Using the Label-Free Concept Bottleneck Model to Improve Fairness in Computer Vision

Daniel Kim

05/11//2024

Abstract

Recent advancements in Artificial Intelligence (AI) have propelled image classification into a rapidly evolving domain. Nevertheless, numerous studies have uncovered the propensity of AI models to introduce or exacerbate biases during decision-making processes. To ensure the safe deployment of these models in critical scenarios, their outcomes must remain impartial concerning sensitive attributes such as gender, race, and disability. In addressing this challenge, we advocate for the utilization of the Label-Free Concept Bottleneck Model (LF-CBM), which conducts image classification based on concealed concepts that obfuscate sensitive information within images. In this paper, we apply LF-CBM and analyze the model's absolute accuracy and performance disparity between different genders. Additionally, we implement dataset pre-processing techniques to mitigate the influence of biased datasets and introduce a "visualization score" to further attenuate potential sources of bias in LF-CBM. Our findings indicate that LF-CBM, augmented with a visualization filter, achieves similar accuracy compared to other state-of-the-art models while mitigating accuracy disparities between genders.

1. Introduction

Image classification models using Artificial Intelligence (AI) and Machine Learning (ML) have expanded tremendously in the past few decades. While they have helped society by automating manual work in various industries, there are also growing ethical concerns around biased results created or amplified by these models. For example, studies found that some facial recognition and classification systems like the COMPAS software [1] misidentify African and East Asian faces 10 to 100 times more than Caucasian faces. As these models are increasingly deployed in critical applications, they must produce unbiased results, especially concerning sensitive attributes such as gender, race, and disability.

There have been studies that aimed to achieve fairness in computer vision models in different ways, such as loss reweighting [13], optimization constraints [15], and adversarial debiasing of images [11]. These methods, however, still pose key challenges in their application. First, the requirement for manual gender labeling and vulnerability to adversarial attacks make them unscalable in many practical settings. Moreover, even when they achieve higher fairness, it is hard to identify the source of their success as they often fail to show the underlying changes that led to the improvements. Hence, we need a way to debias modern image classification models in a secure, scalable manner while achieving transparency and interpretability in the way it does it.

In this paper, we propose a model that expands on the Label-Free Concept Bottleneck Model (LF-CBM), which was introduced by Oikarinen et al. [12] in 2023, to mitigate bias while achieving accuracy similar to other

state-of-the-art models. With LF-CBM, it conducts image classification from the penultimate concept layer, whose neurons represent each concept and are activated by the images. LF-CBM leverages GPT to generate these concepts and CLIP to train the projection from the backbone model to the concept layer, providing a solution that can be applied with no labels or human annotations. We have found a few reasons why LF-CBM can contribute to debiasing models effectively. First, mapping an image to a set of concepts unrelated to gender should, theoretically, remove biases. For example, if we conduct a classification task based on concepts such as "a desk" and "glasses," the model should not be able to infer gender information from an image. Second, LF-CBM's concept layer provides much more interpretability when compared to other computer vision models, which helps us understand and identify the source of biases in the model's internal representation. Lastly, LF-CBM does not require gender information or human labeling, allowing it to be used for any dataset and at scale.

In addition to applying LF-CBM in the gender-bias-mitigating context, we make contributions in two main other ways: 1) dataset pre-processing to examine the impact of biased datasets and 2) introduction of a visualization filter to further remove the potential sources of bias. First, we analyze the model performance in two main datasets: one where male and female images are equally present (balanced) and the other where male and female imbalance from the original dataset is preserved (imbalanced). Comparing the results in these two datasets allows us to examine the effect of dataset bias on the model performance and assess how the model performs in a real-world-like scenario when the dataset is biased. Second, on top of the original LF-CBM architecture, we add another component that filters through the GPT-generated concepts and removes invisible concepts from the layer. Since invisible concepts such as "love" and "affection" are nuanced and more likely to be a source of sentimental bias that could serve as a proxy for gender information, removing them could help with reducing gender bias.

We use accuracy and accuracy parity metrics to analyze the model performance on classification and fairness. Then we compare our model's performance to that of other state-of-the-art models such as ResNet-50 [4] as the baseline. Overall, we show that LF-CBM with a visualization filter achieves similar accuracy while reducing accuracy parity between the genders when compared with the baseline. Furthermore, we demonstrate our model's superior performance on both balanced and imbalanced datasets, indicating that our model can be used to produce debiased results in real-world image classification domains with biased datasets. Our findings not only underscore the effectiveness of LF-CBM with a visualization filter in mitigating gender-based accuracy disparities while maintaining competitive accuracy levels but also highlight its potential applicability in addressing biases in diverse real-world image classification scenarios.

2. Related Work

In this section, we explore literature works that discuss fairness in AI models as their main topic. Specifically, we first look into how fairness is defined in the computer vision field. Then, we delve into different approaches that have been developed to remove biases in AI models, including their strengths and weaknesses.

2.1 Defining Related Work

Before diving into improving fairness, it is important to first define what “fairness” means in the image classification domain. Broadly, computer vision algorithms trained with data are unlikely to produce a *disparate treatment* (direct, intended discrimination) but may induce *disparate impact* (indirect, unintentional discrimination) [9]. Furthermore, there are many options one can use to quantify a model’s disparate impact and evaluate the fairness of the results produced. For example, *equalized odds* [3] compute the difference between the false-positive rates (FPRs) and between the true-positive rates (TPRs) for multiple groups, aiming to ensure different groups get similar positive rates. *Equality of opportunity*, [3] on the other hand, requires that only TPRs be similar across groups, which can be particularly relevant in scenarios where both privileged and underprivileged groups must have an equal chance for a positive outcome. Last but not least, *accuracy parity* compares and aims to minimize the differences in the accuracies between different groups [14]. There are tradeoffs to be made for different measures of fairness, and their effectiveness also depends on the context in which the model is being used.

