COS429 Final Report: Racial Bias in Image Captioning

Benjamin Liu, Evan Wang

May 15, 2024

Abstract

The prevalence of image captioning models in technology and society introduces racial biases, evident in various ways such as amplifying social biases and differing quality of captions and sentiments based on race, necessitating efforts to mitigate these biases for equality and reduced prejudice. An image-to-text model is susceptible to biases inherited from training data, prompting investigation into whether these models exacerbate racial biases rather than focusing solely on biases within existing datasets. Using a dataset of images annotated with skin colors, found that modern image-to-text machine learning models do not show bias for or against a shade of skin color. Modern captioning systems were feared to trend towards increasing bias between skin colors, but our experiments show promise that systems proposed from only two years ago no longer support this trend.

1 Introduction

Image captioning models are becoming increasingly prevalent in technology and society. Search and retrieval, image tagging, and improving accessibility are applications that incorporate automated image captioning, but are also plagued by racial bias [5]. These racial biases can be exhibited in multiple ways, such as amplifying existing social biases, different quality of captions depending on race, and different sentiments/diction depending on race [3] [8]. Mitigating these biases is crucial for protecting equality (of treatment and opportunity), reducing prejudice, and many other normative reasons [1]. We contribute to this effort by investigating racial biases in state-of-the-art image captioning models.

Image captioning using machine learning involves two main steps: feature extraction and sequence modeling. Feature extraction is typically done with a convolutional neural network (CNN) and sequence modeling is typically done using recurrent neural networks (RNN) or transformer-based architectures. Combining the two gives us a image-to-text model that can be trained to caption images. A machine learning model like this is susceptible to the quality of the data that it is trained on. Given bad quality data, such as data with biases in them, the model is likely to inherit the biases from the data and treat them as ground truths [5]. Our project focuses not on the bias in existing datasets, but rather if image captioning models exacerbate racial biases.

2 Related Work

The exploration of racial bias in automated systems is well-documented, with seminal works like Safiya Noble's "Algorithms of Oppression" highlighting how technological biases can perpetuate societal inequalities [5]. In the realm of image captioning, recent studies have pointed out how biases in training data can translate into biased algorithmic outcomes. For instance, Hirota et al. (2022) discuss how societal biases are amplified in image captioning systems, influencing the quality and sentiment of generated captions [3]. This thus provides motivation for our work.

Zhao et al. (2021) provide a critical analysis of racial biases within image captioning models by assessing the differences in captions generated for images of individuals with differing skin tones [8]. Their findings suggest significant disparities, prompting a need for more nuanced investigations into how different models handle racially diverse datasets. Specifically, Zhao et. al. show that modern image captioning systems based around neural networks and transformers have biased vocabulary choice between light and dark skin tones. We build off of this paper by verifying their methods, and investigating the robustness of their

In terms of models, the development of advanced neural network architectures like ResNet and various transformer-based models has significantly impacted feature extraction and sequence modeling in image captioning [2, 4, 7]. These models have been central to recent efforts to mitigate bias by improving the diversity and representation within training datasets.

In addition to neural network advancements, non-deep learning feature extraction methods like Histogram of Gradients (HoG) continue to be relevant for their ability to capture essential image characteristics without the computational complexity of deep learning models. These

techniques are crucial for developing robust systems that are less susceptible to the variances in data quality.

Our work builds upon these foundations by employing novel methodologies to evaluate bias more rigorously. By creating and analyzing 'similar' image pairs that differ primarily in skin tone, we aim to uncover subtle biases that may not be evident through conventional analysis methods. This approach is inspired by but distinct from Zhao et al., as we incorporate additional feature extraction techniques and employ modern image-to-text models to assess bias across different dimensions.

3 Methods

3.1 Preprocessing: Image Feature Extraction

We follow certain steps in order to pre-process our data for bias analysis. Our intermediary goal is to come up with image pairs that are very similar to each other except for skin tone, since these images should ideally have similar captions. A notable difference in captions signals racial bias, as the captions would be different due to skin tone differences.

To begin, we work off of Zhao et. al.'s annotations, which provide a labelling of dark and light skin tone for the COCO 2014 validation dataset [8]. We refer to images with darker-toned people as dark images and images with lighter-toned people as light images, as defined by Zhao et. al. To start off, we mask the people in the relevant images in the COCO 2014 validation data. Zhao et. al. mention this step but do not go into details, and their repository lacks the relevant scripts, so we implement these ourselves in the `mask.py` file. COCO 2014 has instance segmentation data, so we go through the people segmentation category and use the Python Imaging Library to mask the humans, replacing their pixel value by the average color of the rest of the image. This creates our masked dataset.

