A study on the distribution of social biases in self-supervised learning visual models

Kirill Sirotkin, Pablo Carballeira, Marcos Escudero-Viñolo
Video Processing and Understanding Lab, Univ. Autónoma de Madrid, 28049, Madrid, Spain
{kirill.sirotkin, pablo.carballeira, marcos.escudero}@uam.es

Abstract

Deep neural networks are efficient at learning the data distribution if it is sufficiently sampled. However, they can be strongly biased by non-relevant factors implicitly incorporated in the training data. These include operational biases, such as ineffective or uneven data sampling, but also ethical concerns, as the social biases are implicitly present—even inadvertently, in the training data or explicitly defined in unfair training schedules. In tasks having impact on human processes, the learning of social biases may produce discriminatory, unethical and untrustworthy consequences. It is often assumed that social biases stem from supervised learning on labelled data, and thus, Self-Supervised Learning (SSL) wrongly appears as an efficient and bias-free solution, as it does not require labelled data. However, it was recently proven that a popular SSL method also incorporates biases. In this paper, we study the biases of a varied set of SSL visual models, trained using ImageNet data, using a method and dataset designed by psychological experts to measure social biases. We show that there is a correlation between the type of the SSL model and the number of biases that it incorporates. Furthermore, the results also suggest that this number does not strictly depend on the model's accuracy and changes throughout the network. Finally, we conclude that a careful SSL model selection process can reduce the number of social biases in the deployed model, whilst keeping high performance. The code is available at https://github.com/vpulab/SB-SSL.

1. Introduction

Supervised Deep Learning models currently constitute the state-of-the-art in the fields of computer vision and natural language processing. However, the recent developments [17, 24] in the field of Self-Supervised Learning (SSL) —a type of unsupervised learning, are slowly closing the performance gap gained via human guidance usually provided in the shape of target labels. SSL methods aim at solving a pre-formulated pretext task—defined by automatically generated labels, whose solution is expected to require high-level understanding of the data in order to learn descriptive feature embeddings with strong transferability potential.

Human social biases are a well-studied and, in some cases, numerically quantifiable phenomenon [23] that causes unjustified prejudices against social groups based, among others, on aspects such as age, gender and race. Whereas one cannot assign prejudices or preferences to deep learning approaches as these are highly subjective characteristics attributed solely to humans, deep learning methods can wrongly correlate certain concepts if the labeled training data distribution is biased itself [45]. In practice, this leads to the replication of social biases. Several cases have been studied and reported, including: an incorrect gender prediction based on the contextual cues (i.e., location - kitchen, office), rather than on visual evidence associated with the described person [30], a fewer number of automatic high-paying job recommendations for female candidates than for male ones [20] and a promotion of biased suggestions in the dating/political decision-making context [7]. Anticipating these situations, institutional initiatives are being developed internationally to extinguish social biases from the training data, as declared in the Ethics Guidelines for a Trustworthy AI issued by the European Commission, and regulate the use of machine learning methods with potential human implications, as stated in numerous US bills [2, 3, 18] and the legislative documents of other countries [1, 4–6].

Previously, it was demonstrated that supervised learning models are prone to implicitly learn biases from the datasets containing them [9, 27, 30], as these human biases are encapsulated in the target labels. For instance, it has been shown that the earlier versions of ImageNet [21] exposed an imbalanced distribution regarding skin colors, ages and genders, leading to the under-representation of certain groups [43]. Furthermore, datasets collecting raw comments scraped from the web [10, 40] contain explicit biases against certain social groups [31].

SSL approaches, being unsupervised, are expected to be

unaffected by biases-bearing labels. However, as they require large amounts of training data that often prevents its curation, it is not unlikely that the data itself contains some social human biases. In fact, results for a recent study [37], suggest that two of the state-of-the-art unsupervised learning models also contain association biases learned from the data, only in this case it cannot be explained by the choice of class labels, as the unsupervised models do not leverage this information in the training process. This result indicates that at least one top performing SSL model [15] might implicitly learn social biases while training to solve the targeted pretext task. Hence, data should be handled carefully, as the neural network's capacity to avoid inadvertent perpetuation of undesirable social biases is an important quality to consider, alongside classification accuracy, in the design of deep learning models.

This paper addresses this phenomenon and attempts to answer the questions: what is the origin of the biases in the SSL setting? What affects the model's proneness to learn an implicit social bias? What is the relationship between the model's accuracy and the biases it learns? Whereas a preliminary work addresses the first question and hypothesizes on the origins of implicit social biases in a couple of unsupervised models [37], to our knowledge, this is the first attempt to study a wider and more varied set of SSL models.

In particular, the contributions of this paper are:

- We study the association biases acquired by 11 SSL models that share the same ResNet-50 [29] architecture, and vary in terms of pretext task and, thus, their accuracy after transfer learning. The results of this study suggest that the nature of the pretext task influences the number and nature of incorporated biases, and that contrastive models are more prone to acquire biased associations that are implicit in the data.

