Counterfactual Matching for Fair Multimodal Deep Learning in Healthcare
Final Report

{carriech, guangzel, ryanwang, cheny357}@mit.edu

1 Abstract

Multimodal models are becoming increasingly adopted in healthcare to facilitate the diagnostics and decision-making for patients. However, one major aspect of applying deep learning models is investigating their reliability, where bias arising from different demographic groups could lead to the issue of unfairness. This study introduces a novel bias correction methodology that emphasizes counterfactual matching to address these concerns. Applied to the MIMIC-IV and MIMIC-CXR datasets, our method involved quantifying bias in existing Multimodal models and then implementing a causal de-biasing algorithm. Our results show notable improvements in model performance, particularly in reducing disparity across diverse racial groups in the context of lung disease diagnosis.

2 Introduction

The advent of machine learning in healthcare has heralded significant advancements in diagnostics and patient care, enabling providers to leverage vast amounts of multimodal data for informed decision-making. However, bias in artificial intelligence, particularly in sensitive areas such as healthcare, can have profound implications, including perpetuating disparities and affecting the quality of patient care.

A major issue with the application of deep learning in healthcare is *fairness*. Unfairness in language models could also arise from implicit bias coming from different demographic groups (2). Even in the most commonly used medical datasets (i.e. MIMIC), intervention prediction becomes biased by race and marital status (6). This is phenomenon is also observed in representation learning, where Electronic Health Record (EHR) textual representations exhibit bias due to the nature of different gender pronouns in the language (8).

In this project, we focus on the fairness of language models and address and quantify the bias present in multimodal healthcare models, with a focus on those integrating radiological images, textual reports, and demographic data. Our approach involves two strategies: quantifying the extent of bias in existing models and applying a novel causal de-biasing algorithm. This algorithm seeks to neutralize confounders such as demographic indicators to avoid skewed predictions. We investigate bias exhibited in both representation learning and standard classifcation through contrastive learning and modality fusion methods. We challenge the models with counterfactual scenarios, adjusting for demographic probability weights to better reflect minority classes. The efficacy of these interventions is rigorously tested by evaluating model performance, ensuring that our de-biasing efforts do not compromise the accuracy and reliability of the healthcare predictions.

3 Related Work

3.1 Causal Inference in Natural Language Processing

The issue of unfairness usually arises from spurious correlations. The concept of spurious correlation originated from the causal inference framework, where causal methods investigate causal relationships that go beyond correlation.

The causal framework captures the causal effect of a particular intervention on an outcome. The idea of spurious correlation in causal inference is represented by confounding, where confounding in causal inference arises when a particular feature is correlated to both the treatment and the outcome. Similarly, spurious correlations in the learned word representations could introduce the possibility that downstream tasks learn these spurious correlations, ignoring true correlations, leading to untrustworthiness and unfairness of the language model. In recent literature, there are emerging work and discussions on de-biasing methodologies motivated by the causal framework (eg., (1; 2; 13)). In this project, we primarily focus on the matching methodology and showcase its application to language models as a bias correction technique.

3.2 Bias Correction Methodology

To quantify counterfactual fairness, one can observe whether the model produces the same output for each individual under observation and counterfactual (2). Post-processing methods such as feature clipping de-correlate learned representations, by removing confounding dimensions (13). In addition, scoring functions help understand the effect size of different attributes on the target, which quantifies the magnitude of bias (4).

In particular, under the causal setting, we remove all possible association paths from treatment to the outcome counterfactual so that the remaining path only comes from the treatment itself. In our setting, we try to eliminate any "confounders" so that difference in our final predictions comes from different meanings of the text, but not from different gender pronouns, for example. In the next section, we describe our proposed de-biasing methodology and its implementation details.

4 Methods

4.1 MIMIC Dataset

We use MIMIC-CXR and MIMIC-IV data. MIMIC-IV (Medical Information Mart for Intensive Care IV) is a large, publicly available EHR dataset developed by the MIT Laboratory for Computational Physiology, in collaboration with the Beth Israel Deaconess Medical Center. The dataset contains comprehensive, de-identified health data from over 299,000 critical care patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019. MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-Ray) Database is a dataset of chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center.