Fig. 1 and Fig. 2 show an example of this process, where the human is replaced by a mask that removes skin tone from the image.

In order to find "similar" pairs of light and dark images for comparison, we need a metric of comparison, which means we need to extract image features.

We then take two approaches for extracting features for these images. First, we follow Zhao et. al. and extract the last linear layer of ResNet 50 [2] for each image. Here, this code was not

Figure 1: Original image.

Figure 2: Masked image.

present in their repository, so we implement this ourselves in `extract_resnet_features.py`. We follow this approach as ResNet represents modern state-of-the-art feature extraction techniques in computer vision, as it is a deep convolutional neural network that can encode a wide variety of features, from local textures to overall spatial structure.

Second, we experiment with extracting features via Histogram of Gradients (HoG), as a non-deep-learning based approach. HoG is able to capture edge and gradient information, making it able to detect overall shapes and contours in images, while being weak to scale, rotation, noise, illumination, etc. We hypothesize that if overall spatial structure is important for featurizing these dark and light images, and scale and rotation are not as important, then perhaps HoG features will produce meaningful pairings.

We chose to not use Scale-Invariant Feature Transform (SIFT) features for this specific project, as SIFT does not return a singular vector that can be easily compared between images (as ResNet and HoG do, which we can easily modularize our code for). We recognize that an extension could be made to experiment with SIFT descriptors.

We features for both light and dark images, we then pair the images. We follow Zhao et. al.'s approach of using the Gale Shapley algorithm to make a stable matching between light and dark images. We follow this approach because an alternative would be to have groups or clusters of images instead of pairs. However, this would require many decisions in terms of how to create these clusters and groups, and define some threshold similarity for each group. Grouping may lead to tricky slippery slopes where two images pairwise could be very similar, but the most dissimilar images in the group could be very far apart. So, we stick with the stable matching

outcome to pair light and dark images. Finally, to construct our comparison dataset, we filter the **top 20 percent most similar image pairs**, as we do not want to include pairs that are still very far apart, which will naturally exist. This leaves us with 220 pairs of images.

We take two approaches to measure the similarity between feature vectors: cosine similarity and Euclidean distance. We experiment with both, but recognize that to have thorough evaluation between the two choices would require manual labor through a system like Mechanical Turks. We do manually inspect some pairings for a limited analysis.

3.2 Captioning

We employed two different image-to-text models, GIT (Generative Image-to-text Transformer) and BLIP (Bootstrapping Language-Image Pre-training), to generate image captions for each pair of images. We also used VADER (Valence Aware Dictionary and sEntiment Reasoner) for sentiment analysis on the obtained captions. Note that we produce captions on the non-masked images, as we want to introduce skin tone back into the model.

3.2.1 GIT

GIT is a modern image-to-text transformer proposed in 2022. It uses CLIP's [6] vision encoder to condition the model on vision inputs. It is said to perform extremely well on the image captioning task. It is trained on a large dataset of over 1 billion images and has 0.7 billion parameters [7]. We use the huggingface transformers version of GIT, specifically GITForCausalLM. We choose GIT as a relative modern state-of-the-art captioning model, released in 2022. Thus, it represents recent transformer-based computer vision models, and we can verify/extend some of previous papers' work on this more recent model. The architecture for GIT is shown in Fig. 3.

3.2.2 BLIP

BLIP is another modern image-to-text transformer that was also proposed in 2022. It uses noisy web data by bootstrapping captions and is able to achieve state-of-the-art results with image captioning [4]. We use the huggingface transformers version of BLIP, specifically BlipForConditionalGeneration. We chose BLIP to test different model architectures. GIT only consists of one image encoder and one text decoder [7] while BLIP uses more traditional architecture, involving complex multi-modal encoder/decoder structures. In doing so, we test the effect of

