Zero-Shot Gender Classification: Bias Analysis and Prompt Tuning

SIBGRAPI Paper ID: 88

***Abstract*—Multi-modal vision-language models showed promising zero-shot capabilities in several downstream tasks, with no need for fine-tuning or retraining. Different studies have been conducted to better understand how these models were trained, on what data, and how prompt engineering could improve and generalize them. In this paper, we investigate the impact of data and model scaling for zero-shot gender classification using openCLIP and explore how to employ simple prompt tuning methods to improve these results. We perceived a 4.4% improvement in accuracy with the scaling of models, and yet we could not see the same behavior when scaling training data sources. For all models, the original OpenAI's WIT dataset achieved the highest accuracy for some models, even though it is 30 times smaller than the largest CommonPool dataset, composed of 12.3 billion samples. We then explored textual prompt ensembling and aggregation techniques by combining a list of age, race, and gender, and averaging the top-k similarity scores. Our proposed prompts and aggregation methods increased accuracy by a small margin but strongly reduced the difference between the Male and Female accuracy.**

I. Introduction

Large Vision Language Models (VLMs), such as CLIP [1] and ALIGN [2], proved overwhelming zero-shot and few-shot capabilities in several tasks such as image classification, image and text retrieving, geo-localization, and many more. These methods leverage contrastive learning [3] and large-scale datasets comprising millions or billions of image-text pairs automatically scraped from publicly available web sources [4].

By aligning visual and textual information, VLMs allow zero-shot image classification, where several textual prompts (each one containing the name of one category) are compared with the visual features in a *competitive* manner. However, the proper design of the ideal prompt for a given task/dataset is still a challenge [5]–[9]. Another key issue for zero-shot image classification is the complexity/size of the models and the training dataset: as shown in [10], there is a consistent decrease in zero-shot classification errors for several datasets when scaling model, dataset size, and samples seen during training. Their findings are consistent with the hypothesis that larger foundation models, particularly based on Vision Transformers (ViTs), can benefit from more data.

Despite the promising zero-shot capabilities of VLMs, there are still many concerns about the societal impacts in real-world tasks [11]. These Large VLMs are trained without supervision from the images and texts scraped from the web, and are prone to learn the spurious correlations between them. The unlimited flexibility of the class design of such models could result in its misguided use, justifying and amplifying stereotypical and unfair social relations [12], [13]. The nature and the details of the datasets used to train these models are often treasured as business secrets of the highest value by the private companies that built them, leaving researchers and society in the dark. Even for publicly available data such as the LAION-5B dataset [4], the huge amount of data makes a manual curation process virtually impossible. Gadre and colleagues [14] introduced CommonPool, a testbed consisting of 12.8 billion image-text, along with an automatically-filtered version with 1.4 billion samples called DataComp-1B. It is important to note that even automatic methods for filtering datasets on such massive datasets can be unfeasible with limited hardware resources.

This work explores zero-shot capabilities of several CLIP-like models [10] for identifying sensitive information. We narrow the scope of this work to the problem of binary gender classification, aiming to study the impact of the textual prompts and the combinations of models/training datasets. For model scaling, the larger model ViT-g-14 improved overall accuracy by 4.4% compared to ViT-B-32 (95% $\times$ 91%), while reducing the gap between male and female class accuracy to -0.0031 GDif, in contrast to the -0.0606 GDif of ViT-B-32. As for the data scaling, OpenAI's WIT data source performed better than any other source for all model architectures, even though it was one of the smallest ones.

We also present a prompt ensemble approach that combines different human attributes (such as ethnicity and age) and show that this additional information improves the accuracy of zero-shot gender classification for all models while also reducing bias by closing the gap between Male and Female accuracy.

II. Related Work

This work focuses on zero-shot classification of sensitive human information (gender), which involves two main aspects: zero-shot classification and bias analysis. These two main topics will be briefly revised next.

