

The 'Racial Categorization' Challenge for CLIP-based Image Synthesis Systems

Originally published at <https://www.unite.ai/the-racial-categorization-challenge-for-clip-based-image-synthesis-systems/>



New research from the US finds that one of the popular computer vision models behind the much feted DALL-E series, as well as many other image generation and classification models, exhibits a provable tendency towards [hypodescent](#) – the race categorization rule (also known as the '[one drop rule](#)') which categorizes a person with even a small extent of 'mixed' (i.e. non-Caucasian) genetic lineage entirely into a 'minority' racial classification.

Since hypodescent has [characterized](#) some of the ugliest chapters in human history, the authors of the new paper suggest that such tendencies in computer vision research and implementation should receive greater attention, not least because the supporting framework in question, downloaded nearly a million times a month, could further disseminate and promulgate racial bias in downstream frameworks.

The architecture being studied in the new work is [Contrastive Language Image Pretraining](#) (CLIP), a multimodal machine learning model that learns semantic associations by training on image/caption pairs drawn from the internet – a semi-supervised approach that reduces the significant cost of labeling, but which is likely to reflect the bias of the people who created the captions.

From the paper:

'Our results provide evidence for hypodescent in the CLIP embedding space, a bias applied more strongly to images of women. Results further indicate that CLIP associates images with racial or ethnic labels based on deviation from White, with White as the default.'

The paper also find that an image's valence association (it's tendency to be associated with 'good' or 'bad' things, is notably higher for 'minority' racial labels than for Caucasian labels, and suggests that CLIP's biases reflect the US-centric corpus of literature (English language Wikipedia) on which the framework was trained.

Commenting on the implications of CLIP's apparent support of hypodescent, the authors state*:

'[Among] the first uses of CLIP was to train the zero-shot image generation model [DALL-E](#). A larger, non-public version of the CLIP architecture was used in the training of [DALL-E 2](#). Commensurate with the findings of the present research, the Risks and Limitations described in the DALL-E 2 model card [note](#) that it "produces images that tend to overrepresent people who are White-passing".'

'Such uses demonstrate the potential for the biases learned by CLIP to spread beyond the model's embedding space, as its features are used to guide the formation of semantics in other state-of-the-art AI models.'

'Moreover, due in part to the advances realized by CLIP and similar models for associating images and text in the zero-shot setting, multimodal architectures have been [described](#) as the foundation for the future of widely used internet applications, including search engines.'

'Our results indicate that additional attention to what such models learn from natural language supervision is warranted.'

The [paper](#) is titled *Evidence for Hypodescent in Visual Semantic AI*, and comes from three researchers at the University of Washington and Harvard University.

CLIP and Bad Influences

Though the researchers attest that their work is the first analysis of hypodescent in CLIP, prior works have demonstrated that the CLIP workflow, dependent as it is on largely unsupervised training from [under-curated](#) web-derived data, [under-represents women](#), can produce [offensive content](#), and can demonstrate [semantic bias](#) (such as anti-Muslim sentiment) in its image encoder.

The original paper that presented CLIP conceded that in a zero-shot setting, CLIP associates only 58.3% of people with the White racial label in the [FairFace](#) dataset. Observing that FairFace was labeled with possible bias by Amazon Mechanical Turk workers, the authors of the new paper state that ‘a substantial minority of people who are perceived by other humans as White are associated with a race other than White by CLIP.’