- We also perform an analysis of biases acquired in the embeddings at different layers of the models, showing that the number and strength of the biases vary at different model depths. The results of the per-layer analysis suggest that a careful consideration of bias in transfer learning applications can improve the trade-off between bias and accuracy, as the accuracy achieved using embeddings from highly-biased layers is not far from the accuracy achieved by a less-biased embeddings layer.

2. Related work

2.1. Measuring biases of computer vision models

Many of the existing methods for measuring the biased associations are based on the Implicit Association Test (IAT) [23] that measures the differential relationships between a target concept and an attribute. The IAT measures the difference in reaction time of a respondent when correlating concepts and attributes for which biased associations are prone to exist and for which they are not. For instance, a biased test subject strongly associating a concept *flower* with an attribute *pleasant* takes less time to correlate verbal or visual stimuli representing them rather than correlating stimuli representing a concept *insect* with an attribute *pleasant*.

Until recently, association bias tests were mostly used for Natural Language Processing (NLP) [11, 19, 35, 39, 45], but a recent work has extended the Word Embedding Association Test (WEAT) [11] to the image domain, thereby, making it possible to quantify association biases in computer vision models. This approach, named Image Embedding Association Test (iEAT), measures the differential association of the target concepts X and Y with the attributes A and B based on the image embeddings obtained by feeding images representing these concepts and attributes to a trained deep learning model. For instance, let the chosen target concepts be *insect* (X) and *flower* (Y) and the attributes be *unpleasant* (A) and *pleasant* (B). Then, the association test will measure the strength of correlation between *insect* and *unpleasant*, and *flower* and *pleasant* based on the cosine distances between the embeddings of X, Y, A and B. A more detailed explanation is given in Section 3.

2.2. Image embeddings via self-supervised learning models

In this paper, we refer to an image embedding as the features extracted at a given layer of a deep learning model when a particular image is fed to it. These embeddings are accepted as a representative description of the image —subjected to the training target. Usually, one can expect that, at a given layer and for a given architecture, the higher the performance of the learned model is, the more representative the embeddings will be. A common way to obtain image embeddings is by using a network trained in the supervised mode [29, 38]. Alternatively, SSL models can be used if images are to be represented in label scarce scenarios—as medical data requiring expert annotations or data acquired using devices capturing at non-visual modalities. SSL methods, instead of being trained for a label-driven task can be trained by using objectives such as a simple geometric task [22, 32], pseudo labels generated through automatic clustering [12, 44], or promoting proximity of "similar" data points in the feature space [14–17, 24, 28]. These objectives are commonly known as pretext tasks and can be used to arrange SSL models into the following three groups.

Geometric models One of the most straightforward approaches to defining a pretext task is applying a geometric transformation to an input image and training a network to solve it. The three geometric pretext tasks considered in

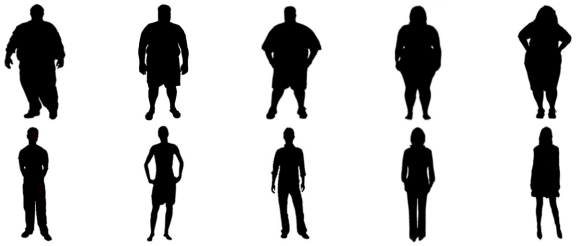


Figure 1. Images representing the concepts of overweight people (at the top) and thin people (on the bottom). The images are used to identify a weight-valence bias in the iEAT and original IATs [23].

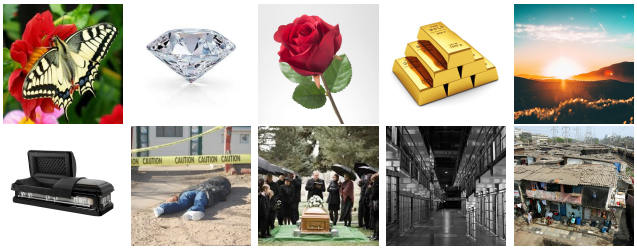


Figure 2. Images representing the concepts of valence: pleasant (at the top) and unpleasant (on the bottom). The images correspond to the verbal stimuli commonly used to describe the valence concepts [23].

this paper are rotation prediction [32], relative patch location prediction [22] and jigsaw puzzles [36]. The rotation prediction pretext task randomly applies one out of 4 rotations: 0°, 90°, 180°, 270°, to each training image sample and trains the network to predict which rotation was applied to a given image. On the other hand, a model trained to predict patch locations is based on randomly sampling two close regions from an input image and training the network to predict their relative spatial location. Finally, when a jigsaw puzzle strategy is followed, the image is divided into tiles, that are then randomly shuffled. Then, the network is trained to predict their original arrangement.