A. Zero-Shot Classification

Large VLMs show promising results for zero-shot image classification [1], where pre-trained weights are used, and the only step required is the prompt design, which must reflect the name of the target classification labels. For example, a two-class problem involving cats and dogs can be seen as the process of encoding an image of such pets and comparing it with encoding a set of different text prompts, such as "a photo of a dog" and "a photo of a cat". The contrastive learning nature of these models will allow us to compute the cosine

similarity between these images and the text prompts to extract the final classification label.

To further explore the capabilities of CLIP in zero-shot tasks, the authors of openCLIP [10] trained several ViT models with different patch sizes and the number of trainable parameters on a variety of open-source datasets, like Laion-400M [15], Laion-5B [4], DataComp-1B, and its unfiltered version with 12.8B image-text pairs called CommonPool [14]. They observed a consistent increase in performance of such models when scaling model parameters, data samples and compute power for multiple downstream tasks. Beyond their findings, they also released more than 120 model checkpoints trained on different dataset combinations. Gadre et al. [14] showed that scaling is important, but the quality of the dataset also plays an important role. In fact, their results in the ImageNet challenge [16] indicated that using a "good" subset of massive datasets can yield better results than training using all data.

Aside from the impact of the model architecture and dataset, other researchers have focused on how to build an adequate text prompt for the desired task. As shown in the original CLIP paper [1], the design of the prompt can significantly impact the accuracy of zero-shot classification. Instead of manually tuning the prompt, CoOp [5] introduced the idea of *prompt learning*, in which the prompts are composed of a learnable root portion that is concatenated with the desired class names, leading to few-shot classification. A follow-up work [6] extends this idea, aiming to better generalize for unseen categories, while Parisot et al. [17] focuses on the problem of handcrafted choice of class names that define queries, mentioning ambiguous names (such as bat, that can be related to the animal or a baseball bat). Despite all the exciting improvements in prompt-learning methods, they require additional training data (typically few-shot).

B. Bias in Vision Tasks

Following the initial bias-related experiments performed in [1], several works have explored social biases present in VLMs. These initial experiments detected racial and gender biases in image classification tasks when prompting CLIP with human and non-human labels, and gender-agnostic profession labels.

Hall and colleagues [13] explore gender bias in VLMs by probing CLIP and two other models built upon it, namely Detic [18] for bounding box detection and LSeg [19] for semantic segmentation. They showed that all models performed better for images containing women and some object types (e.g., bag, necklace and sweater) and images containing men and another set of objects (e.g., necktie, wheel and hat). Furthermore, They found that the bias in word embeddings parallels the biases in the zero-shot vision-language models and, as such, verifies the claims that biases from language models are passed to vision models learned from it.

According to [20], model bias can be classified based on the dependencies between the data attributes. Such attributes can be dependent or independent of each other. Correlations between dependent attributes, such as high cheekbone and sex, are called intrinsic dependences, while the correlation between independent attributes, such as hair color and sex, is called a spurious correlation. The authors claim that debiasing methods are commonly focused on spurious correlations, requiring or not ground-truth labels, and use iterative methods that are computationally costly. Their method is supposed to address all these issues. They developed a method to learn the biases from text and image embeddings of CLIP and deploy their model as a post-processing stage after feature encoding and prior to cosine similarity classifiers.

The authors of [21] propose a calibration loss that minimizes the discrepancy between a pair of prompt embeddings that contain biases. By doing so, the embeddings of both male and female versions of a given prompt should be similar, and as such, debiased. Only focusing on the text embedding was sufficient to improve group robustness of zero-shot models, corroborating the claims of parallelism between VLMs and LLMs biases.

Birhane et al. [12] explored the effect of dataset scaling on racial bias. They used several openCLIP models and the Chicago Face Dataset (CFD) using negative terms (criminal, thief) and non-human (gorilla) classes to discover social biases. Among other findings, they concluded that dataset scaling increased the number of Latino and black faces being labeled with criminal classes on the larger models.

This brief review indicates that VLMs are promising for zero-shot classification, but the (possibly uncurated) use of large datasets might lead to bias in human-related tasks. Next, we present our approach to explore several VLMs for gender classification.