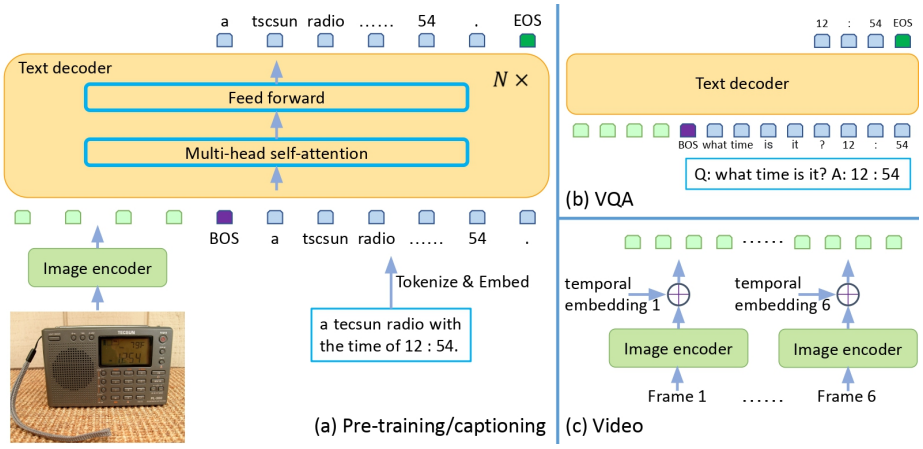


Figure 3: GIT Architecture

the network architecture on bias, seeing if the simplicity of GIT's architecture is the potentail solution to bias. BLIP's architecture is shown in Fig. 4

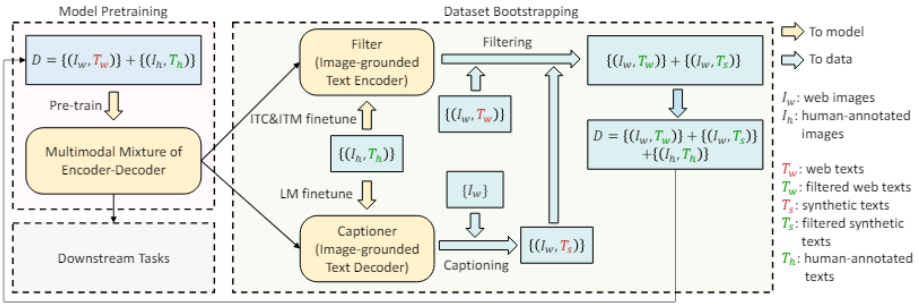


Figure 4: BLIP Architecture

3.3 Caption Analysis

Implemented in `notebooks/sentiment_analysis.ipynb` and `notebooks/vocab_analysis.ipynb`

3.3.1 VADER Sentiment Analysis

VADER returns a sentiment score from -1, indicating strongly negative, to 1, indicating strongly positive. We use VADER as it is trained on sentiments expressed in social media, which is fitting for image captions and short texts. It is also a lightweight and fast system that does not require heavy computational resources.

3.3.2 Hugging Face Pipeline Sentiment Analysis

We also use Hugging Face's pipeline for sentiment analysis, using BERT transformer models that have proven high accuracy and powerful contextual understanding. We contrast this pipeline's non-rule-based approach to VADER's rule-based approach. This pipeline's output is a discrete choice of a "POSITIVE" or "NEGATIVE" label, so analysis takes on a discrete rather than continuous form, as with VADER.

3.4 Word Choice Analysis

Beyond sentiment analysis, we are also interested in general vocabulary bias between dark and light image captions. To do so, we follow and implement Zhao et. al.'s approach of taking a logistic regression model with the top 100 (non-trivial) words as features, and measuring the AUC of the model. If the word choice is truly neutral, the model should predict with random choice and have an AUC of 0.5. Higher predictive power indicates bias.

4 Results

4.1 Image Pairing

To perform a limited evaluation of our image feature extraction and pairing methods, we manually inspect the top 20 most similar pairings for each approach. We look for high quality pairings, example in Fig. 8 and low quality pairings, example in Fig. 9. Low quality pairs entail pairings that do not look similar, and/or seem to have a relevant skin tone difference.

Our results are:

Feature Type	**Similarity Measurement**	**Number of high-quality pairs out of 20**
ResNet	Cosine	10
ResNet	Distance	8
HoG	Cosine	2
HoG	Distance	2

Table 1: Comparison of feature types and similarity measurements

ResNet features do clearly create higher quality pairings, but there is no significance difference between cosine similarity and Euclidean distance as the similarity metric. HoG did not

create very good pairings, as seen in Fig. 5, where the images are not very similar in content, but was in HoG's top 10 most similar pairings.

We visualize the HoG features in Figures 6 and 7, and see that HoG features are not very descriptive or useful for our more complex and nuanced images.

