao

Motivation

Evaluating gender and racial biases in text embeddings for Natural Language Processing (NLP) tasks is a crucial step toward ensuring **algorithm fairness**. The effect of such biases is further exacerbated in **multi-modal** models that combine text and image information in a variety of applications such as image captioning and text-to-image generation.

Algorithm Fairness: Ethical principle that algorithms should treat all individuals and groups fairly to ensure algorithm outcomes and the decisions made from those algorithms are without discrimination.

Multi-Modal Models: Models that can handle and integrate information from different modalities to make predictions.

A popular multi-modal model is OpenAI's **CLIP** [3](Contrastive Language-Image Pretraining). Biases in models like CLIP can lead to skewed outcomes in areas like disease diagnosis, public safety, and decision-making in autonomous systems.

Example: Biased CLIP Embeddings in Stable Diffusion
A popular text-to-image generative model that uses CLIP to generate text embeddings is **Stable Diffusion** [4]. The image generation results for Stable Diffusion 1.4 [6] are given below for the following prompts:

- "High resolution image of a doctor" (Figure 1, Row 1)
- "High resolution image of a criminal" (Figure 1, Row 2)
- "High resolution image of teacher" (Figure 1, Row 3)

The majority of images returned for doctors were of white males, for criminals were of darker-skinned males, and for teachers were of white females - highlighting implicit biases in the model's text embeddings.

Figure 1: Stable Diffusion Example

Background

CLIP is a multi-modal Deep Neural Network trained on 400 million (text, image) pairs to jointly learn visual concepts and natural language embeddings by using two separate encoders:

1. **Vision transformer[1]:** to embed images
2. **Transformer-based language encoder[5]:** to embed tokenized textual descriptions

An overview of the CLIP model's architecture is given in figure 2.

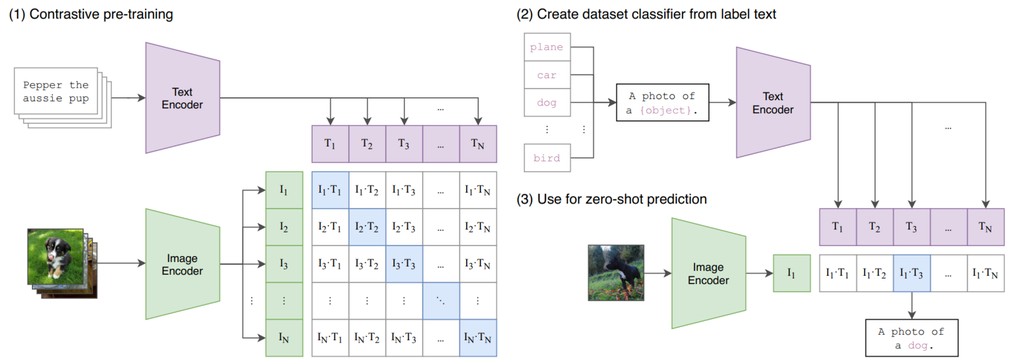

Figure 2: CLIP Architecture

Training:

Use a contrastive learning objective to jointly train two models:

- A text encoder: $\mathcal{E}_T : \mathcal{X}_T \to \mathbb{R}^d$
- An image encoder: $\mathcal{E}_I : \mathcal{X}_I \to \mathbb{R}^d$

on a dataset $\mathcal{D} = \{(\text{image}^{(i)}, \text{text}^{(i)})\}_{i=1}^{n}$ where n is the number of training samples.

Contrastive Learning Objective:

Let $\mathcal{E}_T(\text{text}^{(i)})$ be the text embedding for input pair i's text description and $\mathcal{E}_I(\text{image}^{(i)})$ be the embedding for input pair i's image, CLIP:

1. Maximizes the similarities between positive pairs $(\text{image}^{(i)}, \text{text}^{(i)})$ such that $\mathcal{E}_T(\text{text}^{(i)}) \approx \mathcal{E}_I(\text{image}^{(i)})$
2. Minimizes the similarities between negative pairs $(\text{image}^{(i)}, \text{text}^{(j)})$ for $i \neq j$.

Inference:

CLIP can predict how semantically related a given image and text pair is by measuring the similarity between them in a shared space. Common metrics to measure similarity include:

- Cosine Similarity
- CLIP Score

Cause for Bias Propagation:

Implicit biases in the text embeddings can be reflected in the model's output where images and text are predicted to be semantically related under a biased context.

Main Contributions

Contribution 1: Created a clean dataset that can be used to evaluate model robustness to gender and racial biases that overcomes the lack of open-source profession datasets containing race and gender information.

Contribution 2: Exposed racial and gender biases present in CLIP's text encoder by evaluating its performance on the dataset.

Data Collection

Task: Collect images representing different races and genders in professions historically associated with biased connotations.

Methodology: Scrape Google Images for 5 images per race-gender-profession combination across 12 professions, 2 genders, and 4 races for a total of 480 images.