III. Our Methodology

This work explores the limits of zero-shot gender classification of images from the FairFace dataset [22] using VLMs. We aim to understand better the impact of data and model scaling of VLMs such as CLIP and the potential harm of bad class design in exacerbating the inherent societal bias contained within large vision language models freely trained using the web.

This section describes our methods and the setup used to conduct such experiments. For the proper measurement of model scaling, we selected Vision Transformer (ViT) models of distinct sizes trained on the same data. As for data scaling, we compared the performance of a group of models with the same architecture but trained on distinct data sources with varying dataset sizes and number of samples seen. To address the impact of prompt design in zero-shot gender classification tasks, we propose a prompt tuning technique using an ensemble of auxiliary adjectives along with the original gender class labels. These steps are detailed next.

A. Models, Data sources, and Baseline

The core idea of CLIP [1] is to use a pair of image and text encoders trained so that paired data (image/text) should generate aligned feature vectors regarding the cosine similarity.

Clearly, there are several choices for the encoders and training datasets, which can strongly impact the results.

Although some convolutional image encoders were tested in CLIP, the best results were obtained with ViTs. Different ViTs with an increasing number of parameters were tested in [1], such as ViT-B (base) and ViT-L (large). Even larger models such as ViT-H (huge) and ViT-g (giant) were trained in [10]. In this work, we consider all such variations.

For each architecture, we can choose a variety of training datasets, conventionally called *data sources*, containing pairs of images and the corresponding textual descriptions. The original CLIP paper [1] briefly describes the data scraping process used to obtain their private WebImageText (WIT-400M) dataset, which contains 400 million image/text pairs collected from the Internet. LAION-400M [15] is an open-source dataset with over 400 million text/image pairs scraped from the Common Crawl [23] web dump, filtered using CLIP models for image text similarity match. Based on this initial effort, a larger version of the dataset was constructed, consisting of 5.85 billion CLIP-filtered image-text pairs, of which 2.32B contain text in English. Both 400M and the 2B subset are employed at openCLIP [10] and used in this work. More recently, Datacomp [14] was introduced as a testbed for dataset experiments centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl [23], which they called CommonPool. An automatically filtered version based on CLIP similarity with 1 billion pairs was also released, called DataComp-1B, and results shown in [14] indicate that using such "curated" dataset yields better zero-shot classification results for some datasets than using larger data sources for pre-training.

For each desired combination of architecture and data source, we must also define the total number of samples that are seen during training, which relates to the number of iterations. Clearly, using large models with massive training data requires extensive computational power and time. In this work, we selected pre-trained combinations[1] used in [10]. Table I provides the full list of models (with the corresponding number of parameters), data sources, and number of samples seen in each training session along with the batch size, as reported in [10]. The "openai" dataset refers to the OpenAI's WIT dataset used to train the original CLIP models.

For the performance evaluation of CLIP models in zero-shot gender classification, we employed the FairFace [22] dataset. This face image dataset contains 108,501 images with a balanced number of samples for each race, gender, and age group. The samples were collected from the YFCC-100M Flickr [24] dataset and labeled with seven race groups, nine age groups, and two gender groups, noting that labels were defined by a third-party group and not by self-attribution.

As for the metrics, overall performance was evaluated using the accuracy, a simple yet efficient metric well-suited for balanced data such as the FairFace dataset. To measure gender bias, we computed the male-female accuracy difference (called GDif), noting that lower absolute values indicate less bias. To measure racial and age bias, we used Gap, defined as the difference between the average and worst-Group accuracies [20].