They continue:

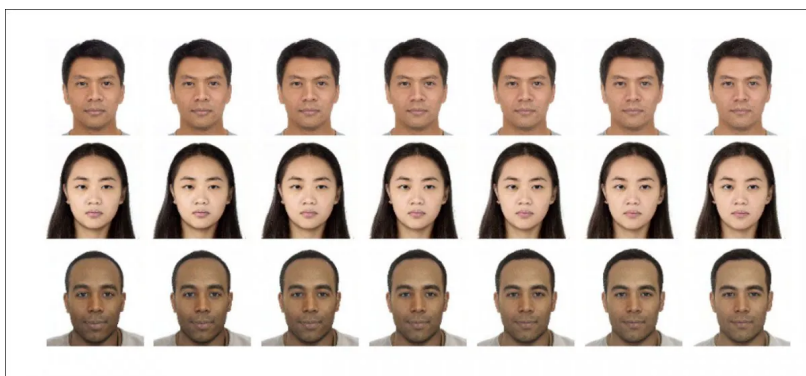
‘The inverse does not appear to be true, as individuals who are perceived to belong to other racial or ethnic labels in the FairFace dataset are associated with those labels by CLIP. This result suggests the possibility that CLIP has learned the rule of “hypodescent,” as described by social scientists: individuals with multiracial ancestry are more likely to be perceived and categorized as belonging to the minority or less advantaged parent group than to the equally legitimate majority or advantaged parent group.

‘In other words, the child of a Black and a White parent is perceived to be more Black than White; and the child of an Asian and a White parent is perceived to be more Asian than White.’

The paper has three central findings: that CLIP evidences hypodescent, by ‘herding’ people with multiracial identities into the minority contributing racial category that applies to them; that ‘White is the default race in CLIP’, and that competing races are defined by their ‘deviation’ from a White category; and that [valence bias](#) (an association with ‘bad’ concepts) correlates to the extent that the individual is categorized into a racial minority.

Method and Data

In order to determine the way that CLIP treats multiracial subjects, the researchers used a [previously-adopted](#) morphing technique to alter the race of images of individuals. The photos were taken from the [Chicago Face Database](#), a set developed for psychological studies involving race.



Examples from the racially-morphed CFD images featured in the new paper’s supplementary material. Source: <https://arxiv.org/pdf/2205.10764.pdf>

The researchers chose only ‘neutral expression’ images from the dataset, in order to remain consistent with the prior work. They used the Generative Adversarial Network [StyleGAN2-ADA](#) (trained on [FFHQ](#)) to accomplish race-changing of the facial images, and created interstitial images that demonstrate the progression from one race to another (see example images above).

Consistent with the previous work, the researchers morphed faces of people who self-identified as Black, Asian and Latino in the dataset into faces of those who labeled themselves as White. Nineteen intermediate stages are produced in the process. In total, 21,000 1024x1024px images were created for the project by this method.

The researchers then obtained a projected image embedding for CLIP for each of the total 21 images in each racial morph set. After this, they solicited a label for each image from CLIP: ‘multiracial’, ‘biracial’, ‘mixed race’, and ‘person’ (the final label omitting race).

The version of CLIP used was the [CLIP-ViT-Base-Patch32](#) implementation. The authors note that this model was downloaded over a million times in the month prior to writing up their research, and accounts for 98% of the downloads of any CLIP model from the [Transformers library](#).

Tests

To test for CLIP’s potential proclivity towards hypodescent, the researchers noted the race label assigned by CLIP to each image in the gradient of morphed images for each individual.

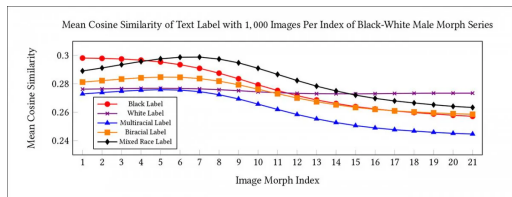
According to the findings, CLIP tends to group people in the ‘minority’ categories at around the 50% transition mark.

Minority Group Associations by Mixing Ratio					
Morph Series	75%	55%	50%	45%	25%
Asian-White Female	98.6	92.9	89.1	84.6	52.7
Black-White Female	97.6	79.7	69.7	58.2	18.6
Latina-White Female	92.2	81.5	75.8	71.4	50.3
Asian-White Male	84.4	55.3	47.2	40.1	11.1
Black-White Male	96.5	67.5	53.8	42.5	10.8
Latino-White Male	77.4	52.7	45.2	39.2	19.9

At a 50% mixing ratio, where the subject is equally origin/target race, CLIP associates a higher number of 1000 morphed female images with Asian (89.1%), Latina (75.8%) and Black (69.7%) labels than with an equivalent White label.