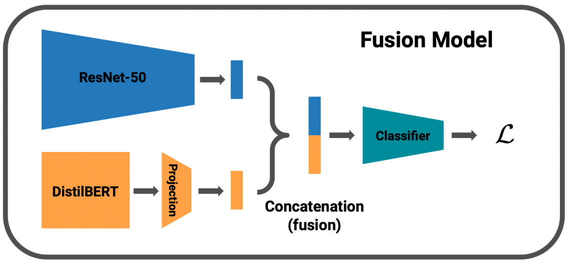

Figure 1: Multimodal fusion pipeline used in standard training. Representations from the image and textual encoder are concatenated together to be used for downstream analysis.

We select a subgroup with interested columns of IDs, view position, label, images, and final reports from MIMIC-CXR data and join it with MIMIC-IV EHR data on the study and subject IDs to get demographic information (*gender*, *marital status*, *age*, and *race*). We grouped all detailed race types into general 5 types of *Black*, *Asian*, *Hispanic*, *White* and *Others*. We use a 5000 patient subset containing demographic information and pleural effusion of the lungs as our target due to sample size. We evaluate our models on a disjoint subset of 2000 sampled patients.

4.2 Multimodal Methods

Our first goal is quantifying the bias in various multimodal methods. We aim to examine a simple modality fusion strategy (**Figure 1**). Let $\mathcal{D} = \{(v_{im}^{(i)}, v_{text}^{(i)}, Z_{dem}^{(i)} = z_{dem}^{(i)})\}_{i=1}^{N}$, the radiology image, radiology report, and demographic data for individual i, respectively.

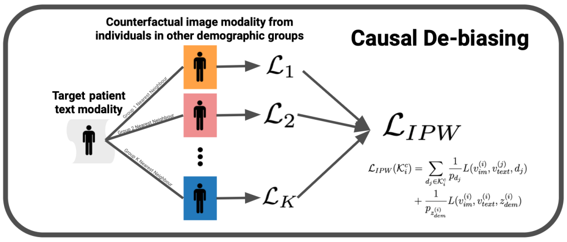
Counterfactual image modality from individuals in other demographic groups
Causal De-biasing
Target patient text modality
Group 1 Nearest Neighbour
Group 2 Nearest Neighbour
Group K Nearest Neighbour
$\mathcal{L}_1$
$\mathcal{L}_2$
$\mathcal{L}_K$
$\mathcal{L}_{IPW}$
$\mathcal{L}_{IPW}(\mathcal{K}_i^c) = \sum_{d_j \in \mathcal{K}_i^c} \frac{1}{p_{d_j}} L(v_{im}^{(i)}, v_{text}^{(j)}, d_j) + \frac{1}{p_{z_{dem}^{(i)}}} L(v_{im}^{(i)}, v_{text}^{(i)}, z_{dem}^{(i)})$

Figure 2: Visual representation of Algorithm 1. Target patients are matched with counterfactual individuals from different demographics using Word2Vec representations, marital status, age, and gender. Losses are calulated on a demograhic basis, and aggregated together using inverse probability weighting.

$Z_{dem}^{(i)}$ denotes the demographic group indicator for each individual i, where $z_{dem}^{(i)} \in \mathcal{K} = \{d_1, \cdots, d_k\}$ be all possible demographic groups and k is finite. Let $\mathcal{D}_{dem}^{(d_j)}$ be the corresponding subsets of $\mathcal{D}$ with respect to the demographic group d_j. Let $\mathcal{D}_{im}, \mathcal{D}_{text}, \mathcal{D}_{dem} := \cup_j \mathcal{D}_{dem}^{(d_j)}$ be subsets of $\mathcal{D}$ that only contain the respective modalities $v_{im}, v_{text}, Z_{dem}$.