[1]https://github.com/mlfoundations/open_clip

TABLE I
MODEL ARCHITECTURES AND TRAINING DATASETS

Architecture (#P)	Data source	Data source size	# samples	batch size
ViT-B-16 (149M)	openai	400M	13B	32k
	laion400m_e32	407M	13B	33k
	datacomp_xl_s13b_b90k	1B	13B	90k
	commonpool_l_s1b_b8k	1.28B	1B	8k
	laion2b_s34b_b88k	2.3B	34B	88k
ViT-B-32 (151M)	datacomp_m_s128m_b4k	14M	128M	4k
	commonpool_m_s128m_b4k	128M	128M	4k
	openai	400M	13B	32k
	laion400m_e32	407M	13B	86k
	datacomp_xl_s13b_b90k	1B	13B	90k
	laion2b_s34b_b79k	2.3B	34B	79k
ViT-L-14 (428M)	openai	400M	13B	32k
	laion400m_e32	407M	13B	86k
	datacomp_xl_s13b_b90k	1B	13B	90k
	laion2b_s32b_b82k	2B	32B	82k
	commonpool_xl_s13b_b90k	12.8B	13B	90k
ViT-H-14 (986M)	laion2b_s32b_b79k	2B	32B	79k
ViT-g-14 (1,367B)	laion2b_s34b_b88k	2B	34B	88k

As the baseline zero-shot classification strategy, we follow the approach described in [1]. The input image is fed to the image encoder and L2-normalized, generating an image feature vector $\boldsymbol{I}$. For each desired class y_i, a *canonical name* `class_i` is defined, and a text prompt in the form "`A photo of a/an <class_i>`" is fed to the text encoder and normalized, generating features $\boldsymbol{T}_i$. The chosen category $\hat{y} = y_{\hat{i}}$ is the one that maximizes the cosine similarity s_i, i.e.,

$$\hat{i} = \arg\max_i s_i, \quad s_i = \boldsymbol{I}^T \boldsymbol{T}_i. \tag{1}$$

For binary gender classification, we use only two categories. The canonical class names were `man` and `woman`.

B. The effect of model and data scaling

To evaluate model scaling, we choose a fixed data source and scale the ViT model. Since the only available dataset for all model checkpoints was LAION-2B, this dataset was chosen in the analysis. All tested models are listed in Table I.

For data scaling, we must choose a constant model architecture and scale the training data. Although the results of [10] indicate a scaling-law improvement based on the dataset size, recent findings [14] indicate that data quality rather than size can be more important. Hence, it is essential to evaluate the effect of filtered data sources on gender classification. In our experiments, we explored both filtered DataComp and unfiltered CommonPool data sources, as well as a range of data sources with varying sizes as listed in Table I. We performed the analysis using ViT-B-32, ViT-B-16, and ViT-L-14, for which pre-training with several data sources is available.

C. The effect of auxiliary adjectives and ensembles

The core idea behind zero-shot classification is to find the best alignment between a set of candidate text prompts and one input image. Since training datasets are massive and mostly

web-crawled, it is hardly possible to *guess* which textual description was used for a given image.

We hypothesize that more complete text prompts, including adjectives unrelated to gender but that might have been used to describe images of human faces, can improve zero-shot classification results. More precisely, we explore prompt templates that also include age and race information besides the gender class names.

By combining a list of age, race, and gender words, the list of possible prompts was generated from all possible combinations. The prompt was built based on the aforementioned prefix and a combination of race + gender, and age + race + gender. We also added prompts with only the gender classes to leverage the original prompt. The list of terms for each attribute is provided below.

```
age_list = ['young', 'old', 'middle-aged']
race_list = ['black', 'white', 'hispanic',
    'latino', 'indian', 'asian', 'arabic']
gender_list = ['woman', 'man']
```

By combining each list, we were able to create distinct prompts such as `a photo of a woman` (baseline gender prompt, or GP), `a photo of a black woman` (race + gender, or RGP), or `a photo of a young black woman` (race + age + gender, or RAGP).

In the latter two prompt-tuning strategies (RGP and RAGP), there are several possible prompts for the same class, and we propose a *score aggregation* strategy to define the winning class. Let $s_i^j = \boldsymbol{I}^T \boldsymbol{T}_i^j$ denote the cosine similarity between the image feature $\boldsymbol{I}$ and the text feature $\boldsymbol{T}_i^j$ related to the J^{th} prompt for class i. We propose to compute the average of the top-k scores for each class i as the final class-related scores.