Clustering-based representation learning A more sophisticated approach to deep unsupervised learning is based on the classical clustering methods that are used to group unlabeled data into clusters according to some homogeneity criteria. An obvious way to incorporate clustering into the pretext task formulation is to perform clustering after each model update step. The generated labels are then used as pseudo-labels to evaluate the model in a supervised manner. These labels would, in turn, change the embeddings at the next step as the newly generated labels may differ from the labels at the previous step. This is the strategy followed by Deep Clustering (DC) [12], that suffers from instability during the training process due to the random permutation of labels at each step. To tackle the issue of labels permutation and instability, Cluster Fit [42] relies on using a teacher network to define the pseudo-labels. Differently, in Online Deep Clustering (ODC) [44] the labels are updated using mini-batches and this process is integrated into the model update. This way, the embeddings and labels evolve together and the instability inherent in DC is eliminated.

Contrastive models Top performing SSL models are driven by pretext tasks using contrastive losses [26]. Although exact implementations vary from model to model, the main idea remains the same: to learn representations that map the *positives* close together and push apart the *negatives*. The *positive* samples might be chosen based on modifications of patches in the same image or applying different augmentations obtained from the same image.

Non-Parametric Instance Discrimination (NPID) [41] treats each input image (instance) as belonging to a unique class and trains the classifier to separate between each instance via the noise-contrastive estimation [25]. The motivation for it comes from the observation that supervised learning approaches return similar embeddings for related images. Specifically, it is often the case that the second top scoring predicted class at the end of the model is semantically close to the first one following a human interpretation. Therefore, the network is expected to learn the semantic similarity between classes without explicitly having it as the objective.

Momentum Contrast (MoCo) [28] leverages a dynamic dictionary where a *query* and associated *keys* represent image encodings obtained with an encoder network. If a *query* and a *key* come from the same image, they are considered to be a *positive* pair, otherwise a *negative* one. The *queries* and the *keys* are encoded by separate networks and the *key* encoder is updated as a moving average of the *query* encoder, enabling a large and consistent dictionary for learning visual representations.

Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [14], building on the principles of contrast learning, introduces a series of design changes that allow it to outperform MoCo [28] not requiring a memory bank. Among these changes are a more careful choice of data augmentation strategies, addition of a non-linearity between the embeddings and the contrastive loss, and increased batch sizes and the number of training steps. Further improving on the results of SimCLR [14], the second version of Momentum Contrast model (MoCo v2) [16] acknowledges its efficient design choices and takes advantage of an MLP projection head and more data augmentations.

Bootstrap Your Own Latent (BYOL) [24] reaches a new state-of-the-art on ImageNet linear classification while avoiding one of the greatest challenges that other contrastive models face: a need for negative pairs. BYOL circumvents this problem by generating the target representations with a randomly initialized model and then using them for its on-

line training. By iteratively updating the target network, the online network is expected to learn better and better representations.

Finally, SwAV [13] describes a hybrid clustering-contrastive method that avoids the computation of pairwise distances between positive and negative samples by clustering the data in consistency-enforced clusters of the different image augmentations. Thereby, defining positive samples according to cluster memberships and reducing the distance storage requirements of the other contrastive methods.

3. Methodology

Given the set of SSL models described in Section 2.2, we apply the iEAT framework [37] introduced in Section 2.1 to each model and investigate the presence of association biases. iEAT takes a set of input embeddings $\{x, y\}$ of images representing the target concepts (X, Y) and a set of input embeddings $\{a, b\}$ representing the measured attributes (A, B). For instance, the target concepts *overweight people* and *thin people* are represented by the example images on Figure 1, while the attributes *pleasant* and *unpleasant* can be visualized by the images representing the *valence* concept (see examples in Figure 2). The null-hypothesis tested by iEAT states that X=*overweight people* embeddings are as similar as Y=*thin people* embeddings to (A, B)=(*pleasant*, *unpleasant*) embeddings, or that the dissimilarities are alike. The rejection of the null-hypothesis would mean that one target concept is more correlated with one attribute than the other target concept, thus, detecting an association bias. iEAT tests the null-hypothesis by a permutation test and a metric quantifying the differential association $s(X, Y, A, B)$, defined as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B), \quad (1)$$

where:

$$s(t, A, B) = \underset{a \in A}{mean} \cos(t, a) - \underset{b \in B}{mean} \cos(t, b) \text{ for } t = \{x, y\}. \quad (2)$$

The permutation test randomly shuffles the labels of the set of embeddings representing target concepts (X, Y), creating 10000 randomly permuted sets (or the maximum number of permutations allowed by the set size). Then, the differential association (Eq. 1) of each one of these permuted sets is measured. The p-value collects the percentage of permuted sets resulting in a larger or equal differential association than the original set. The null-hypothesis can be rejected (and thus a bias detected) with high probability if the p-value is below a certain threshold. The strength of the bias can be measured as the effect size (d-value) —measures the separation between the two distributions of the distances of the two sets of target concept samples to the attribute samples [11]:

$$d = \frac{\underset{x \in X}{mean}\, s(x, A, B) - \underset{y \in Y}{mean}\, s(y, A, B)}{\underset{t \in X \cup Y}{std}\, s(t, A, B)}. \quad (3)$$

The full bias-detection pipeline for a given network model, therefore, consists of the extraction of deep feature embeddings of 4 image sets representing 2 target concepts and 2 attributes (i.e., Office-Home vs. Male-Female) with the same model, and running the permutation test described above on these embedding sets.