2.2 Adjacent Efforts for Fairness in Computer Vision

Several sources contribute to biases in machine learning models, including those in datasets, missing data, algorithmic objectives, and "proxy" attributes for sensitive features [9]. There have been a few bias detection models and techniques developed in other domains, such as loss reweighting and adversarial learning [13]. Nonetheless, they often fail to work successfully in the image domain due to a few reasons. First, the intricate layers of input and hidden data in computer vision models pose challenges to fairness, as biases in training data may elude detection and the complexity of image features hampers understanding and identification of biases. Furthermore, labels for these image data are unclear and unscalable. Thus, it’s important to look deeper into past work on improving fairness in the image classification domain specifically. One proposed way is to introduce fairness constraint during optimization, which decreases the magnitude of bias amplification up to 47.5% [15]. Another approach is adversarial debiasing of images [11], which essentially removes unwanted features corresponding to protected variables while keeping information that is useful to recognize objects or verbs. These methods still introduce other challenges, however. For example, while these methods achieved fairness objectives, they do not explain the underlying changes in the internal representation that led to their improvements, making the results hard to interpret. Furthermore, they require gender annotations in their data and have only conducted their study on the original train/val/test split on the imSitu dataset, which was not partitioned concerning gender. They also are vulnerable to adversarial attacks and need large amounts of labeled data. In this paper, we tackle these challenges by focusing on mitigating gender biases within computer vision.

3. Preliminaries: Label-Free Concept Bottleneck Model

The basis of our model used the Label-Free Concept Bottleneck Model (LF CBM), a neural network that performs classification based on the concepts [7]. Creating an LF-CBM model takes 4 main steps: 1) create and filter the initial concept set, 2) compute embeddings from the backbone and the concept matrix on the training dataset, 2) learn projection weights to create a Concept Bottleneck Layer (CBL), and 4) learn the weights of the sparse final layer for the prediction. The diagram of this entire process is shown in Figure 3.1. In this paper, we briefly discuss each step. For more in-depth technical details, one should read Oikarinen et al.'s paper on LF-CBM.

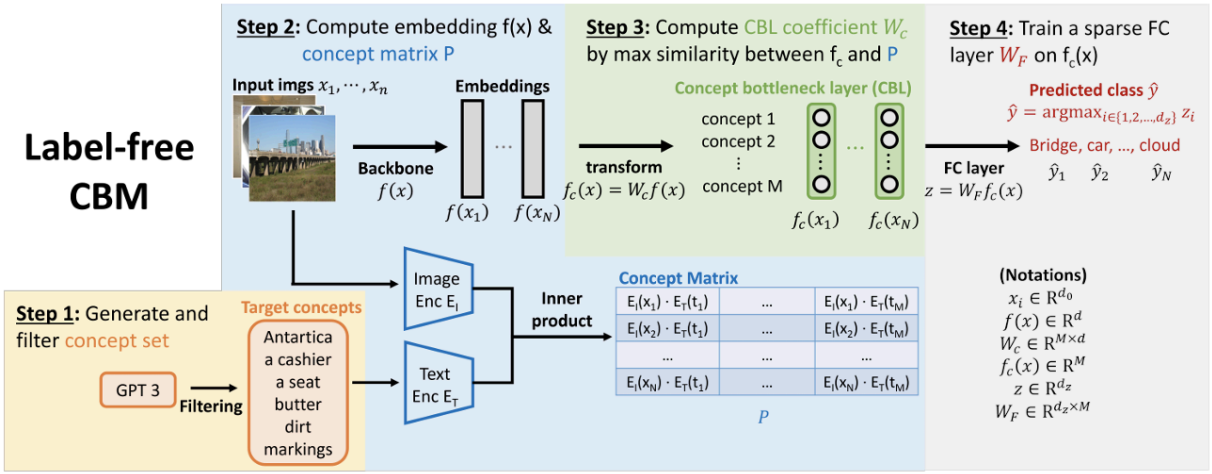


Figure 3.1: Label-Free Concept Bottleneck Model (LF-CBM) 4-Step Workflow Overview [7]

3.1 Concept Generation

We generated the concept set via LLM, specifically GPT-3 using the OpenAI API [2]. Specifically, we asked GPT-3 the questions introduced in the original LF-CBM paper, such as "List the things most commonly seen around a {class}," for each of the 200 classes. Then, we filtered the generated concepts using the original criteria: *too long concept*, *concepts too similar to classes, concepts too similar to each other, concepts not present in training data,* and *concepts that can't be projected accurately*. Furthermore, for our specific study, we filtered out the concepts that contained gender-specific information (such as "male" or "lady").

3.2 Learning Concept Bottleneck Layer (CBL)

For the backbone of our model, we used the architecture of ResNet-50 as in the original paper [76. Similarly, we utilize CLIP-Dissect to create a concept matrix on this training dataset [8]. Then, we created a Concept Bottleneck Layer (CBM) by training a projection matrix by minimizing the *cos-cubed* similarity between the concept matrix and the concept activation based on this projection matrix.