Consider that s_i^j are given in descending order w.r.t. j for each class i. The aggregated score $\bar{s}_i^k$ is given by

$$\bar{s}_i^k = \frac{1}{k} \sum_{j=1}^{k} s_i^j. \tag{2}$$

For $k = 1$, this strategy simplifies to a Winner Takes All (WTA) using cosine similarity of *all* prompts using Eq. (1). We evaluate the results of the WTA approach and other choices for k.

Both model and data scaling experiments were replicated with the addition of the aggregation techniques, selecting the best method for each model and data source.

IV. Results and Discussion

This section discusses the results obtained from our experiments on data and model scaling, prompt tuning, and aggregation techniques. We used multiple openCLIP models to zero-shot classify images of faces from FairFace into one of the two target gender labels and evaluated its accuracy on each age and ethnicity subgroup. Based on previous work, we expected to see a progressive improvement in performance as data and models scaled. Also, we aimed to improve performance by employing new prompt tuning and aggregation techniques.

A. Model and Data scaling

We start by analyzing the impact of model scaling on gender classification. Different ViT architectures with the same LAION-2B training dataset were evaluated over the FairFace dataset. Table II shows the overall and per-gender accuracy values. The accuracy for the gender classification of samples grouped by race can be seen in Table III, and grouped by age in Table IV – the per-group average values are also reported. We can see a clear improvement in almost all metrics as model architecture scales, corroborating the scaling laws presented by the work of [10]. ViT-H-14 had the best overall accuracy, while ViT-g-14 had the lowest GDif score. We can see a negative GDif value for all model architectures except ViT-H-14, indicating higher accuracy for the female gender for most models. For the ethnic subgroups, all models performed poorly for black people. As for the age subgroups, the worst group was in the 0-2 age span, which is expected since gender-specific traits are still underdeveloped.

TABLE II
Accuracy by gender for each model architecture with LAION-2B data-source.

Model	accuracy	Male	Female	GDif
ViT-B-16	0.9271	0.9011	0.9562	-0.0551
ViT-B-32	0.9104	0.8819	0.9425	-0.0606
ViT-L-14	0.9469	0.9275	**0.9686**	-0.0411
ViT-H-14	**0.9546**	**0.9589**	0.9498	0.0091
ViT-g-14	0.9542	0.9527	0.9558	**-0.0031**

TABLE III
Accuracy by race with LAION-2B data source. Best result in bold.

Model	Avg.	E. Asian	White	Latino	S. Asian	Black	Indian	M. Eastern	Gap
ViT-B-16	0.9268	0.9213	0.9468	0.9421	0.9187	0.8760	0.9261	0.9570	0.0508
ViT-B-32	0.9101	0.9071	0.9362	0.9248	0.8996	0.8451	0.9103	0.9479	0.0650
ViT-L-14	0.9468	0.9471	0.9650	0.9507	0.9449	0.8965	0.9479	0.9760	0.0503
ViT-H-14	**0.9546**	**0.9561**	**0.9679**	0.9612	**0.9541**	**0.9113**	**0.9558**	0.9760	**0.0433**
ViT-g-14	0.9544	**0.9561**	0.9635	**0.9649**	0.9527	0.9100	0.9545	**0.9793**	0.0444

TABLE IV
Accuracy by age for each model with LAION-2B data source. Best result in bold.

Model	Avg.	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	70+	Gap
ViT-B-16	0.9069	0.7688	0.8024	0.8620	0.9552	0.9678	0.9667	0.9623	0.9533	0.9237	0.1381
ViT-B-32	0.8853	0.7186	0.7743	0.8129	0.9479	0.9614	0.9579	0.9435	0.9190	0.9322	0.1667
ViT-L-14	0.9339	0.8392	0.8473	0.8865	0.9712	0.9781	0.9808	0.9736	0.9626	**0.9661**	0.0947
ViT-H-14	**0.9421**	0.8593	0.8776	**0.9102**	**0.9739**	0.9815	0.9786	0.9686	**0.9720**	0.9576	**0.0828**
ViT-g-14	0.9378	0.8191	**0.8791**	0.9035	0.9736	**0.9820**	**0.9823**	**0.9749**	0.9688	0.9576	0.1187