We evaluate the SSL models on the data provided by the authors of the iEAT framework[1] [37]. This data encompasses visual stimuli for sets of target concepts such as race, gender and age. The dataset contains 3 to 55 psychologists-selected images per concept taken from well-established IAT tests [23], CIFAR-100 dataset [33] or the web. Following the approach outlined above, we collect the p- and d-values representing the likelihood and strength of each of the 39 association biases proposed in the iEAT framework (see supplementary material for a complete list).

All models used in this paper were trained on ImageNet2012 [21] and share the same backbone architecture: ResNet-50 [29]. The weights for the pretrained networks were taken from the OpenSelfSupervised Framework[2] and VISSL[3] with training hyper-parameters listed in Table 1. We evaluate embeddings obtained from the first max pooling layer (layer 1 hereinafter), as well as the embeddings obtained after each ResNet block (layers 2-5) and the final Global Average Pooling (GAP). To achieve a more comprehensive overview of the presence of biases in the SSL models, we do not limit our choice to the state-of-the-art architectures and select networks that conceptually represent different approaches for SSL: Rotation prediction (Rotation) [32], Relative patch location prediction (Relative Location, RL) [22], Jigsaw puzzles [36], SwAV [13], ClusterFit [42], ODC [44], NPID [41], MoCo_v1 [28], MoCo_v2 [16], SimCLR [14] and BYOL [24], as well as a randomly initialized ResNet-50 (random) and a fully supervised ResNet-50 (supervised) [29].

4. Experimental results

This section summarizes the results of the bias detection on the deepest ResNet-50 embeddings, as well as on embeddings of its intermediate and shallow layers.

As stated in Section 3, the p-values are obtained with the permutation test. Due to the inherent randomness of this test, we repeat each experiment three times to confirm the consistency of the results and average the obtained p- and d-values (statistical error results are provided in the supplementary material). Moreover, we evaluate three instances

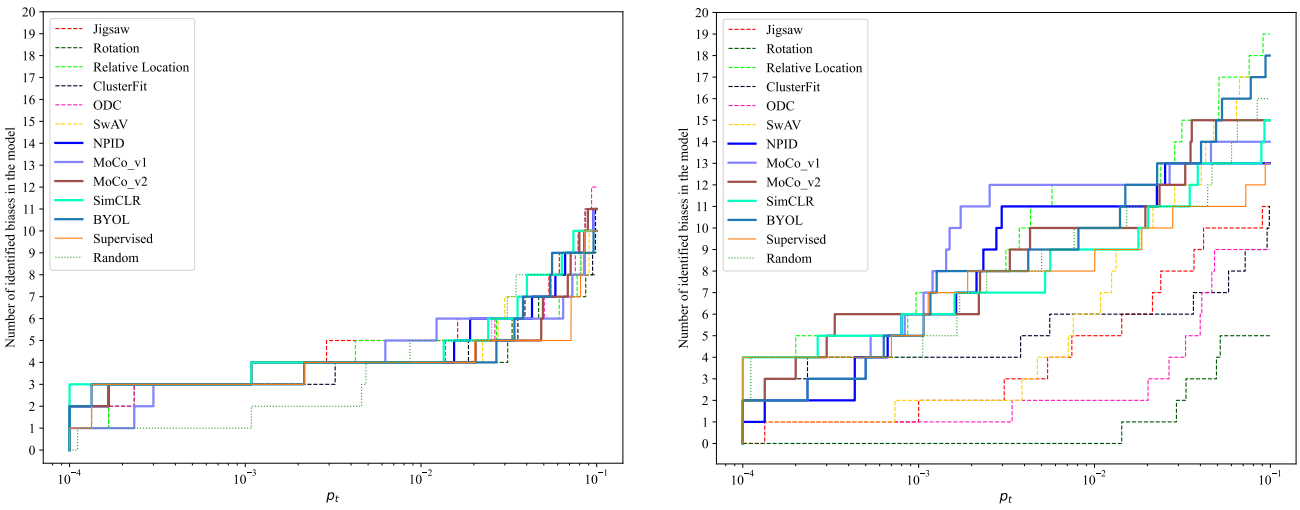
[1] distributed under CC BY-NC-SA 4.0 license

[2] https://rb.gy/iz1xlg

[3] https://rb.gy/rveiso

Table 1. Hyperparameters used to train the SSL models and the fully supervised ResNet-50.