We will use a ResNet-50 (3) for encoding an image representation v_{im}, and a DistilBERT transformer (12; 10) to encode a text representation v_{text}.We will perform fusion modeling by concatenating v_{text} and v_{im} giving $v_{fusion} = [v_{text}||v_{im}]$, and using v_{fusion} as input to a classifier head.

4.3 Causal De-Biasing Algorithm

For each individual i, we first define the *counterfactual* scenario for each of the remaining $k - 1$ demographic groups as $v_*^{(j)}$, the corresponding counterfactual outcome in demographic group $d_j \in \mathcal{K}_i^c = \{d_1, \cdots, d_k\} \setminus \{z_{dem}^{(i)}\}$ (**Figure 2**). For example, for each individual i, the counterfactual individual in the demographic group d_j is found using a matching algorithm, finding the nearest neighbor $v_{text}^{(j)}$ to $v_{text}^{(i)}$ in $\mathcal{C}_{text}^{i,j} = \{v_{text}^{(j)} \in \mathcal{D}_{text}$ such that $z_{dem}^{(j)} \in \mathcal{D}_{dem}^{(d_j)}\}$. In order to improve the identification of a counterfactual, we use Word2Vec (7) as a pre-processing step on the reports, such that we incorporate the textual similarities in addition to the demographic information (marital status, age, and gender) between individuals in this matching step.

Let $L(v_{im}, v_{text})$ be a loss function that takes in the multimodal representations. We choose L to be Cross Entropy Loss where the first step generates v_{fusion}. In practice, L can be anything taking in multiple modalities, for example CLIP loss (9). We use inverse probability weighting with $p_{z_{dem}^{(i)}} = P(Z_{dem}^{(i)} = z_{dem}^{(i)})$, the probability of the individual i being in a given demographic $z_{dem}^{(i)}$, giving minority classes more weight.

$$\mathcal{L}_{IPW}(\mathcal{K}_i^c) = \sum_{d_j \in \mathcal{K}_i^c} \frac{1}{p_{d_j}} L(v_{im}^{(i)}, v_{text}^{(j)}, d_j) + \frac{1}{p_{z_{dem}^{(i)}}} L(v_{im}^{(i)}, v_{text}^{(i)}, z_{dem}^{(i)})$$

Algorithm 1: Causal De-Biasing

```
$W \leftarrow$ Word2Vec($\mathcal{D}_{text}$)
$W_{match\_idx} \leftarrow$ Matrix($N, k$)
for $i$ in $1 : N$ do
    for $j$ in $1 : k$ do
        $W_{match\_idx}[i, j] \leftarrow$ Match($\mathcal{D}_{text}[i], \mathcal{C}_{text}^{i,j}$)
    end for
end for
$f_\theta \leftarrow$init multimodal model
for $epoch$ in $1 : N_{epoch}$ do
    Train($f_\theta, W_{match\_idx}$)
end for
```

Algorithm 2: Train Epoch

```
$\mathcal{J} \leftarrow 0$
for $i$ in $1 : N$ do
    $\mathcal{K}_i^c = W_{match\_idx}[i, :]$
    $\mathcal{J} \leftarrow L_{total} + \mathcal{L}_{IPW}(v_{im}^{(i)}, v_{text}^{(i)}, \mathcal{K}_i^c)$
end for
$\mathcal{J} = \frac{1}{N}\mathcal{J}$
$\theta \leftarrow \theta - \eta\nabla\mathcal{J}$
```

4.4 Model Evaluation

We will primarily evaluate our models by investigating the classification performance of each model using accuracy and AUC. Accuracy will be used for an overarching look at the data (including true negatives), while AUC will be used to judge performance given imbalanced data settings. Rather than focusing on overall performance, we will look at performance stratified on the demographic data that we de-bias on (i.e. racial data).

4.5 Interpretability

We investigate whether de-biasing affects model interpretability through both textual and imaging modalities. We investigate textual interpretability through attention scores of the last layer of the DistilBERT encoder. We investigate imaging interpretability through saliency maps, taking gradients through the multimodal modal back to the input (11).