The results for the data scaling experiment are reported in Table V, and data sources are listed in increasing order regarding their size. We note that using OpenAI's WIT private dataset outperformed all other sources in almost all subcategories for all model architectures despite being one of the smallest datasets. This finding is going in the opposite direction of [4], [14], noting that different datasets were evaluated. We cannot fully examine the reason for these results due to the private nature of OpenAI's WIT dataset; we can only know that complex scraping and filtering techniques were used to optimize these models.

TABLE V
DATA SCALING FOR THE VIT-B-16, VIT-B-32 AND VIT-L-14. DATA SOURCES ORDERED BY SIZE

datasource	accuracy	Male	Female	GDif
ViT-B-16				
openai	**0.9502**	**0.9432**	**0.9582**	**-0.0150**
laion400m_e32	0.9204	0.8940	0.9500	-0.0560
datacomp_l_s1b_b8k	0.7894	0.6628	0.9314	-0.2686
datacomp_xl_s13b_b90k	0.8664	0.8194	0.9190	-0.0996
commonpool_l_s1b_b8k	0.7875	0.8073	0.7652	0.0421
laion2b_s34b_b88k	0.9271	0.9011	0.9562	-0.0551
ViT-B-32				
datacomp_m_s128m_b4k	0.6089	0.5366	0.6900	-0.1534
commonpool_m_s128m_b4k	0.6282	0.6785	0.5717	0.1068
openai	**0.9360**	**0.9235**	**0.9500**	**-0.0265**
laion400m_e32	0.8916	0.8652	0.9213	-0.0561
datacomp_xl_s13b_b90k	0.8347	0.8011	0.8723	-0.0712
laion2b_s34b_b79k	0.9104	0.8819	0.9425	-0.0606
ViT-L-14				
openai	**0.9622**	**0.9605**	0.9642	**-0.0037**
laion400m_e32	0.9285	0.8923	**0.9692**	-0.0769
datacomp_xl_s13b_b90k	0.9177	0.8795	0.9607	-0.0812
laion2b_s32b_b82k	0.9469	0.9275	0.9686	-0.0411
commonpool_xl_s13b_b90k	0.9181	0.9244	0.9111	0.0133
laion2b_s34b_b79k	0.9104	0.8819	0.9425	-0.0606

These findings guided our next experiments by defining the best-performing combinations of models and data sources. As such, we will continue to explore how to improve these top models by experimenting with prompt tuning and similarity aggregation. For the next experiments, we will use pre-training with the OpenAI data source for ViT-B-32, ViT-B-16, and ViT-L-14. For ViT-L-14 and ViT-H-14, we will use LAION-2B, which is the only data source available.

B. The effect of auxiliary adjectives and ensembles

As described in Section III-C, we also performed experiments by augmenting the prompts using race and age-related adjectives. In the first experiment, we used a WTA strategy to find the winning class considering all prompts. The results shown in Table VI do not indicate a clear trend for the accuracy values, but it shows a consistent drop in the gender bias metric GDif (except for ViT-H-14).

TABLE VI
GENDER CLASSIFICATION COMPARISON BETWEEN PROMPT STRATEGIES.

Model	Prompt	Accuracy	Male	Female	GDif
ViT-B-16 openai	GP	0.9502	0.9432	**0.9582**	-0.0150
	RGP	0.9513	**0.9605**	0.9411	0.0194
	RAGP	**0.9532**	0.9541	0.9522	**0.0019**
ViT-B-32 openai	GP	0.9360	0.9235	**0.9500**	-0.0265
	RGP	0.9366	**0.9539**	0.9171	0.0368
	RAGP	**0.9396**	0.9463	0.932	**0.0143**
ViT-L-14 openai	GP	**0.9622**	**0.9605**	**0.9642**	-0.0037
	RGP	0.9592	0.9589	0.9595	**-0.0006**
	RAGP	0.9588	0.9603	0.9572	0.0031
ViT-H-14 laion2b	GP	**0.9546**	0.9589	0.9498	**0.0091**
	RGP	0.9518	0.9420	**0.9628**	-0.0208
	RAGP	0.9531	**0.9608**	0.9444	0.0164
ViT-g-14 laion2b	GP	0.9542	0.9527	0.9558	-0.0031
	RGP	0.9531	0.9504	**0.9560**	-0.0056
	RAGP	**0.9555**	**0.9567**	0.9541	**0.0026**