Figure 5: Top Similar Image Pairing via HoG Features

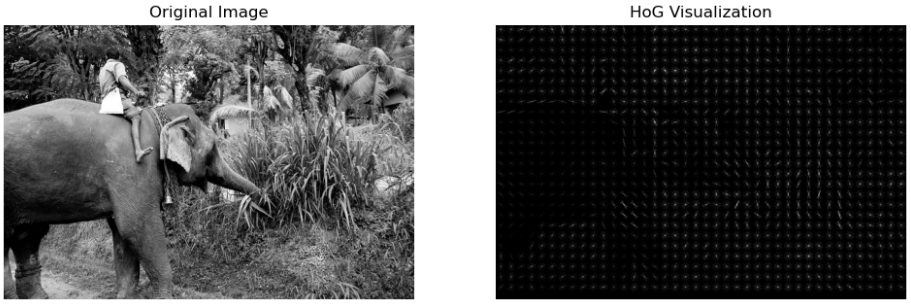


Figure 6: Top Similar Image Pairing via HoG Features

Thus, for simplicity and effectiveness, going forward, we use pairings based on ResNet features and cosine similarity.

4.2 Quantitative Semantic Results

Using GIT and BLIP, we generated captions for pairs of similar images with differing skin tones and evaluated the semantic differences between the two captions.

For a baseline, we evaluate the differences for the human COCO captions. Human captions caption light images with a mean VADER score of 0.0527 ± 0.01, and dark images with a mean

Figure 7: Top Similar Image Pairing via HoG Features

Figure 8: High quality image pairing generated by ResNet features, cosine similarity. Images have extremely similar content, with one dark-skinned person and one light-skinned person.

VADER score of 0.034 ± 0.01. So we do see some potential bias here, but the p-value of this difference is 0.22, not statistically significant. For the Hugging Face pipeline sentiment analysis, we use chi-squared since the outputs are discrete, and get an insignificant p-value of 0.63. We note that this disagrees with Zhao et. al. who find statistically significant difference in human captions. This may be due to different specific implementations of image pairing, as well as our choosing the top 20 percentile instead of their 40 percentile.

GIT sentiment for dark images: 0.0466 ± 0.006, light images: 0.0427 ± 0.006, p-value of 0.67. No statistical significance.

Figure 9: Low quality image pairing generated by ResNet features, cosine similarity. Images do have similar content but people are not central enough to have relevant skin tone differences.

BLIP sentiment for dark images: 0.057 ± 0.007, light images: 0.042 ± 0.006, p-value of 0.06. Interestingly, the average sentiment of BLIP was more positive for dark images than light images, although still not statistically significance at an alpha level of 0.05.

Our results for GIT can be seen in Fig. 10. We also used BLIP to generate captions on the same pairs of images to compare the performance between the two models. The results for BLIP can be seen in Fig. 11.

The graphs show a significant portion of the images having completely neutral captions for both BLIP and GIT. Some captions leaned positively and only a few captions leaned slightly negative. The positively and negatively leaning captions are also evenly distributed between the light and dark skin toned images. This shows an overall neutral captioning system, with little bias between light and dark skin tones.

4.3 Quantitative Vocabulary Results

As a baseline, for human captions, our model achieves an AUC of 0.47 for classifying between light and dark image captions. Interestingly, this contradicts Zhao et. al.'s findings, which may be due to our strictier threshold of the top 20 percentile of image pairs.

Surprisingly, we find similar lack of predictive power for our image captioning models. GIT captions have an AUC of around 0.47 as well, and BLIP captions have an AUC of around 0.40. These are certainly surprising, and goes against Zhao et. al.'s findings, where their automated

