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

Luis Henrique Simplicio Ribeiro, Isha Vaish, Alison Yao
Harvard University
lsimplicioribeiro@g.harvard.edu,ishavaish@g.harvard.edu,
yuhan_yao@g.harvard.edu

Abstract

Evaluating gender and racial biases in text embeddings for Natural Language Processing (NLP) tasks is a crucial step towards ensuring algorithm fairness. The effect of such biases is further exacerbated in multi-modal models, which combine both text and image information and are used in a variety of applications such as image captioning and image generation. In this study, we conduct a deep dive into a widely used multi-modal model for image captioning and text-to-image generation applications: OpenAI's Contrastive Language-Image Pretraining (CLIP) model. In particular, we analyze the relationship between CLIP's text and image embeddings concerning professionals in occupations historically laden with racial and gender biases.

1 Introduction

CLIP [8] is multi-model Deep Neural Network trained to jointly learn visual concepts and natural language embeddings by using two separate encoders. Images are converted into numerical embeddings using a vision transformer [4] while text descriptions are tokenized and converted into embeddings using a transformer-based [11] language encoder. CLIP uses a contrastive learning objective where it learns to bring corresponding image and text embedding pairs closer together in a shared embedding space. Once trained, CLIP can predict how semantically related a given image and text pair is by measuring the distance (similarity) between them in the shared space. A diagram of the CLIP model's architecture and basic process is given in figure 1 [8]. If the text embeddings carry implicit biases, these biases can be reflected in the model's output where images and text are predicted to be semantically related under a biased context.

An example of this can be demonstrated using Stable Diffusion [9]: a popular text-to-image generation model that uses CLIP to encode its image generation prompt. Figure 2 shows the Stable Diffusion results when prompted to return a "high resolution image of a doctor." All the images returned were of white males; the results clearly reflect the historic racial and gender biases associated with the doctor profession.

By thoroughly evaluating the possible gender and racial biases in CLIP's text encoder, we can identify areas of improvement and warn users of the possible consequences of biased text encodings. Biases in such models can lead to skewed outcomes in areas like disease diagnosis, public safety, and decision-making in autonomous systems. Evaluating biases in multi-modal models encourages the development of fairer algorithms/methodologies and is essential in preventing unfair treatment based on gender or race.

To analyze the implicit biases present in CLIP's text embedding model, we use CLIP to predict the most likely image and profession pairings given a hand-crafted dataset containing images of individuals with different races and gender across different professions. Then for each profession, we can compare which race and gender is more likely to be associated with it. If CLIP's text embedding method is indeed biased, we expect to see common profession-related racial and gender stereotypes reflected in the model's similarity scores between each profession (text) and the corresponding professionals (images for that profession categorized by race and gender).

2 Related Work

Social biases in multi-modal models have been studied recently. Agarwal et al. [1] from OpenAI utilized bias probes to reveal that class design significantly influenced CLIP's performance and that the model could excel in niche tasks without task-specific training data. Their experiments uncovered biases related to race, age, and gender, emphasizing the need for thoughtful class design to mitigate un-

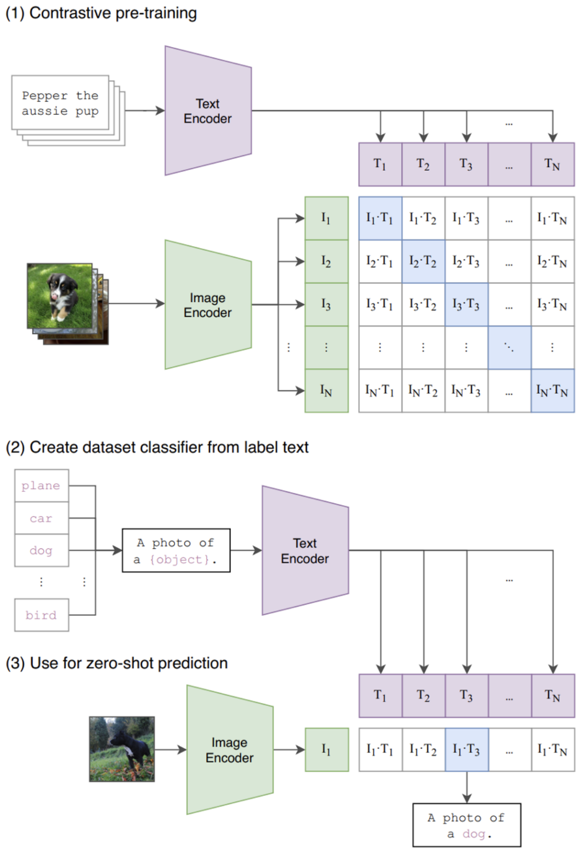
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

Figure 1: CLIP Architecture and Process

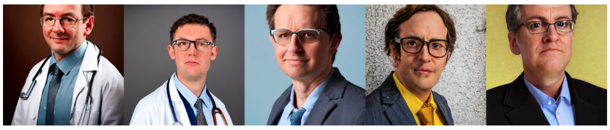

Figure 2: Stable Diffusion Model Results

wanted biases. The paper called for a community exploration to comprehensively characterize models like CLIP and develop evaluations that encompass their capabilities, biases, and overall impact on deployment.

Agarwal et al. [1] drew inspiration from the paper of Schwemmer et al. [10] for model evaluation. The paper used the photos of US members of Congress and Twitter images posted by these politicians to investigate systemic biases in image recognition systems, particularly in the classification of men and women. It reveals how these systems, such as Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Recognition, exhibit gender bias in labeling even uniform political images.