	Jigsaw	RL	ClusterFit	Rotation	NPID	ODC	MoCo v1	SimCLR	MoCo v2	BYOL	SwAV	Sup.
Batch size	256	512	256	512	256	512	256	4096	256	4096	4096	256
Epochs	105	70	105	70	200	440	200	200	200	200	200	90
Base lr	0.1	0.2	0.1	0.2	0.03	0.06	0.03	0.3	0.03	0.3	0.3	0.1
ImageNet accuracy (best layer)	48.57	49.31	53.63	54.99	56.61	57.70	61.02	66.61	67.69	71.61	73.85	74.12



(a) 2nd ResNet block

(b) GAP ResNet layer

Figure 3. Number of biases for different values of p_t. Biases detected for lower values of p_t are statistically more significant. Contrastive models are plotted with thick solid lines, geometric models and clustering-based models with dashed lines.

of the random ResNet-50 model, in order to account for the random weight initialization. Thus, the p- and d-values reported for the random model are averaged for three instances and three permutation tests.

4.1. Bias-analysis for the GAP embeddings

The first analysis of the presence of social biases on different SSL models is performed using the embeddings at the deepest layer of the CNN architecture, i.e., after the GAP layer of ResNet-50, (GAP embeddings hereafter). GAP embedings generally convey a high performance in transfer learning scenarios (although not necessarily the highest). Thus, given the reduced dimensionality of the GAP embeddings with respect to those extracted at previous layers (which leads to faster training of a classifier on top of them) they are a common choice for transfer learning applications.

A bias is considered to be present in the model if its statistical significance, measured with the p-value yielded by the permutation test, is below a certain threshold p_t. Given the absence of a universally correct p_t value, a possible approach to categorize bias detection is to split them into groups of significance, as done in previous works [37]. Here, we take a similar approach, and explore the biases acquired by a model for p_t values in the $[10^{-4}, 10^{-1}]$ interval —which sits in the range of high statistical significance. Figure 3b shows the number of biases found at the GAP embeddings of the thirteen considered models with respect to the p-value threshold. From Figure 3b one can observe a clear separation in the number of detected biases between two groups of models: contrastive SSL models yield more biases than geometric/clustering-based models, with the exception of the RL model. For example, at $p_t = 10^{-2}$ no bias is detected for the rotation model and only 2 biases are detected for ODC, while the number of biases for contrastive models ranges from 8 to 12. This holds for any value of the threshold, showing the reliability of this conclusion. This is also shown on the large differences in the number of acquired biases among the three groups of models depicted in Figure 4. Especially for the *intersectional* biases (the most common ones) detected at the GAP embeddings. Results for all embeddings and biases are in the supplementary material.

We quantitatively assess the difference in the number

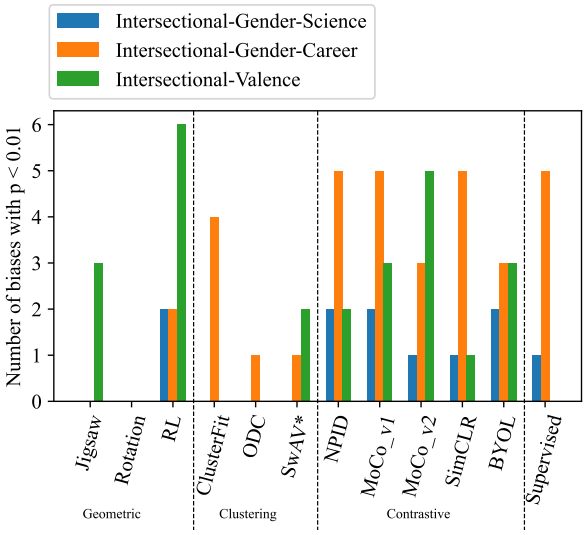


Figure 4. Numbers of intersectional biases detected in the embeddings of the Global Average Pooling layer with $p_t < 0.01$. Note that a hybrid clustering-contrastive model SwAV [13] is labeled as a clustering method for better readability of the figure.

of biases acquired by contrastive and non-contrastive models, by a statistical analysis of the biases present in two equal-sized sets of models: i) contrastive (NPID, SimCLR, MoCo_v1, MoCo_v2, BYOL) and ii) non-contrastive (Jigsaw, Rotation, ODC, RL, ClusterFit):

1. Initially, p_t is set to 0, and is gradually increased to 0.1 (i.e., p_t is moved in the direction of the x-axis of Figure 3b).
2. For each p_t value, we:
   2.1. Compute δ_{orig}: difference between the number of biases for the contrastive and non-contrastive sets.
   2.2. Permute the labels between the contrastive and non-contrastive sets, and, for each permutation k, compute δ_k: the difference between the number of biases for the two sets generated in the permutation.
   2.3. For each permutation k, check if ($\delta_{orig} < \delta_k$), and use this to estimate the probability of randomly permuted sets having more biases than the original sets.
3. Finally, we average the probabilities computed for each step of p_t.