4.6 Implementation Details

We train our 90 million parameter fusion model on 16GB Nvidia V100 GPUs available on Google Colab. We use a batch size of 6 due to compute limitations. We use AdamW (5) to optimize our model with a learning rate of 1×10^{-5} for 15 epochs and pick the model with the best validation loss. We use the standard L_2 norm as the distance measure for the causal de-biasing algorithm mentioned in Algorithm 1. For the base model, we will use ResNet-50 from `torchvision`, and distillBERT from `transformers`. We will also use `sklearn` for defining train, validation, and test splits (70%/15%/15%). We use `torch` for all other processing.

Code is publically available on GitHub https://github.com/rjunw/causal-mm-bias/.

5 Discussion of Results

5.1 Overall Performance of Causal De-biased Model

In this section, we discuss performance of the proposed causal de-biased fusion model on MIMIC-IV data and compare it with the standard fusion model. In particular, we focus on the stratified performance of both models for each demographic group.

In **Table 1**, we show the fusion AUC score of the de-biased fusion model and standard fusion model for each demographic group. In general, when training sample size increases, AUC for both de-biased fusion model and standard fusion model increases. If we consider marginalizing across demographic groups, the performance of the de-biased fusion model is comparable to that of the standard fusion model for different training sample sizes.

Within each demographic group, the de-biased fusion model performs mostly at least as well as the standard fusion model except for Asian group with a training sample size of $N_{\text{train}} = 700$, across different training sample sizes. With less data, it may be hard to find a representative counterfactual, which may actually hurt performance, rather than improve. However, when $N_{\text{train}} = 3500$, the AUC of the de-biased fusion model is higher than that of the standard fusion model in all demographic groups. Overall, the performance improvement of the de-biased fusion model mostly occurs in rarer demographic groups such as Hispanic and Other Races. For example, for Hispanic group, the causal de-biased model yields a fusion AUC of 0.988, which is larger than the standard fusion model, for $N_{\text{train}} = 1400$ and $N_{\text{train}} = 3500$. This performance improvement potentially comes from incorporating the information of the counterfactual through the matching algorithm described in section 4.3 and utilizing a weighted loss with a larger weight assigned to minority classes.

We do notice that most performance improvements are small. This is likely due to the fact that the performance of the standard fusion model already achieves a high AUC. In the case where there is a slightly larger performance disparity such as when $N_{\text{train}} = 3500$, our proposed causal de-biased fusion model is able to correct the bias and bridge the disparity in each demographic group. Therefore, our results show that our proposed causal de-biasing algorithm has the potential to reduce bias.

5.2 Variability in Causal De-Biased Models Over Sample Sizes

Furthermore, in **Table 2**, we show the standard deviation of all fusion AUC scores across different sample sizes in each demographic group for both the de-biased fusion model and standard fusion model. The standard deviation is calculated on the AUC scores across different training sample sizes to show the fluctuation of performance across different training sample sizes. The overall standard deviation (i.e., marginalizing across demographic groups) of

Model	Black	Asian	Hispanic	White	Other	Overall
Normal $N_{\text{train}} = 700$	0.952	0.970	0.978	0.965	0.994	0.965
De-biased $N_{\text{train}} = 700$	0.952	0.928	0.978	0.957	**1.000**	0.959
Normal $N_{\text{train}} = 1400$	0.964	0.990	0.968	0.964	0.994	0.967
De-biased $N_{\text{train}} = 1400$	0.961	**1.000**	**0.988**	**0.967**	**1.000**	**0.970**
Normal $N_{\text{train}} = 2100$	0.966	0.959	0.988	0.972	1.000	0.973
De-biased $N_{\text{train}} = 2100$	0.972	**0.990**	0.988	0.973	1.000	**0.976**
Normal $N_{\text{train}} = 2800$	0.970	0.970	0.978	0.979	1.000	0.978
De-biased $N_{\text{train}} = 2800$	0.973	**1.000**	0.978	0.979	1.000	**0.980**
Normal $N_{\text{train}} = 3500$	0.938	0.969	0.975	0.960	0.949	0.977
De-biased $N_{\text{train}} = 3500$	**0.972**	**1.000**	**0.988**	**0.981**	**1.000**	0.977

Table 1: Fusion AUC on demographic subsets within MIMIC-IV data.