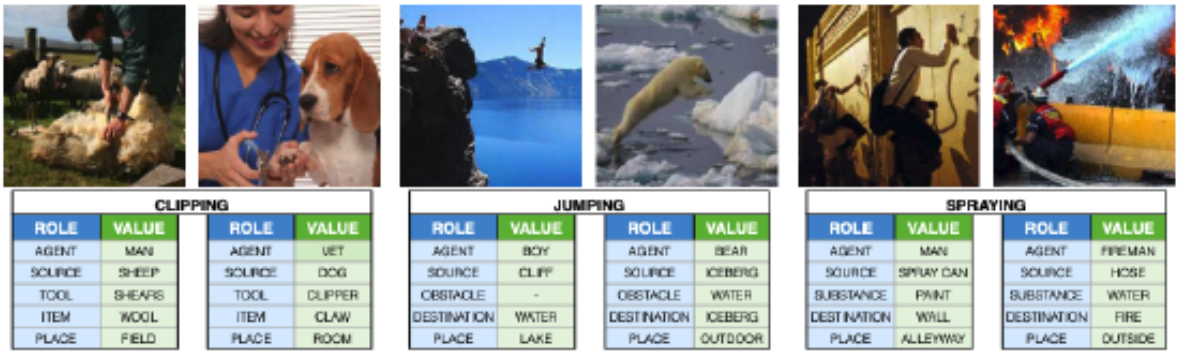
3.3 Learning the Sparse Final Layer

As the last step, we learned the final predictor with the fully connected, sparse linear layer. As with the original paper, we used GLM-SAGA solver [11] for optimization, which generally results in 0.7%-15% of the weights of the model being nonzero.

4. Methods

In this section, we present two main contributions made by this paper. First, we preprocessed the dataset such that it's easy to study the impact of a biased dataset on the result. Second, we applied an additional visual filter to the existing LF-CBM architecture to mitigate a potential source of bias in the model.

4.1 Dataset Preprocessing



CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Figure 4.1: Sample images of the imSitu dataset. Below each image are the annotations with the verb at the top, activity-specific roles in the blue column, and values for the roles in the green column. *Agent* role often provides gender information of the image.

We mainly conducted our experiments on a multi-class classification task, which classifies among more than two classes. Specifically, we wanted to compare our model's accuracy and accuracy parity (difference in accuracies between the groups) with other State-of-the-Art models. To do so, we examined the performance of our model on the imSitu-200 dataset [6]. The imSitu dataset focuses on predicting real-life activities based on an image, as well as its associated semantic roles like actors, objects, substances, and location. The example data in the imSitu-200 dataset can be found in Figure 4.1.

4.1.1 Dataset Modification

Biases in training datasets can be a source of biases for the AI model itself [8]. The imSitu-200 dataset itself also presented a noticeable gender imbalance, particularly favoring males. To isolate and analyze the impact of

biased datasets on the performance of our model, we trained the LF-CBM model with two different datasets: balanced and imbalanced. By training the model with a dataset that ensured a 50/50 balance between the two genders, we isolated the impact of biases in the dataset on the model's performance and identified other sources of biases in the model. On the other hand, by training the model with the imbalanced dataset, we aimed to study the robustness of our model in a more realistic scenario, where the dataset often will contain biases.

To create a balanced dataset, we first ensured that the number of male and female samples for each class was equal by cutting each gender's dataset size to the lower of the two. For example, if an action *climbing* originally had a dataset of 48 male and 36 female images, we would process them such that there are 36 male and 36 female images. Then, we split the dataset into training/validation/testing splits of 60/20/20 using a stratified split based on gender and classes. This way, we made sure that the gender split of 50/50 was preserved across the splits. This procedure is illustrated in Figure 4.2 below.

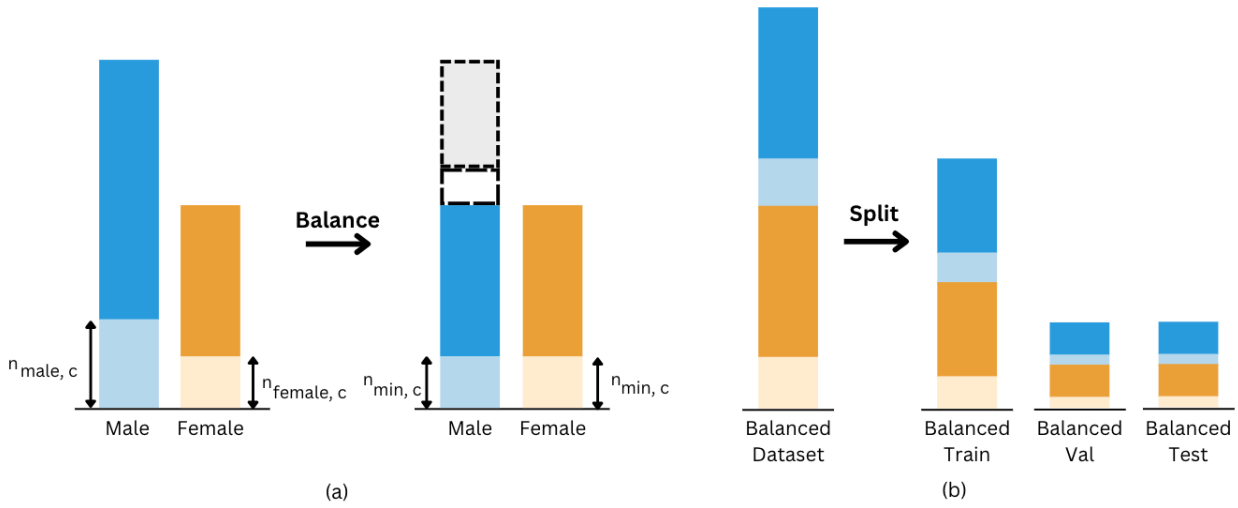


Figure 4.2: Visualization of the creation of a balanced dataset. (a) shows how the dataset gets balanced by taking $n_{min,\ c}$ samples for each class c. (b) shows the splitting of the balanced dataset into train/validation/test ratio of 60/20/20.