This test yields a p-value of 0.049 that validates the premise by the rejection of the null-hypothesis: "contrastive models *are not more biased* than other models".

Finally, if we compare the bias data in Figure 3b with the accuracy data in Table 1, there is no direct link between the model's classification accuracy and the number of biases it incorporates. Indeed, BYOL and MoCo acquired more biases than the more accurate supervised model. Moreover, the least accurate model (RL) is one of the models that incorporates the highest number of biases.

4.2. Bias detections in the random model

Performing the bias analysis on the embeddings of the baseline random ResNet-50 model, we discovered a high number of biases. While it is implausible that a randomly initialized model can consistently contain certain biases, the bias detections themselves are possible in conditions that relate to the specific test data. We hypothesize that the bias detections in the random models come from the correlations in the test data caused by strong similarities between some low-level features. To test this hypothesis we randomly permute the pixels in the images representing two target concepts (i.e., Weapon, Tool), while leaving the images representing the two attributes (i.e., Black, White) unchanged, and repeat the bias test. This allows to remove the high-level visual concepts and most of the low-level features (such as textures) from the images, preserving only the distribution of pixel values. We perform this test for 13 social biases that are detected in the random model and observe that after the permutation of pixels, 11 of them remain present (see supplementary material for complete results) and, in some cases, even have lower p-values (Lincoln-Trump vs. Pleasant-Unpleasant).

4.3. Per-layer analysis

Anticipating that the strength and the number of biases varies for different layers of the network architecture due to increasing semantic interpretability in the internal CNN representations [8], the bias-detection procedure is carried out on the feature embeddings extracted from all ResNet blocks. Our findings are partially depicted in Figure 3, that presents the cumulative number of biases varying p_t in the embeddings of the 2nd ResNet block and Global Average Pooling layer. Figure 5 complements these results by summarizing the number and cumulative strength of biases in the embeddings extracted from all ResNet blocks. The strength of a bias refers to the d-value (Section 3), and the cumulative strength is the sum of the d-values of all detected biases.

The results shown on Figure 3a indicate that the biases are also detected in the feature embeddings of shallow ResNet layers that semantically resemble low-level features. However, the biases detected in the shallow layers mostly repeat for all models. For example, in Figure 3a four out of five biases at $p < 10^{-2}$ are common for all models and are race-related. On the other hand, the feature embeddings extracted from deeper layers, as shown on Figure 3b, result in more biases given the same value of the threshold. This statement holds for all the models except for the ro-

tation prediction model. Model-wise, the number of biases with the same degree of statistical significance is more uniform in the shallow layer embeddings and begins to differ towards the end of the network, with contrastive SSL and supervised models having a larger amount of biases.

Dissecting the bias detections at $p < 10^{-2}$ (Figure 5) for the embeddings of different ResNet blocks we gain insight into the distribution of the number and strength of the biases identified at different model depths. Overall, the cumulative strength of detected biases is smallest around the 3rd and 4th blocks, and it grows at the 5th block. One can observe that some models (e.g., rotation, jigsaw and supervised) deviate from this pattern. Finally, for each model, the embeddings of the 1st ResNet block yield some of the highest cumulative strength values.

4.4. Biases in the downstream tasks

Additionally, we perform a preliminary experiment that further exemplifies the practical consequences of a contrastive model being more biased than a clustering one. Specifically, we analyze two SSL models: SimCLR (contrastive) and ODC (clustering). According to both p- and d- values of the Gender-Career bias (see Table 7 in Supplementary Material) the classification of people's occupation (career) is more biased by gender in SimCLR than in ODC. To validate this result, we transfer the knowledge of the SSL models to a relationship classifier for image subjects, and assess the differences in the classification accuracy for male and female subjects. To this aim, we train two linear classifiers on top of SimCLR and ODC features, to predict people's relationship ("Friends", "Family", "Couple", "Professional"), using the training set of the People in Social Context dataset [34]. We evaluate the predictions of both models on the test set, where the people's gender has been manually annotated in 400 images (male-only: M, female-only: F, and male/s and female/s together: M+F). Table 2 shows the classification accuracy for each relationship category, for the whole test set (A) and for the M, F and M+F gender subsets. The results indicate that whereas SimCLR performs better overall (in line with the results in Table 1), it is less accurate than ODC in the classification of female professionals, while being better in the classification of male professionals. Hence, the Gender-Career bias detected in the SSL backbone is transferred to the downstream task and noticeably affects the results favoring "Professional" predictions towards male subjects.

Table 2. Classification accuracy of the Professional-Family-Friends-Couple categories according to the genders of people present in the image. M stands for "Male-only", F stands for "Female-only", A stands for "Any gender"-males, females or both together, M+F stands for males and females together.