Finally, the last experiment aims to evaluate the impact of the aggregation strategy. Starting from the RAGP strategy, we used the original aggregation strategy presented in [1], [10] that consists of averaging and normalizing the textual features of *all* prompts related to a given class[2], called "OpenAI". We also used the proposed score-based top-k aggregation strategy given by Eq. (2) and experimented with different values for k. We noted that each model's optimal k value varies, and that $k = 10$ produces consistent results for all models.

TABLE VII
COMPARISON BETWEEN OUR TOP-k TECHNIQUE AND OPENAI'S AGGREGATION METHOD FOR RAGP.

Model	Mode	Accuracy	Male	Female	GDif
ViT-B-16	OpenAI	0.9540	**0.9610**	0.9461	0.0149
	top-10	0.9550	0.9537	**0.9564**	-0.0027
	top-17	**0.9557**	0.9563	0.9551	**0.0012**
ViT-B-32	OpenAI	0.9394	**0.9513**	0.9260	0.0253
	top-10	0.9407	0.9460	0.9347	0.0113
	Avg Sum	**0.9421**	0.9472	**0.9365**	**0.0107**
ViT-L-14	OpenAI	0.9618	**0.9603**	0.9636	**-0.0033**
	top-10	0.9621	0.9565	**0.9684**	-0.0119
	top-13	**0.9623**	0.9570	0.9682	-0.0112
ViT-H-14	OpenAI	0.9534	0.9498	**0.9576**	-0.0078
	top-10	0.9549	0.9542	0.9556	**-0.0014**
	top-7	**0.9554**	**0.9563**	0.9543	0.0020
ViT-g-14	OpenAI	0.9538	0.9475	0.9609	-0.0134
	top-10	0.9554	0.9492	**0.9622**	-0.0130
	top-2	**0.9566**	**0.9570**	0.9562	**0.0008**

Table VII shows the result using aggregation with "OpenAI" and top-k, for $k = 10$ and the per-model optimal k value. We note that the top-10 aggregation results are slightly but consistently better than OpenAI, while reducing gender bias by closing the gap between the accuracy of Male and Female subgroups (see the GDif metric). These results indicate that performing aggregation of similarities seems a better option for zero-shot gender classification than the aggregation of the embeddings.

Our experiments showed us that just raw power scaling is not enough to build fair and efficient VLMs, and only by benchmarking downstream tasks can we truly grasp the impact of these trends. The proper class design plays a significant role in mitigating spurious correlations, and we have only just begun to understand how we can probe and measure the intrinsic societal biases learned by weakly supervised vision-language models. Our work on zero-shot gender classification helps pave the way for a fair and efficient deployment of pre-trained models in sensitive, real-world solutions.

V. CONCLUSION

Multi-modal vision-language models showed promising zero-shot capabilities in a myriad of different downstream tasks without the need for any fine-tuning or retraining. Different studies have been conducted to better understand how these models were trained on what data, and how they could be improved and generalized by prompt engineering.

We observed the impact of data and model scaling for zero-shot gender classification using openCLIP, and how to employ simple prompt tuning methods to improve these results. We

[2]https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

chose the FairFace dataset so that we could leverage the solid baseline from recent studies that employed it. However, we acknowledge the lack of a broad exploration of different face datasets currently available, especially those that go beyond the binary specification of gender.