To create an imbalanced dataset we first divided the original imbalanced dataset into the same 60/20/20 split using a stratified split while making sure there existed consistent gender imbalance from the original dataset across the splits. Then we cut down the dataset size to match that of the balanced set while preserving the gender imbalance from the original dataset. This way, we could isolate the impact of dataset size on the results produced from these two different datasets. This procedure is illustrated in Figure 4.3 below.

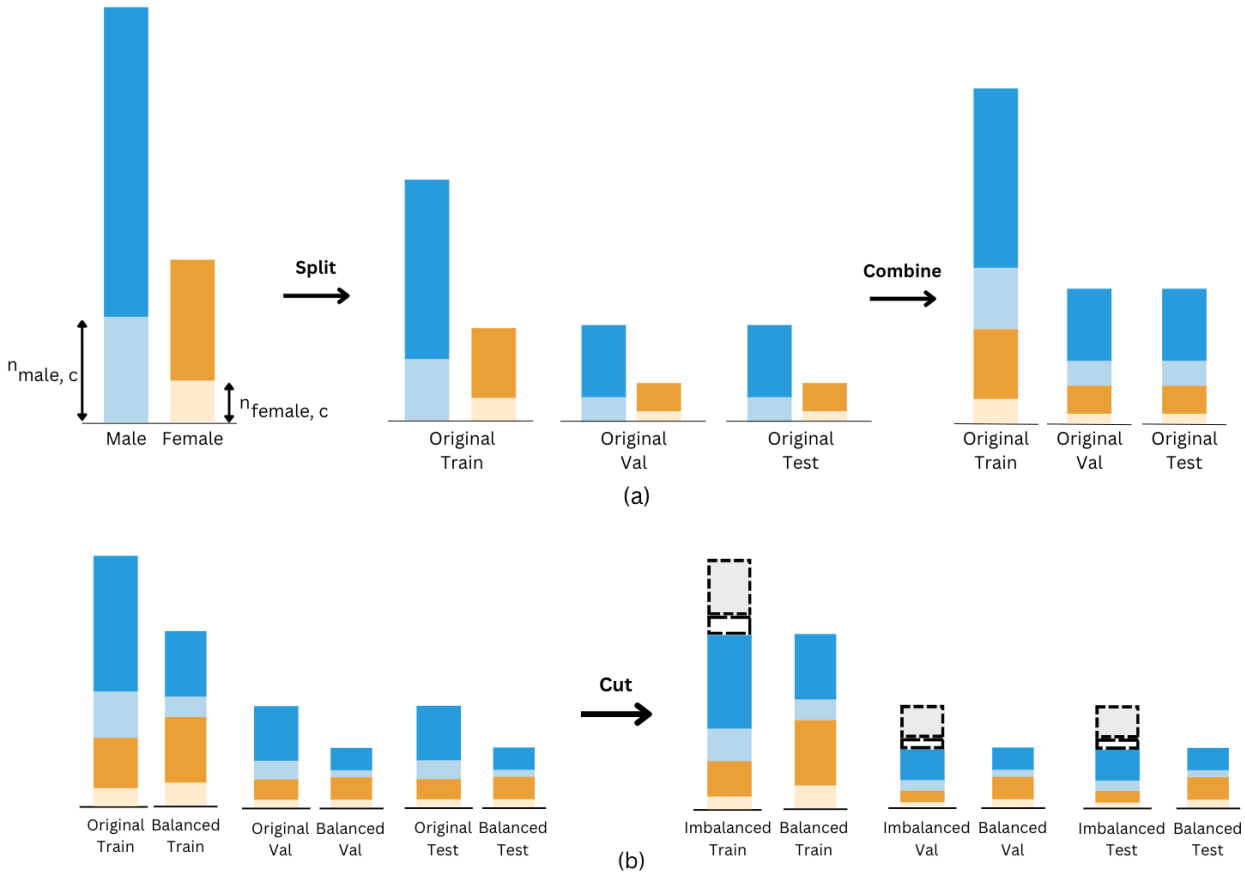


Figure 4.3: Visualization of the creation of an imbalanced dataset. (a) shows how the dataset first gets split into train/validation/test ratio of 60/20/20. (b) shows how the split datasets from the original dataset get cut to match the size of balanced datasets, thus becoming imbalanced datasets.

4.2 Final Architecture with Visual Filter

While LF-CBM provided a decent basis for our model, some concepts generated during this process were more vulnerable to introducing biases into the model than others. One characteristic of these more vulnerable concepts is their invisibility. Specifically, concepts such as “love” and “affection” that are not as visual could be a source of biases when associating each image with these concepts. Hence, we introduced another constraint for the generated concepts, where non-visual concepts got filtered out in the process. The change in the model architecture with this introduced filter is demonstrated in Figures 4.4 and 4.5.

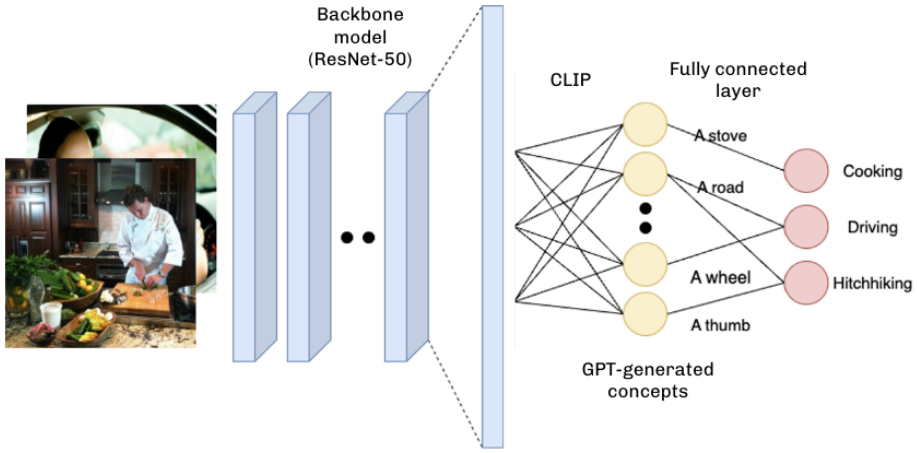


Figure 4.4: LF-CBM architecture before the introduction of the visual filter.