	ODC, accuracy (%)				SimCLR, accuracy (%)			
	A	M	F	M+F	A	M	F	M+F
Professional	69.2	74.6	81.8	56.3	75.2	82.5	72.7	66.7
Family	51.4	46.9	50.0	56.3	55.6	56.3	37.5	59.4
Couple	44.0	59.1	28.6	33.3	54.0	68.2	42.9	42.9
Friends	40.9	38.6	45.0	41.3	50.9	50.0	45.0	54.4

5. Discussion on the experimental results

5.1. Relation between the number of biases and SSL learning strategy

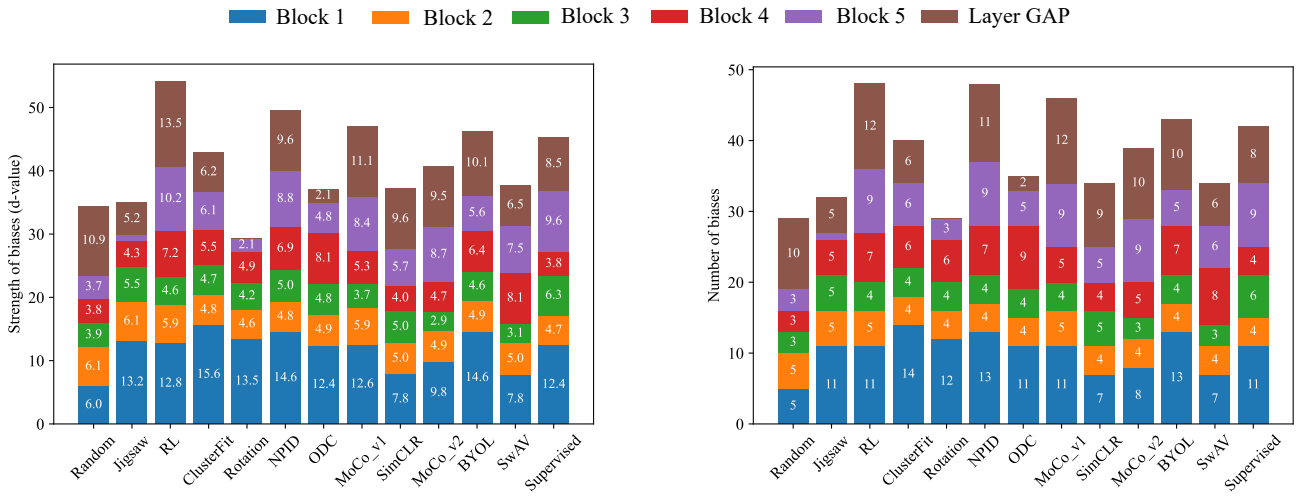
Figures 3b and 5 suggest that deep embeddings obtained with contrastive SSL models show more biases than the ones computed with geometric and clustering-based models. We hypothesize that the reason for this difference might lie in the nature of the contrastive loss function. A contrastive loss promotes the similarity between features of two images representing a concept and an attribute (not necessarily related) if the images are similar. Instead, a geometry-based loss function will not amplify this circumstantial similarity as much as a contrastive one.

Moreover, we state that a higher classification accuracy of the SSL model does not necessarily result in a higher number of social biases incorporated into it. Figure 5 provides a good example of this by showing that one of the least accurate, among studied, model (RL) yields the highest cumulative strength of biases detected in it. In addition to it, ODC demonstrates an inferior to NPID cumulative bias-strength, whilst being more accurate (based on the accuracy of linear classifiers trained on the features of the 5th ResNet block).

The aforementioned conclusions might have an important application in the deployment of deep learning models for tasks that have an impact on human processes: in the context of transfer learning, one must not solely rely on final accuracy when choosing a model or the layer-depth to extract embeddings from. Based on the ImageNet performances, we argue that it might be beneficial, for some models, to give preference to embeddings resulting in a slightly lower accuracy but significantly reducing the strength of identified biases. For instance, for two linear classifiers trained on the embeddings from the 5th block and the GAP layer of NPID, the difference in classification accuracy on ImageNet is only 0.01%. Meanwhile, cumulative strength of biases of these two layers differs by 9%. Furthermore, the classifier trained on the GAP embeddings of NPID is 3.18% more accurate than the classifier trained on the GAP embeddings of ODC, but the number of *intersectional* biases acquired by them differs significantly (see Figure 4).

5.2. Distribution of biases along the layers

Although the presence of biases in initial layers of the model might be counterintuitive, we believe that it can be



(a) Cumulative strength of biases with p-value < 0.01

(b) Number of biases with p-value < 0.01

Figure 5. Cumulative strength (on the left) and number (on the right) of biases detected in different layers. The number of biases correlates with the cumulative strength (see additional plots in supplementary material). Models are ordered according to their classification accuracy on ImageNet.

explained through a correlation between the low-level characteristics of the test data and the nature of the filters learned in the shallow layers of CNNs, i.e., similar data issues affect the first layer and the random models as described in Section 4.2. For example, many of the biases that are consistently detected in the 1st block embeddings of all models relate to the *skin tone* and *valence* or *weight* and *valence*. Considering that images representing the concept of *pleasantness* contain brighter pixels (like the images of *white skin tone*) and images representing the concept of *unpleasantness* contain darker pixels (like the images of *dark skin tone*), it could be expected that the correlation identified between corresponding embeddings might be caused by these data factors (as we explore in the supplementary material) and not by the meaning of the depicted concepts.