The results show that female subjects are more prone to hypodescent under CLIP than men, though the authors hypothesize that this may be because the web-derived and uncurated labels that characterize female images tend to emphasize the subject’s appearance more than in the case of men, and that this may have a skewing effect.

Hypodescent at a 50% racial transition was not observed for the Asian-White male or Latino-White male morph series, while CLIP assigned a higher cosine similarity to the Black label in 67.5% of the cases at a 55% mixing ratio.



The mean cosine similarity of Multiracial, Biracial and Mixed Race labels. The results indicate that CLIP operates a kind of 'watershed' categorization at varying percentages of racial mix, less often assigning such a racial mixture to White ('person', in the rationale of the experiments) than to the ethnicity that has been perceived in the image.

The ideal objective, according to the paper, is that CLIP would categorize the intermediate racial mixes accurately as 'mixed race', instead of defining a 'tipping point' at which the subject is so frequently consigned entirely to the non-White label.

To a certain extent, CLIP does assign the intermediate morph steps with Mixed Race (see graph above), but eventually demonstrates a mid-range preference to categorize subjects as their minority contributing race.

In terms of valence, the authors note CLIP's skewed judgement:

'[Mean] valence association (association with bad or unpleasant vs. with good or pleasant) varies with the mixing ratio over the Black-White male morph series, such that CLIP encodes associations with unpleasantness for the faces most similar to CFD volunteers who self-identify as Black.'

Correlation of Minority Group Label with Unpleasantness			
Female Series	Pearson's ρ	Male Series	Pearson's ρ
A-W Female	-0.22 (10^{-90})	A-W Male	-0.43 (10^{-90})
B-W Female	0.41 (10^{-90})	B-W Male	0.48 (10^{-90})
L-W Female	0.05 (10^{-13})	L-W Male	0.01 (.26)

The valence results – the tests show that minority groups are more associated with negative concepts in the image/pair architecture than for White-labeled subjects. The authors assert that the unpleasantness association of an image increases with the likelihood that the model associates the image with the Black label.

The paper states:

'The evidence indicates that the valence of an image correlates with racial [association]. More concretely, our results indicate that the more certain the model is that an image reflects a Black individual, the more associated with the unpleasant embedding space the image is.'

However, the results also indicate a negative correlation in the case of Asian faces. The authors suggest that this may be due to pass-through (via the web-sourced data) of positive US cultural perceptions of Asian people and communities. The authors state*:

'Observing a correlation between pleasantness and probability of the Asian text label may correspond to the "model minority" stereotype, wherein people of Asian ancestry are lauded for their upward mobility and assimilation into American culture, and even associated with "good behavior".'

Regarding the final objective, to examine whether White is the 'default identity' from CLIP's point of view, the results indicate an embedded polarity, suggesting that under this architecture, it is rather difficult to be 'a little white'.

Correlation (Pearson's ρ) of race or ethnicity label with "person" label (over 21,000 images)						
Racial or Ethnic Group	A-W Female	B-W Female	L-W Female	A-W Male	B-W Male	L-W Male
Source Racial or Ethnic Group	0.13	0.33	0.21	0.11	0.40	0.33
White	0.81	0.75	0.82	0.60	0.74	0.64

Cosine similarity across 21,000 images created for the tests.

The authors comment:

'The evidence indicates that CLIP encodes White as a default race. This is supported by the stronger correlations between White cosine similarities and person cosine similarities than for any other racial or ethnic group.'

*My conversion of the authors' inline citations to hyperlinks.

First published 24th May 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More