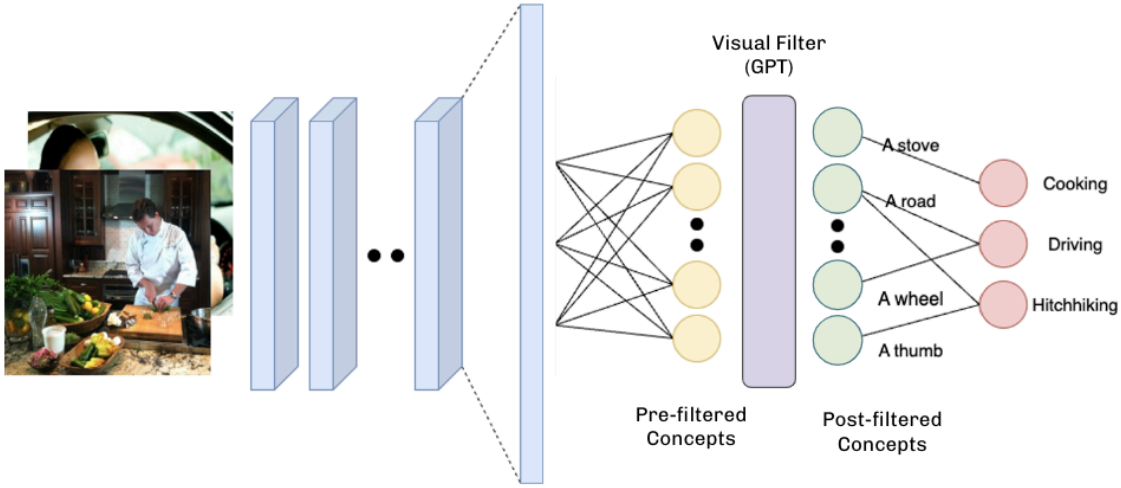


Figure 4.5: Final model architecture with both LF-CBM and the visual filter.

4.2.1 Visualization Score Generation

The line at which a concept is considered *visual* or *non-visual* is not as clear. To minimize human subjectivity and biases when determining the visibility of each concept, we used GPT-3 to generate a visualization score for each concept. This visualization score, which runs on a scale of 1-10, indicates how easily visible a certain concept is, where a higher score corresponds to more obvious visibility. After experimenting with a few prompts, we decided to generate visualization scores by asking GPT-3 the following question: “For human eyes, when looking at a picture, how detectable is (on the scale of 1 to 10) {class}?”. Further study and examples of the outputs are detailed in Section 5.4.

4.2.2 Visualization Score Cutoffs

With this visualization score attached to each concept, we still needed to decide at what score a concept was visible enough to remain in the concept set. Specifically, for every concept c and its visualization score v_c from the original concept set C, we needed to specify a visualization score cutoff score i such that our new concept set C_n only contains concepts whose v_c was greater than or equal to i. There are two main tradeoffs when considering what the cutoff should be. First, too low a cutoff score may be inefficient in filtering out the non-visual concepts from the original concept set, hence exposing our model to learning biases with the new concept set that it would have learned with the original concept set as well. On the other hand, too high a cutoff score may significantly limit the number of concept neurons in CBL, which could lower the expressibility and accuracy of the entire model. To determine the best cutoff score to be used, we designate i as a hyperparameter and study the impact of it on the performance of our model, both in accuracy and accuracy parity.

5. Experiment Results

In this paper, we offer a comprehensive analysis of the primary performance metrics of the models under comparison, focusing on accuracy and accuracy parity. Additionally, we present an investigation into the interpretability aspects concerning the generation of visualization scores and the inference process of the LF-CBM model

5.1 Datasets

To assess the efficacy of our methodology, we conduct training and evaluation procedures on the imSitu-200 dataset [6], renowned for its emphasis on predicting real-world activities depicted in images, alongside their corresponding semantic roles encompassing actors, objects, substances, and spatial context. As expounded upon in Section 4.1, preprocessing of the imSitu-200 dataset yields two supplementary datasets: balanced and imbalanced. Consequently, our analysis encompasses a comparative evaluation of model performance across three distinct variants of the imSitu-200 dataset: the original (full), balanced, and imbalanced configurations.

5.2 Training & Model Setup

We conducted comprehensive evaluations encompassing both our proposed models and a baseline counterpart to facilitate a comparative assessment of the novel models presented herein against the State-of-the-Art paradigm. The baseline model was instantiated utilizing the ResNet-50 [4] architecture, chosen for its established efficacy in computer vision tasks owing to its deep structure featuring skip connections, which effectively alleviates vanishing gradient challenges and promotes the training of deeper networks, thereby enhancing feature extraction and representation learning from image data. The model utilized the Adam optimizer [5] and Cross Entropy Loss function for its training. The specified hyperparameters for the baseline model encompass a learning rate of 0.001, momentum of 0.8, step size of 5, gamma of 0.1, and epochs of 25, which were the hyperparameters that yielded the best accuracy based on our grid search. Subsequently, results were generated for eleven distinct models for comparison against the baseline: one model adopting the LF-CBM architecture sans a visual score filter layer, and ten variants incorporating the LF-CBM architecture alongside a visual score filter layer, each corresponding to a unique visualization score cutoff ranging from 1 to

10. Across all models, a consistent iteration count of 80 and sparsity lambda value of 0.0007 were applied. Training and evaluation procedures were executed utilizing MIT's HPC Supercloud infrastructure [10], with the duration of each training run ranging from several minutes to up to 20 hours contingent upon the model architecture and dataset dimensions.