Model	Black	Asian	Hispanic	White	Other	Overall
Normal	0.011	0.013	0.007	0.007	0.022	0.006
De-biased	0.031	**0.009**	**0.005**	0.010	**0.000**	0.008

Table 2: Standard deviation of fusion AUC scores across training sample sizes on demographic subsets within MIMIC-IV data.

AUC scores is lower for the standard fusion model. However, the standard deviation of AUC scores across different training sample sizes is lower for the de-biased fusion model in rarer demographic groups including Asian, Hispanic, and Other Races. This indicates that the de-biased fusion model yields more stable prediction results whereas there is more fluctuation using the standard fusion model in such rarer demographic groups. This potentially informs that the de-biased fusion model could have a larger generalizability in rarer demographic groups whereas the standard fusion model could be sensitive to new data.

5.3 Causal De-Biasing Time Complexity

In order to reduce the time of the algorithm, we perform matching as a pre-processing step, building up a lookup table of matches based on labels such as *gender*, *marital status*, and *age*, as well as Word2Vec representations. As such, theoretically, the time complexity during training increases from $\mathcal{O}(N)$ to $\mathcal{O}(NK)$ where N is the number of epochs and K, the number of categories we are de-biasing over. Empirically, we see around $2K$ difference, but this will likely vary depending on compute available.

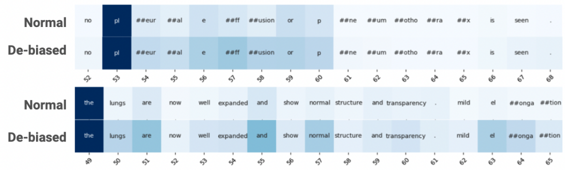

Figure 3: Attention scores calculated from the final layer of DistilBERT encoder. Differences between de-biased and normal training manifest in token attribution strength, rather than which tokens are being attributed to.

5.4 De-Biasing May Improve Model Interpretability of All Modalities

We demonstrate how attention scores shift after casual de-biasing using two examples (**Figure 3**). Interestingly, we do not see a difference in the specific tokens contributing to a given prediction. However, we do notice a large relative difference in the strength of word attributions. Specifically, in the de-biased models, there is stronger attribution to works such as *effusion* in the first example, and *lungs*, *expanded*, *and*, *normal*, and *elongation* in the second example. This suggests that after de-biasing, our model may be focusing on more relevant secondary signals in textual data, rather than just the strongest signal. Since the strongest signals

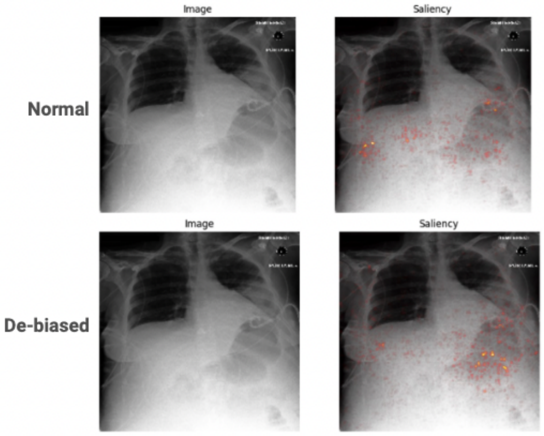

Figure 4: Saliency maps of a pleural effusion case using standard training and casual de-biasing. Standard training results in gradients focused on spurious features such as thoracic catheters (chest tubes), whereas there is less focus on these features after de-biasing.

did not change between normal and de-biased models, we hypothesized that there might also be a difference in what areas of the imaging modality are contributing to final model predictions.

We investigated pixel attributions by taking gradients with respect to the inputs, creating saliency maps (**Figure 4**). Interestingly, we find that there is a slight difference in pixel attributions between the two models. In the normal training paradigm, the model has highest pixel attribution towards spurious features such as the thoracic catheters on the left and right sides of the lungs. In the de-biased model, there is more attribution to the lung cavity, rather than specifically on the catheters, suggesting again that de-biasing may influence the model to focus on less spurious features, similar to the textual modality.