Regarding the distribution of the biases, besides the aforementioned behaviour in the first layer, biases grow in strength and quantity as one advances along the contrastive models, strongly correlating the number and intensity of the acquired biases with the classification potential of the embeddings at each layer of a given model—generally the deeper the better as reported in OpenSelfSupervised Framework: the more specialized the embeddings are within a model, the more and stronger biases are acquired. The same trend is observable in the RL model, and in a more subtle way in the ODC model. Biases in the supervised model are more evenly distributed along the layers and strongly increase in the last layers, maybe because these are closer to the label-guided classification layer. The rotation model shows a different behaviour with lower and less intense biases evenly distributed along the model without representative biases in the GAP layer.

6. Conclusion

In this work, building on the existing approaches, we study the presence of common social biases in three types of SSL models: geometric, clustering-based and contrastive. We show that the number of detected biases does not depend on the SSL model's classification accuracy but on its type, with contrastive models yielding the highest number of biases. Moreover, we show that the presence of biases is not constant across different layers of a model, and that this layer-distribution of biases changes across models. Given these findings, we suggest that the number and strength of biases should be taken into account, alongside the resulting accuracy, when performing transfer learning on (supervised or SSL) pre-trained models. Specially for tasks that have an impact on human processes, this educated selection would result in models with a better trade-off in terms of accuracy and bias. Nevertheless, not all open questions have been answered yet: the sources of biases that stem from training data need to be isolated, and the influence of the dataset used during the models training needs to be investigated more closely. In fact, this study considered an ample number of models, although all of them trained only on ImageNet. This limitation opens avenues for further exploration of biases that arise in models trained using different datasets.

Acknowledgements: This work was supported by the Consejería de Educación e Investigación of the Comunidad de Madrid under Project SI1/PJI/2019-00414.

References

[1] Kenya's High Court Delays National Biometric ID Program. 1

[2] SB 6280 - 2019-20. Concerning the use of facial recognition services. 1

[3] SB S5140B. Relates to the use of biometric identifying technology. 1

[4] Indian Supreme Court decision, K. S. Puttaswamy v. Union of India, Supreme Court of India, Writ Petition (Civil) No. 494, 2012. 1

[5] Edward Bridges v. Chief Constable of South Wales Police and Secretary of State For The Home Department, EWHC 2341 (Admin), Case No. CO/4085/2018, September 4, para. 59, 2019. 1

[6] Jamaican Supreme Court Decision, Julian Robinson v. Attorney General of Jamaica JMFC Full 04, 2019. 1

[7] Ujué Agudo and Helena Matute. The influence of algorithms on political and dating decisions. *Plos one*, 16(4):e0249454, 2021. 1

[8] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[9] Avrim Blum and Kevin Stangl. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? In *1st Symposium on Foundations of Responsible Computing*, volume 156, pages 3:1–3:20, 2020. 1

[10] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019. 1

[11] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 2, 4

[12] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, pages 132–149, 2018. 2, 3

[13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020. 4, 6

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 4

[15] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255, 2020. 2

[16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3, 4

[17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1, 2

[18] Kate Conger, Richard Fausset, and Serge F Kovaleski. San Francisco bans facial recognition technology. *The New York Times*, 14, 2019. 1

[19] Christine Basta Marta R Costa-juss and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *GeBNLP 2019*, page 33, 2019. 2

[20] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014. 1

[21] Jia Deng, Alex Berg, Sanjeev Satheesh, Hao Su, Aditya Khosla, and Fei-Fei Li. Large scale visual recognition challenge. *www. image-net. org/challenges/LSVRC/2012*, 1, 2012. 1, 4

[22] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1430, 2015. 2, 3, 4

[23] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998. 1, 2, 3, 4

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020. 1, 2, 3, 4

[25] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 3

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006. 3

[27] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, volume 29, page 3323–3331, 2016. 1

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 4

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 4

[30] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787, 2018. 1

[31] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Craig Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of ACL*, 2020. 1

[32] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2, 3, 4

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[34] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2659, 2017. 7

[35] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 622–628, 2019. 2

[36] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, 2016. 3, 4

[37] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 701–713, 2021. 2, 4, 5

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 2

[39] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 6–12, 2020. 2

[40] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018. 1

[41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3, 4

[42] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. 3, 4

[43] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020. 1

[44] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6688–6697, 2020. 2, 3, 4

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017. 1, 2