Mandal, Little and Leavy [6] employed the Word Embeddings Association Test (WEAT) from natural language processing to quantitatively measure and analyze stereotypical associations in CLIP's predictions related to character descriptions and occupations. The study reveals evidence of gender bias, showcasing how CLIP's unfiltered training data from the internet can lead to the propagation of societal biases. The authors not only assess bias within CLIP but also introduce a methodological framework for evaluating biases in large-scale multimodal models, emphasizing the importance of addressing harmful biases in applications ranging from social media to education.

3 Methodology

In general, our overall methodology can be split into three parts: data collection, model inference, and model evaluation.

3.1 Data Collection

During the data collection phase, we web-scraped images from Google Images. The images consisted of individuals with different races and genders across a list of professions. For each profession-race-gender combination, we scraped 5 images. From these 5 images, we chose 1 image that we thought represented the profession-race-gender combination the best. If none of the images are suitable, we Google the combination ourselves and handpick a good image with no watermark and no cartoon representation.

3.2 Model Inference

During the inference phase, each image in our scraped dataset represents an individual of a specific race and gender in a specific profession while each profession represents a text label. Using OpenAI's CLIP model, for each profession, we get a text embedding for the profession label and image embeddings for each image of an individual in that profession.

3.3 Model Evaluation

During the evaluation phase, we calculate the cosine similarity between the profession's text embedding and the embeddings of each image of an individual in that profession. Then for each profession, we can compare the similarity scores between the corresponding text and image embeddings. If the similarity scores between a profession and image are higher for some specific races and genders than others, we can conclude that there is bias present.

For example, if the text embedding for "CEO" has a higher cosine similarity with the image embedding of a male CEO than with the image embedding of a female CEO, we can conclude that there is a gender bias present.

4 Experimental Setup

4.1 Data

To web-scrape from Google Images, we used the python package google_images_search [2]. To use this package, we had to obtain a Google Search API key and create a Custom Search Engine [5]. Our web-scraped dataset contains images of individuals from the following professions, genders, and races as described below.

- Professions ∈ ["Doctor", "Teacher", "Scientist", "Nurse", "Police Officer", "CEO", "Construction Worker", "Lawyer", "Criminal", "Janitor", "Model", "Athlete"]
- Gender ∈ ["Female", "Male"]
- Race ∈ ["White", "Black", "Asian", "Latin American"]

We chose the list of professions based on our background knowledge of career-related stereotypes and we chose to include the most populous races in the United States as outlined by the United States Census [3]. This led to a final dataset of 96 images: 1 per profession-gender-race combination.

4.2 Model

For our model, we used the open-source version of OpenAI's CLIP model hosted on the HuggingFace Transformers library [7]. The pre-trained weights we used are from the "openai/clip-vit-base-patch32" version of the model. For each profession-image pairing, we used the pre-trained model's encode_text() and encode_image() methods to extract the text and image embeddings.

4.3 Metrics

To evaluate whether or not there is bias present in CLIP's text embedding method, we compare the CLIP score (cosine similarities) between the profession's text embedding and the image embedding across the different genders and races. The CLIP score between a text embedding vector ($\mathbf{A}$) and an image embedding vector ($\mathbf{B}$) is given in equation 1.

$$\text{score}(\mathbf{A}, \mathbf{B}) = 100\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \tag{1}$$

5 Preliminary Results and Analysis

Here is a baseline example of biases in the CLIP model. As we can see, the softmax of the CLIP score of the white male CEO is the highest, meaning that keywords in this combination are the most associated. In terms of gender, female CEOs have a much lower score than the male CEOs.

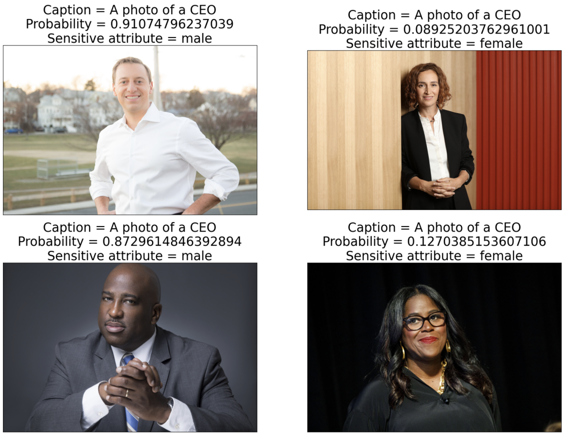
Caption = A photo of a CEO
Probability = 0.91074796237039
Sensitive attribute = male
Caption = A photo of a CEO
Probability = 0.08925203762961001
Sensitive attribute = female
Caption = A photo of a CEO
Probability = 0.8729614846392894
Sensitive attribute = male
Caption = A photo of a CEO
Probability = 0.1270385153607106
Sensitive attribute = female

Figure 3: Comparing image-text scores

We will continue to analyze the rest of the 96 images we collected in the near future.

6 Conclusion

To be continued...

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[2] Ivan Arar. Google images search. https://https://pypi.org/project/Google-Images-Search/, 2022.

[3] United States Census Bureau. Quick facts. https://www.census.gov/quickfacts/fact/table/US/PST045222, 2022.

[4] 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.

[5] Google. Programmable search engine. https://developers.google.com/custom-search/v1/overview, 2023.

[6] Abhishek Mandal, Suzanne Little, and Susan Leavy. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 416–424, 2023.

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

[8] 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.

[9] 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.

[10] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020.

[11] 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.