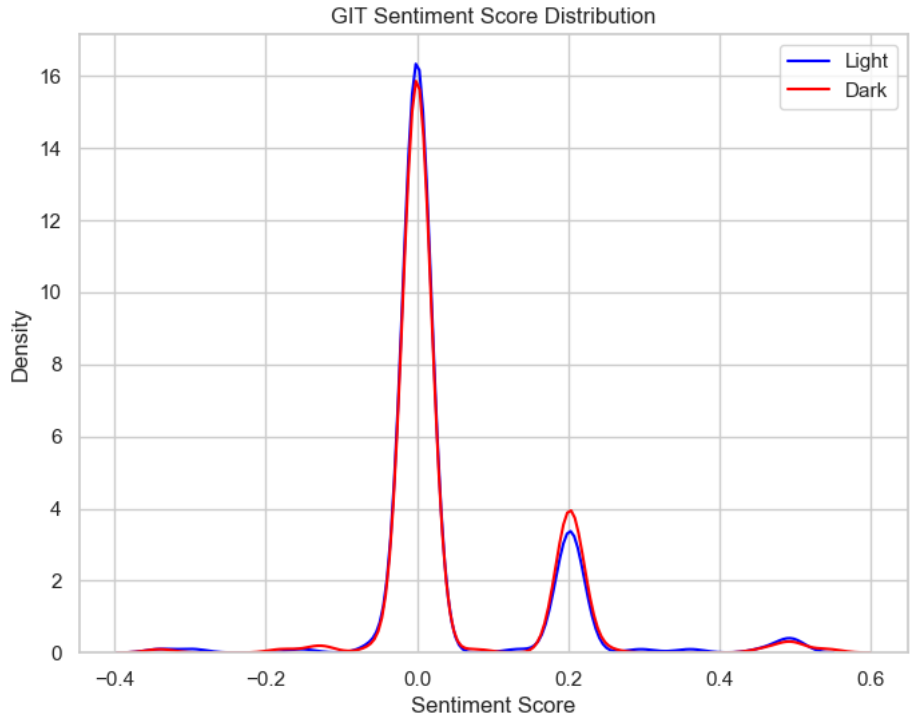


Figure 10: Frequency of sentiment score using VADER evaluation between the captions of paired images obtained via GIT. The majority of images had a sentiment score of 0, and there is very little difference between the distribution for light skin toned images and the distribution for dark skin toned images.

captioning models exacerbated the bias here (i.e. automated captions increased AUC). More implications will be examined in the discussion section.

4.4 Qualitative Results

We see some examples of pairs of images and their generated captions in Fig. 12. These images show successful pairing of images for light and dark skin tones. We also see that the captions for all four images are neutral. There is no negative language used for the dark skin toned images. This pattern was found in all of our image pairs and their captions. Looking deeper into the generated captions, we also found that all the captions avoided race or skin color, only mentioning color when describing backgrounds or clothes, never people.

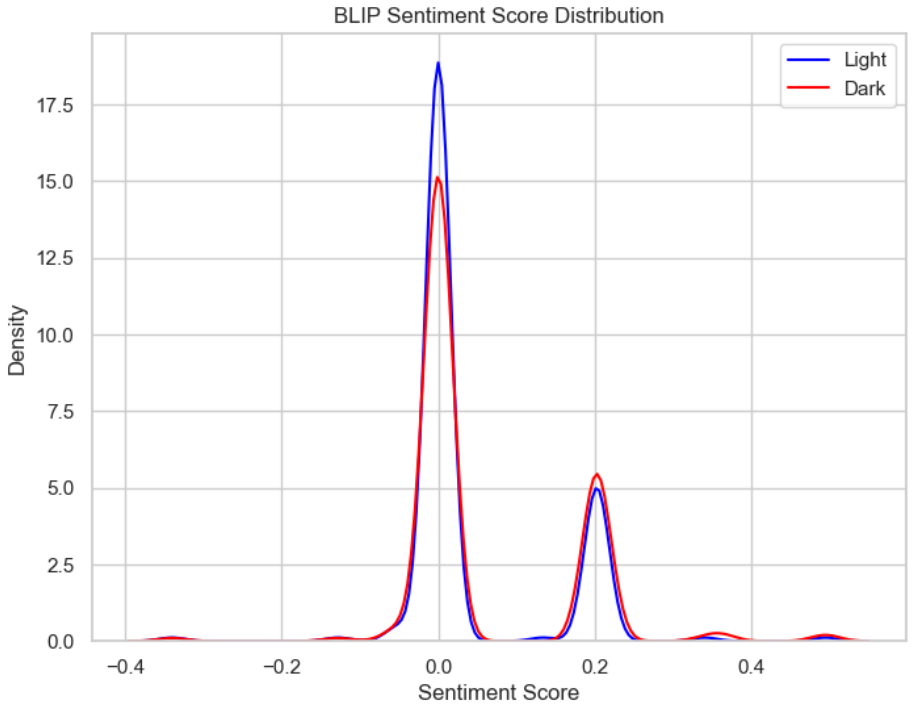


Figure 11: Frequency of sentiment score using VADER evaluation between the captions of paired images obtained via BLIP. The distribution for BLIP captions is very similar to that of GIT captions, with little to no statistical significance in the difference between light skin toned images and dark skin toned images.