5.3 Results (I): Accuracy and Accuracy Parity

Table 5.1 presents a comprehensive overview of the performance summary encompassing the baseline (ResNet-50) alongside various LF-CBM models. As discussed earlier, the performance evaluation of each model is delineated through three pivotal metrics: accuracy, aggregate accuracy parity, and per-class accuracy parity. The term "Baseline" denotes the ResNet-50 architecture model employed as a benchmark for comparative analysis. The designation "CBM" signifies the LF-CBM model configuration devoid of a visual score filter, while the "Visual" models denote LF-CBM models incorporating a visual score filter, with the numbers within parentheses denoting the respective cutoff scores utilized for filtration. Furthermore, annotations "Full," "Balanced," and "Imbalanced" indicate the datasets utilized for training and evaluation purposes. Key insights derived from the results are outlined as follows:

	Accuracy	**Accuracy Parity**	
	Total	**Aggregation**	**Per-Class**
Baseline Full	31.26	3.13	14.19
CBM Full	**31.62**	3.05	16.14
Visual Full (≥ 8)	30.24	2.62	14.13
Visual Full (≥ 7)	30.75	2.34	13.94
Visual Full (≥ 6)	31.01	2.04	15.16
Visual Full (≥ 5)	31.22	**1.96**	**13.78**
Baseline Balanced	**29.27**	1.20	13.93
CBM Balanced	29.10	2.25	14.13
Visual Balanced (≥ 8)	28.59	1.71	14.49
Visual Balanced (≥ 7)	28.77	**0.51**	13.45
Visual Balanced (≥ 6)	28.59	2.25	**13.00**
Visual Balanced (≥ 5)	28.75	1.71	15.05
Baseline Imbalanced	29.68	**1.52**	**14.19**
CBM Imbalanced	**30.07**	2.58	16.14
Visual Imbalanced (≥ 8)	29.10	3.31	14.87
Visual Imbalanced (≥ 7)	30.17	2.17	14.85
Visual Imbalanced (≥ 6)	29.64	3.07	14.83
Visual Imbalanced (≥ 5)	30.03	2.27	15.59

Table 5.1: Performance comparison. For each dataset (Full, Balanced, Imbalanced), the highest accuracy, the lowest aggregate accuracy parity, and the lowest per-class accuracy parity are bolded.

1. **Minimal Impact on Accuracy:** The achieved accuracies across the baseline and various LF-CBM models demonstrate negligible discrepancies on each dataset, indicating that our proposed models generally do not induce any discernible degradation in comparison to the baseline.
2. **Consistent Accuracy Parity Trends:** As anticipated, training sets with balanced class distributions exhibit the lowest levels of aggregate and per-class accuracy parity. This underscores the notion that under ideal conditions, where models are trained and evaluated with balanced datasets, equitable performance among different gender groups can be achieved.
3. **Mitigated Accuracy Parity via LF-CBM and Visualization Score:** Notably, the LF-CBM models, particularly those utilizing visualization cutoffs of 5 and 6, showcase the most favorable outcomes in terms of minimizing per-class accuracy parity for full and balanced datasets, respectively. This observation highlights the potential of our models to yield fairer outcomes while upholding state-of-the-art accuracy levels.

5.4 Results (II): Visualization Score Results

As a significant aspect of our investigation, we derived visualization scores for each concept. These scores, generated by GPT-3, are delineated on a 1-10 scale, reflecting the perceived visibility of a given concept, with higher scores indicating greater visibility. In this section, we furnish concrete instances of visualization scores alongside their aggregate distribution, aimed at elucidating the interpretative underpinnings of these scores.

5.4.1 Visualization Score Example

Validating the utility of visualization scores produced by GPT-3 as a metric for assessing concept visibility necessitates their alignment with human perceptions. Demonstrating the accuracy and alignment of these generated scores with the human understanding of object visibility is paramount. To this end, Table 5.2 is presented, featuring 10 concepts alongside their respective visualization scores generated by GPT-3 and those assigned by humans, herein represented by the researchers themselves. Examination of the table reveals a notable correspondence between the visualization scores generated by GPT-3 and those determined by humans. However, as part of future endeavors, the establishment of a more robust method for defining human baseline visualization scores is proposed, a prospect elaborated upon in the subsequent Future Work section.

	GPT-3	Human
laptops	9	10
a mouse	8	9
a trigger	3	4
transportation	5	5
a dolly	7	7
verb	1	1
two or more people talking to each other	10	8
a look	6	5
two handles on the top	4	4
eye contact	10	7

Table 5.2: Visualization score example. The left column contains 10 randomly sampled concepts and the middle column shows the visualization score generated for each concept. The right column shows the visualization scores assigned by humans.