The increased attribution on more relevant features after de-biasing could be due to the matching process in which we can formulate each matched patient as a different *view* of the target patient. This formulation is analogous to data augmentation, as such, our de-biasing method may improve generalization by reducing model focus on non-relevant, spurious correlations.

6 Limitations and Future Work

One limitation of our project is the training sample size given our computing limit. In the future, one may consider larger training sample sizes if computing resource is available. In addition, in §5, we saw that the prediction with the causal de-biased fusion model is more stable than the standard fusion model in some rarer demographic groups. This potentially indicates that our proposed causal de-biasing algorithm could be extended and generalized to build more generalizable multimodal models (e.g. using the contrastive learning paradigm (9)) or large language models.

Future studies could consider a more complex fusion model and investigate how the performance of the de-biased fusion model changes with different levels of model complexity. In addition, one could include more features in a wider range of socio-economic and cultural factors, in addition to *gender*, *marital status*, and *age*, offering a more holistic view for counterfactual matching in the causal de-biasing algorithm for counterfactual matching in the causal de-biasing algorithm. In real setting, those features could be high-dimensional. Thus, another consideration is the choice of distance measure used to find each observation's counterfactual. One could investigate how the performance of the causal de-biasing algorithm changes with different choices of distance measures. Also, testing our algorithm across more challenging diseases and medical conditions would help in assessing its generalizability and effectiveness in diverse medical contexts. Given the time limit of this project, we left them for future work.

For long-term practice, we could conduct longitudinal studies to track the effectiveness and impact of de-biased models in clinical settings over time and explore the integration of these de-biased models into existing clinical workflows, which would provide valuable insights into their real-world applicability and regulation.

7 Conclusion

In this study, we tackle the issue of bias in multimodal healthcare models by developing and applying a causal de-biasing algorithm to the integrated radiological images, textual reports, and demographic data. Our methodology involved quantifying existing biases and implementing counterfactual adjustments based on demographic data. We utilize modality fusion techniques in our approach. The results demonstrate improvements in model fairness, as evidenced by more balanced and improved performance across different racial groups in predicting pleural effusion of the lung. Notably, the de-biasing may actually improve model interpretability by reducing spurious correlations in model predictions.

8 Impact Statement

The significance of this study transcends its immediate application to multimodal healthcare models. By successfully applying a causal debiasing algorithm into transformer-based and neural network models, we provide a new and model-agnostic solution of using counterfactual matching to deal with bias issues across different biased groups. This versatility is particularly valuable in fields where data is inherently imbalanced or where certain groups are underrepresented.

Furthermore, the inclusion of a causal inference framework in our multimodal healthcare model addresses a critical gap in current AI applications - the lack of interpretability and explainability. Current machine learning models including large language models rarely perform causal inference, which makes their application in the healthcare field doubtful. Our work not only enhances the causal reasoning capabilities of these models but also contributes to the ongoing discourse on transparency and accountability.

This research highlights the need for continuous evaluation and improvement of AI systems to ensure they remain fair and effective. Moreover, our findings contribute to the broader conversation on emphasizing the responsibility of AI practitioners to consider the societal implications of their work and strive for technologies that are equitable and beneficial for all in future work.

References

[1] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2022.

[2] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. Fair machine learning in healthcare: A review, 2022.

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

[4] Sepehr Janghorbani and Gerard de Melo. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models, 2023.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[6] Chuizheng Meng, Loc Trinh, Nan Xu, and Yan Liu. MIMIC-IF: interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *CoRR*, abs/2102.06761, 2021.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[8] Joshua R. Minot, Nicholas Cheney, Marc Maier, Danne C. Elbers, Christopher M. Danforth, and Peter Sheridan Dodds. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance. *CoRR*, abs/2103.05841, 2021.

[9] 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.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[13] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *CoRR*, abs/2109.05433, 2021.