We explored the effect of auxiliary adjectives by improving the starting prompt "`A photo of a <man/woman>`" with the addition of both ethnic and age adjectives. Our approach improved the average accuracy by 0.003% while also closing the gap between Female and Male classification accuracy by 56,81

We followed by comparing the original OpenAI's feature-based aggregation strategy with our proposed score-based top k method. By averaging the top-k similarity scores, we were able to slightly improve performance whilst reducing gender bias by a considerable margin. For example, we improved the average accuracy by 0.198% but strongly improved gender bias, reducing GDif by 79,50%. These results showed us the potential of simple textual prompt tuning techniques and the power of the proper class design in deploying robust and fair Vision-Language Models to perform sensitive human detection tasks. With our classification accuracy measurements, we were able to efficiently explore data and model scaling and the impact of prompt tuning techniques. However, this simplistic approach may not completely cover all the inner details of how VLMs work and how they can be actively learning spurious social correlations.

Our narrowed scope can only positively impact the end user by aiding its classification with the proper class design. We still lack a generalized probe of social biases for vision-language models and must discuss how to cut it at its source by exploring techniques and guidelines for both users and developers of such powerful machines. We pledge for a more transparent addressing of what kind of data is used in the automated training processes, what are the sources of such data, and what are the tools currently employed to prevent the proliferation of harmful correlations done by these large models freely trained using the whole wide web. While we cannot have a say in these matters, we can only hope to better assist the end user in avoiding the pitfalls introduced by these large volumes of automated scrapes from the Internet.

We aim to expand this line of work by exploring other face datasets, focusing on those with self-identification of personal information such as race, gender, and age. To the best of our knowledge, there is no robust and publicly available face dataset that goes beyond the simplistic definitions of binary genders, providing an interesting line of future works.

Given the private nature of both models and datasets, this work aims to help shed some light on the impact of data and model scaling, while showing the critical role of textual prompts in order to correctly identify images from faces, a simple task that the promising results can persuade users to real-world deployments. It is for the sake of all marginalized and endangered societal groups that this kind of research has become of the utmost importance, given the rapidly growing interest in large vision-language models with out-of-the-shelf capabilities. We hope that by expanding this line of work, we can help raise societal awareness of the fair and careful use of such powerful tools.

REFERENCES

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[3] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision*. Springer, 2020, pp. 776–794.

[4] C. Schuhmann *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[5] K. Zhou *et al.*, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[6] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 816–16 825.

[7] C. Simon, P. Koniusz, and M. Harandi, "Meta-learning for multi-label few-shot classification," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3951–3960.

[8] W. Berrios *et al.*, "Towards language models that can see: Computer vision through the lens of natural language," *arXiv preprint arXiv:2306.16410*, 2023.

[9] R. Zhang *et al.*, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.

[10] M. Cherti *et al.*, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818–2829.

[11] A. Hundt *et al.*, "Robots enact malignant stereotypes," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 743–756.

[12] A. Birhane *et al.*, "The dark side of dataset scaling: Evaluating racial classification in multimodal models," *arXiv preprint arXiv:2405.04623*, 2024.

[13] M. Hall *et al.*, "Vision-language models performing zero-shot tasks exhibit gender-based disparities," *arXiv preprint arXiv:2301.11100*, 2023.

[14] S. Y. Gadre *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] C. Schuhmann *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[16] O. R. et al., "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[17] S. Parisot, Y. Yang, and S. McDonagh, "Learning to name classes for vision and language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 477–23 486.

[18] X. Zhou *et al.*, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 350–368.

[19] B. Li *et al.*, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[20] S. Dehdashtian, L. Wang, and V. N. Boddeti, "Fairerclip: Debiasing clip's zero-shot predictions using functions in RKHSs," *arXiv preprint arXiv:2403.15593*, 2024.

[21] C.-Y. Chuang *et al.*, "Debiasing vision-language models via biased prompts," *arXiv preprint arXiv:2302.00070*, 2023.

[22] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[23] "Common crawl," 2023. [Online]. Available: https://commoncrawl.org/overview

[24] B. Thomee *et al.*, "YfCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.