5.4.1 Visualization Score Distribution

This section delves deeper into the distribution of the generated visualization scores across the entirety of the concept set. In Figure 5.3, we present an illustrative depiction of the distribution showcasing the frequency of concepts across the 10 potential scores. Analysis of this distribution yields two primary observations. Firstly, the score distribution exhibits a left-skewed pattern, characterized by a prevalence of concepts assigned scores of 6 or higher. This skewness could be attributed to the nature of the prompts utilized during the initial generation process, which emphasized aspects of visibility through phrases such as "List the things most commonly seen around" and "Give visual superclasses for...". Secondly, notable proportions of concepts are observed to possess visualization scores below 5, underscoring the diversity within the concept set. This observation underscores the potential for substantial variations in model outcomes based on the application of visualization score filters with different cutoff thresholds, thereby instigating meaningful distinctions in model results.

Figure 5.4 illustrates the cumulative distribution of visualization scores assigned to the generated concepts. Each bar represents the cumulative count of concepts that would be retained if the visualization score indicated on the x-axis were chosen as the visualization score cutoff. Consistent with expectations, a higher score cutoff, indicative of a more stringent criterion, correlates with a reduction in the number of retained concepts.

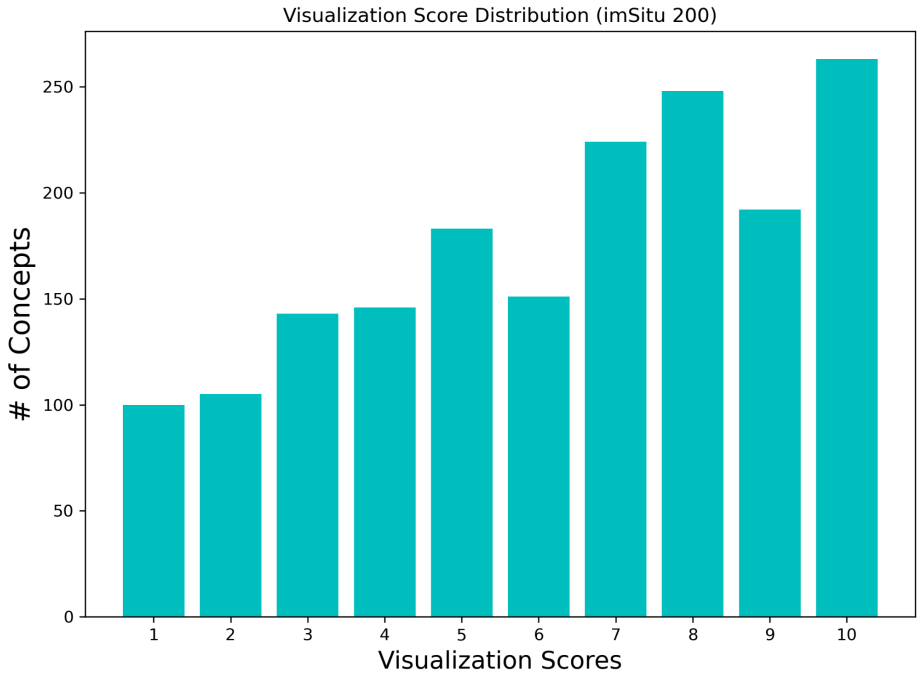


Figure 5.3: Histogram that displays the distribution of the visualization scores of the generated concepts. Notable proportions of concepts are observed to possess visualization scores below 5, underscoring the diversity within the concept set.

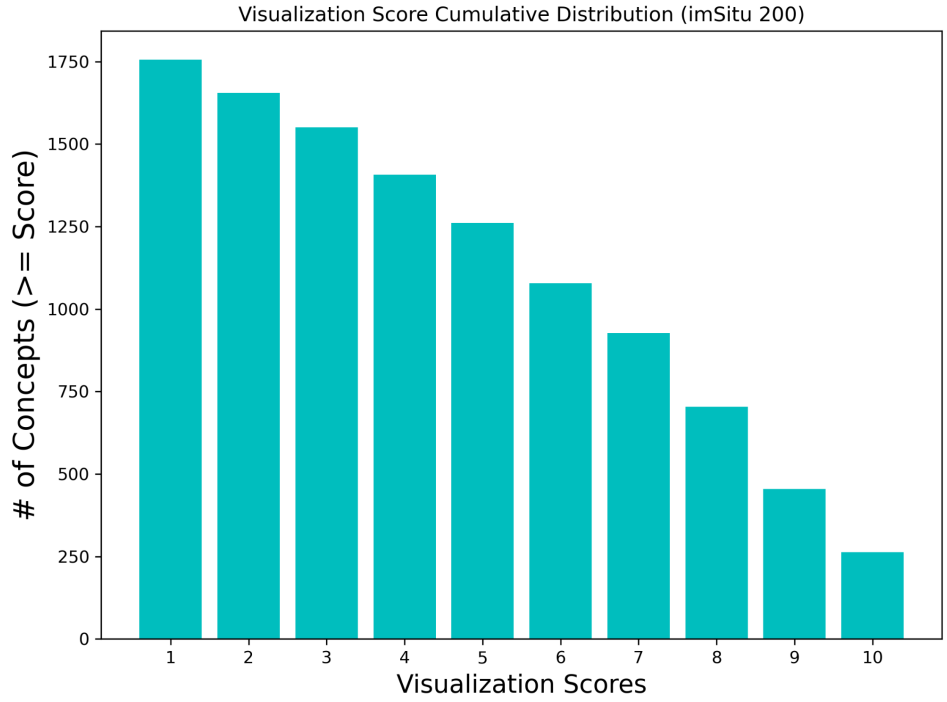


Figure 5.4: Histogram that displays the cumulative distribution of the visualization scores of the generated concepts.

5.5 Results (III): LF-CBM Analysis

In this section, we delve into the interpretability facilitated by CBM, allowing for insightful visualizations. Through focused examination of specific examples, we offer accessible demonstrations of the inference process, thereby enhancing comprehension of the CBM's interpretive capabilities.