- **Profession:** Doctor, Teacher, Scientist, Nurse, Police Officer, CEO, Construction Worker, Lawyer, Criminal, Janitor, Model, Athlete
- **Gender:** Male, Female
- **Race:** White, Black, Asian, Latin American

Model Inference

Model: Open-source CLIP hosted on HuggingFace [2] with pre-trained weights from version "openai/clip-vit-base-patch32"

Similarity Metric: The CLIP score equation for evaluating semantic similarity of a (text, image) pair is given below.

$$\text{CLIPscore}(\text{text}, \text{image}) = \max\left(100\frac{z_{\text{text}} \cdot z_{\text{image}}}{\|z_{\text{text}}\|\|z_{\text{image}}\|}, 0\right)$$

where $z_{\text{text}} = \mathcal{E}_T(\text{text})$ and $z_{\text{image}} = \mathcal{E}_I(\text{image})$, with $z_{\text{text}}, z_{\text{text}} \in \mathbb{R}^d$ are the respective text and image embeddings for a given (text, image) pair.

Experiments: Run experiments to identify if certain professions are more associated with certain races and genders with the methodology described below.

1. Create (text, image) pairs for each image in the dataset where the text prompt describes the profession and reads as follows: "A photo of a [profession]".
2. Calculate the CLIP score for each (text,image) pair.
3. To compare the effect of race on the model's predictions, average the CLIP score by race for each profession: $s_{\text{profession,race}}$
4. To compare the effect of gender on the model's predictions, average the CLIP score by gender for each profession: $s_{\text{profession,gender}}$

If the CLIP score $s_{\text{profession,race}}$ is different across race or if $s_{\text{profession,gender}}$ is different across gender, it is likely that CLIP's text embeddings are biased.

Results

Model Results Across Race

The results (Table 1) are consistent with U.S. racial stereotypes. For example the average CLIP scores by race indicate that

- Asians are the most associated race for professions such as CEO, doctor, lawyer, nurse, and teacher.
- Latin Americans are most associated with the construction worker profession.
- White individuals are most associated with professions such as model, police officer, and scientist.
- Black individuals are the most associated with the athlete profession.

Figure 3 shows a specific example of the construction worker profession; the images chosen are the ones which have the median CLIP score for Construction workers of each Race.

Profession / Race	Athlete	CEO	Worker	Criminal	Doctor	Janitor	Lawyer	Model	Nurse	Cop	Scientist	Teacher
Asian	26.12	**26.56**	28.72	**26.43**	**30.43**	28.76	**27.51**	24.95	**30.24**	29.48	28.3	**28.46**
Black	**27.05**	25.09	27.7	25.28	28.84	28.39	26.04	24.41	29.19	28.99	26.98	26.13
Latinx	25.4	26.19	**28.83**	26.38	28.6	28.39	26.26	25.16	29.85	29.24	25.7	27.2
White	26.21	25.99	27.35	25.34	29.44	**29.74**	27.43	**26.13**	29.69	**29.75**	**28.75**	27.86

Table 1: Average CLIP score across race

Figure 3: Examples of racial bias

Model Results Across Gender

The results (Table 2) are consistent with gender stereotypes in the U.S. For example, the average CLIP scores across gender indicate that

- males are more associated with the doctor profession while females are more associated with the nurse profession.
- males are most associated with professions such as CEO, lawyer, scientist, police officer, teacher, and construction worker than females.

Figure 4 shows a specific example of the construction worker profession; the images chosen are the ones which have the median CLIP score for Construction workers of each Gender. In this example, all of the males have higher CLIP scores than the females.

Profession / Gender	Athlete	CEO	Worker	Criminal	Doctor	Janitor	Lawyer	Model	Nurse	Cop	Scientist	Teacher
Female	**26.26**	25.0	27.14	**25.94**	28.69	28.17	26.7	**25.43**	**30.65**	28.88	26.63	27.02
Male	26.14	**26.92**	**29.16**	25.77	**29.96**	**29.47**	**26.92**	24.9	28.84	**29.86**	**28.24**	**27.8**

Table 2: Average CLIP score across gender

Figure 4: Examples of gender bias

Possible Reason For Biased Text Embeddings

A common reason for biases in models, especially in text encoders, is a lack of a representative training dataset. CLIP's training data may have a skewed distribution of images for each profession across different races and genders. Unfortunately, this skewed distribution may reflect the racial and gender stereotypes associated with certain professions making it difficult to find a representative dataset.

Dataset Constraint

For some professions, the results don't match the typical stereotypes. For example, the results of the criminal category in both tables go against prevalent stereotypes and challenges preconceived notions, highlighting the complexity of how these categories are represented in the data. Potential reasons for this are given below.

- **Insufficient data:** 5 images per race, gender, and profession combination may not be sufficient for achieving a highly representative sample.
- **Lack of image diversity:** Finding a diverse set of images for each profession, gender, and race combination proved to be challenging as some combinations do not have many high-quality (non-blurry, no watermarks, et cetera) images on the internet.

These dataset constraints may contribute to unexpected patterns in CLIP scores. Future work should aim for a more comprehensive and diverse dataset to ensure a more accurate reflection of real-world associations.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] OpenAI. Clip: Vision-text embedding model (vit-base-patch32). `https://huggingface.co/openai/clip-vit-base-patch32`, 2021. Hugging Face Model Hub.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[6] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.