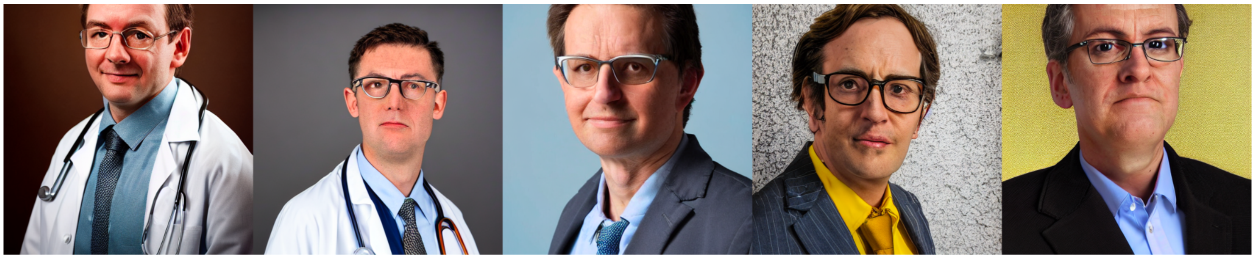
Evaluating Gender and Racial Bias in Text Embeddings of Multimodal Models: A CLIP case-study

Alison Yao, Luis Henrique Simplicio Ribeiro, Isha Vaish

Background

Evaluating gender and racial biases in word embeddings has been a hot topic in the Natural Language Processing (NLP) field. The effect of these biases is further exacerbated in multimodal 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 propose to do a deep dive into OpenAI's CLIP (Contrastive Language-Image Pretraining) model, a widely used model for image captioning and text-to-image generation applications. CLIP works by jointly learning visual concepts and natural language embeddings by using two separate encoders [1]. Images are converted into numerical embeddings using a vision encoder such as a convolutional neural network (CNN) while text descriptions are tokenized and converted into embeddings using a transformer-based language encoder. CLIP uses a contrastive learning objective, where it learns to bring corresponding image and text embeddings closer together into a shared embedding space. 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.

Stable Diffusion is a popular text-to-image generation model that uses CLIP to encode the image generation prompt [2]. Below we give an example, where we prompted the Stable Diffusion model to give us "a high resolution photo of a doctor". As you can see from the output below, all of the images that were produced are white males.



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 multimodal models encourages the development of fairer algorithms/methodologies and is essential in preventing unfair treatment based on gender or race [3].

Dataset

We are going to create a custom dataset by web-scraping the internet. Inspired by [4], we plan to Google images of occupations with gender and racial biases and scrape the first couple of images in the search results. This way, we can create a database of text and image pairs for evaluating the CLIP model. For example, a biased occupation could be a doctor and we would scrape images of a black doctor and a white doctor or a female doctor versus a male doctor.

Methodology & Measurement of Success

To identify racial and gender biases present in the embeddings learned by CLIP, we will measure how likely the model is to associate certain professions with certain groups of people. This would allow us, for instance, to identify whether text embeddings for a sentence containing the word doctor are more related to an image of a black or white person by calculating the cosine similarity of the text embedding with the embedding of the white doctor image and the black doctor image. This means for each profession, we can rank the images by their similarity scores. If the majority of images with the highest similarity scores belong to a certain race or gender, the model is associating that occupation/label with a certain group of people and is thus perpetuating pre-existing biases. We plan on using the HuggingFace open source version of CLIP. We don’t anticipate any resource issues.

References

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[2] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[3] Agarwal, Sandhini, et al. "Evaluating clip: towards characterization of broader capabilities and downstream implications." *arXiv preprint arXiv:2108.02818* (2021).

[4] Mandal, Abhishek, Suzanne Little, and Susan Leavy. "Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques." *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. 2023.