Figure 5.5 illustrates the weight distributions corresponding to four distinct classes within a model trained on the imSitu dataset. Notably, the depicted weights reveal a coherent alignment between the concepts utilized for classification and the identified classes. For instance, prominent concepts with elevated weights for the "pedaling" class include "a bike rack," "a cyclist," and "a pedal," indicative of the model's effective association between concepts and classes.

Figure 5.6 depicts sample-level visualizations, offering insights into the model's classification methodology. For the image depicting "drinking," the model effectively identifies the presence of glass and liquor. Nevertheless, additional concepts, such as a bottle and a cup of coffee, are also identified despite their absence in the image, suggesting the potential learning of proxies rather than semantic concepts. This observation raises the possibility of the model associating "a cup of coffee" with correlated information, such as hand positioning, rather than the tangible presence of a cup of coffee itself.

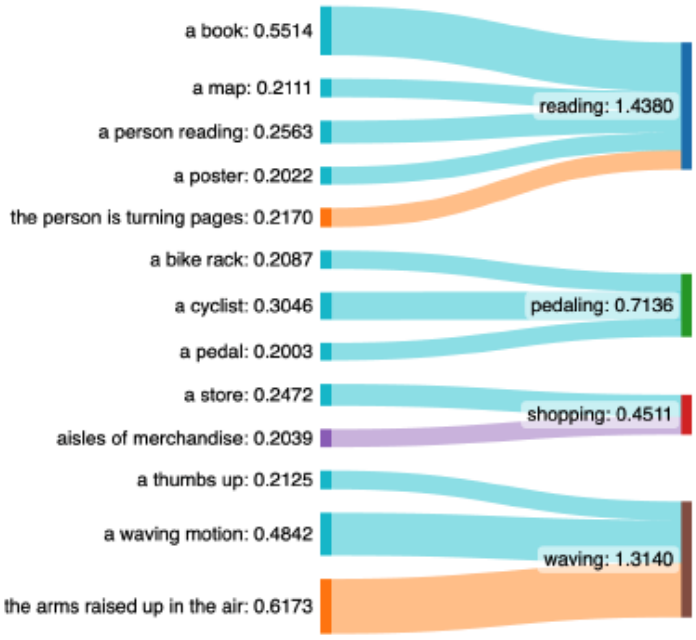


Figure 5.5: Sankey diagram of the CBM weights for four classes in the imSitu dataset trained on the balanced training set

Figure 5.6: A sample image with the prediction made by the model and the contribution of each concept in making the prediction.

As showcased with the examples above, CBM's concepts association with images and labels provides human-interpretable insights into the inference process and potential sources of biases and incorrect predictions.

5.6 Discussion

Despite previous efforts to address biases in computer vision models, scalability and interpretability remain significant challenges. Our proposed LF-CBM model, augmented with visualization scores, offers a scalable and comprehensible solution to achieving fairness in computer vision. By harnessing GPT-3 to generate concepts and the CLIP model to associate images with these concepts, our approach eliminates the need for manual image labeling and is applicable even in large-scale domains. Moreover, as detailed in Section 5.5, the visualization of CBM's inference process enables human understanding of decision-making, shedding light on both successful bias mitigation and the origins of unfair inferences. However, our model relies on access to third-party models like GPT-3 and CLIP, making it vulnerable to changes in their availability or pricing. Additionally, while scalable, our model necessitates concept generation for each domain, posing technical hurdles. Exploring the co-occurrence of concepts and gender could further enhance bias mitigation, as concepts may serve as proxies for gender-related information. Future research should investigate the correlation between gender and concepts, linking this data with concept importance in classification to uncover and address biases in CBM.

6. Conclusion

In this paper, we introduced *LF-CBM* as a tool to mitigate biases in image classification tasks and understand the potential sources of biases better from its superior interpretive capabilities. Moreover, we conduct dataset pre-processing to create *balanced* and *imbalanced datasets,* which are used to study the model's performance in both realistic and ideal settings. We also introduce the *visual filter*, which builds upon the established LF-CBM

architecture further to remove potential sources of biases in the concept layer. Our results emphasize both the efficacy of LF-CBM with a visual filter in reducing gender-based accuracy gaps while upholding competitive accuracy and its potential to tackle biases across various real-world image classification scenarios.

References

1) Angwin, Julia, et al. "Machine bias." Ethics of data and analytics. Auerbach Publications, 2022. 254-264.
2) Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
3) Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).
4) He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
5) ‘Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
6) Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. “Situation Recognition: Visual SemanticRole Labeling for Image Understanding”. In: Conference on Computer Vision and PatternRecognition. 2016.
7) Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T. W. (2023). Label-Free Concept Bottleneck Models. arXiv preprint arXiv:2304.06129
8) Oikarinen, Tuomas, and Tsui-Wei Weng. "Clip-dissect: Automatic description of neuron representations in deep vision networks." arXiv preprint arXiv:2204.10965 (2022).
9) Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), 1-44.
10) Reuther, Albert, et al. "Interactive supercomputing on 40,000 cores for machine learning and data analysis." 2018 IEEE High Performance extreme Computing Conference (HPEC). IEEE, 2018.
11) Wang, Tianlu, et al. "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
12) Wong, Eric, Shibani Santurkar, and Aleksander Madry. "Leveraging sparse linear layers for debuggable deep networks." International Conference on Machine Learning. PMLR, 2021.
13) Wu, Chuhan, et al. "Fairness-aware news recommendation with decomposed adversarial learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 5. 2021.
14) Zafar, Muhammad Bilal, et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." Proceedings of the 26th international conference on world wide web. 2017.
15) Zhao, Jieyu, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." arXiv preprint arXiv:1707.09457 (2017).