5 Discussion

From our results, we can see that there is minimal difference in sentiment and vocabulary when comparing similar pictures with different skin tones. These results show an improvement over previous models that did struggle with racial biases. Because we tested on a modern image captioning model, we show an improvement over time of biases in image captioning models. This is likely due to an improvement in data quality over time. This is because past models struggled with racial bias due to biased datasets [5]. The lack of racially charged terms is a direct improvement over previous models, which were found to have offensive language and racial descriptors [8].

Comparing the performance between GIT and BLIP, we see that both models produce captions that stay relatively neutral in sentiment. There is no negative racial bias against darker skin toned images that we would expect from a biased image captioning model. This is

expected for a modern model like GIT and BLIP, as the availability of data for image-to-text models is much less of a barrier now. GIT, being trained on over 1 billion images, acquires its data from COCO, Conceptual Captions, SMU, Visual Genomem, ALT200M, and more [7]. The diversity in the data used and the recency of the data (some being from 2021) alleviates the racial bias issues that we used to see. In addition, particularly for sentiment analyses, it makes sense that modern image captioning models have been trained/fine-tuned to avoid any harsh polarizing sentiments in their captions, which is why we observe very neutral captions generated all around.

Moreover, addressing the lack of reproducibility from Zhao et. al. We think that the percentile of pairs chosen matters significantly in terms of image pairing quality. Even among the top 20 most similar pairings, we observe some bad quality pairings. Moreover, there seems to be a skin-tone annotation issue. For instance, ResNet features created the following pair in Fig. 13. However, this is evidently the same person and scene, so they should not have been labelled as separate skin tones by the skin tone annotation dataset from Zhao et. al. Annotation issues like these may lead to underestimation of bias in our captioning systems, and may explain the under 0.5 AUC we saw from the logistic regression vocabulary models.

Moreover, by expanding the percentile of similar pairings even more, we would allow for lower and lower quality of pairings, that don't genuinely reflect the idea of picking images that differ only in skin tone, thus biasing results in that way as well.

5.1 Limitations

Due to limited time and computational resources, we were only able to test one modern captioning system. When attempting other captioning models, such as Llava and BLIP-2, we ran into complications, such as running out of memory and other errors, that we were unable to resolve in time. Given more time and resources, we would like to test these other modern image captioning models to see if any implicit racial bias exists within them.

Crucially, to combat the annotation issues, and create genuine high quality dark-light image pairings, we would require a lot of manual inspection and labor. This is perhaps one important direction for further research to take, but was not feasible for our analysis.

5.2 Conclusions

Image-to-text models for image captioning have struggled to avoid racial biases in the past, but modern models today have largely alleviated that issue using high quality data. Our findings show that imaging captioning systems tend to create neutral captions, with little to no favoritism towards images with light skin toned subjects. This is a result of a large movement towards de-biasing the data we use to train our machine learning models, showing successful progress and a promising future for the future of machine learning data quality.

6 Resource Availability

Our code for this project can be found at `https://github.com/benliu961/COS429`

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning, 2022.

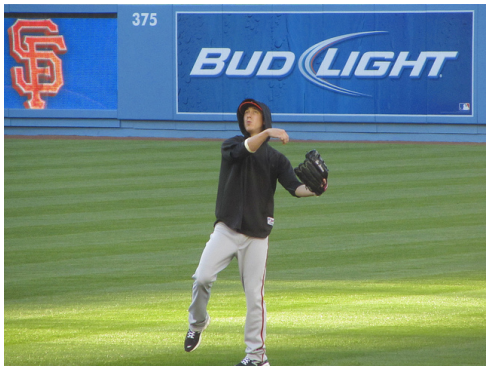
[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[5] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[6] 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.

[7] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022.

[8] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning, 2021.



(a) GIT: a man in a baseball uniform is throwing a baseball.
BLIP: a young boy in a baseball uniform throwing a baseball

(b) GIT: a man walking on a baseball field with a little boy.
BLIP: a baseball player walking across a field with a little boy

(c) GIT: a woman bent over on a tennis court.
BLIP: a woman bending her knees on a tennis court

(d) GIT: a woman in a pink dress holding a tennis racket.
BLIP: a woman holding a tennis rade

Figure 12: Examples of similar image pairs and their generated captions with and BLIP. The left images are the light skin toned images and the right images are the dark skin toned images.

Figure 13: Image Pair according to ResNet features. But this is clearly the same person-doesn't make sense that they were labelled as both light and dark skin tone by the dataset.