Fairness, Debiasing and Privacy in Computer Vision and Medical Imaging

Carlo Alberto Barbano[1,2], Edouard Duchesnay[3], Benoit Dufumier[3], Pietro Gori[2] and Marco Grangetto[1,*]

[1]Computer Science Dept., University of Turin, Italy
[2]LTCI, Télécom Paris, IP Paris, France
[3]NeuroSpin, CEA, Université Paris-Saclay, France

Abstract

Deep Learning (DL) has become one of the predominant tools for solving a variety of issue, often with superior performance compared to previous state-of-the-art methods. DL models are often able to learn meaningful and abstract representations of the underlying data; however, they have also been shown to often learn additional features in the data, which are not necessarily relevant or required for the desired task. This could pose a number of issues, as the additional features can contain bias, sensitive or private information, that should not be taken into account (e.g. gender, race, age, etc.) by the model. We refer to this information as *collateral*. The presence of collateral information translates into practical issues when deploying DL models, especially if they involve users' data. Learning robust representations which are free of biased, private, and collateral information can be very relevant for a variety of fields and applications, for example for medical applications and decision support systems. In this work we present our group's activities aiming at devising methods to ensure that representations learned by DL models are robust to collateral features, biases and privacy-preserving with respect to sensitive information.

Keywords

Fairness, Debiasing, Privacy, Deep Learning, Representation Learning

1. Introduction

In this section we describe the latest research activities in the field of fairness and privacy for deep learning at the EIDOSLAB[1] research group [1], the computer vision and image processing laboratory in the Computer Science department of the University of Turin. The lab is also a member of the Italian Association for Computer Vision, Pattern Recognition and Machine Learning [2].

Trustworthiness, fairness and ethics have become increasingly important topics in deep learning. As deep learning models become more prevalent in computer vision tasks, including medical imaging, concerns about fairness and bias have come to the forefront. Biases can creep into these models in a variety of ways, including the selection and preparation of training data, the design of the model architecture, and the optimization of the model parameters. These biases can have real-world consequences, especially in medical imaging where decisions made based on the results of these models can impact patients' lives. As a result, researchers have been actively working on methods for debiasing these models and improving their overall fairness. This includes techniques such as data augmentation, regularization, and adversarial training, among others. In this context, debiasing methods aim to reduce disparities and ensure that these models perform equally well across different demographic groups. It is crucial to address fairness and bias in deep learning models in computer vision, especially in medical imaging, to ensure equitable and effective care for all patients. More specifically, one of the most important aspect of fairness and debiasing in deep learning models, particularly in medical imaging, is the potential for the model to learn additional information that can introduce bias or compromise the privacy and security of sensitive information. For example, a model may unintentionally learn information about a patient's race age or gender, which could then be inadvertently used to make decisions that unfairly advantage or disadvantage certain groups. Another problem which affects medical imaging is the noise related to the acquisition site, which in multi-site datasets can prevent the model from correctly generalizing to new data from different acquisition sites. Additionally, medical imaging often contains sensitive and personal information about patients (e.g. gender, age, race) which must be handled with care to ensure patient privacy and prevent potential data breaches. Thus, it is essential to not only address bias and fairness in these models but also to consider the potential risks

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.
carlo.barbano@unito.it (C. A. Barbano); edouard.duchesnay@cea.fr (E. Duchesnay); benoit.dufumier@cea.fr (B. Dufumier); pietro.gori@telecom-paris.fr (P. Gori); marco.grangetto@unito.it (M. Grangetto)
0000-0001-9512-0440 (C. A. Barbano); 0000-0002-4073-3490 (E. Duchesnay); 0000-0002-8253-2363 (B. Dufumier); 0000-0003-3456-8529 (P. Gori); 0000-0002-2709-7864 (M. Grangetto)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://eidos.di.unito.it

associated with the information they learn and how it is handled. Referring to all the above cases, we define as *collateral* any information that is not necessarily required for the desired task, but that is picked-up by the model. This concept which was conceptualized by John Dewey as *Collateral Learning*, describes the accidental learning that occurs in and outside the classroom [3]. Based on this definition, and extending it to the deep learning context, we say that collateral learning occurs when a model learns more information than intended. In order to be *robust*, DL models should not be affected by the collateral learning problem.

1.1. Representation Learning

A more throughout understanding of how deep models can learn powerful representations can certainly be helpful in all the above cases. Learning fair and robust representations of the underlying samples, especially when dealing with biased data or sensitive information, is the main objective of the activities described in this work. In the recent years, the topic of representation learning has increasingly gained traction in the deep learning community. Contrastive learning has become the most widespread approach for this purpose, and many losses and frameworks have been proposed [4, 5, 6, 7]. Contrastive learning approaches aim at pulling positive samples representations (e.g. of the same class) closer together while repelling representations of negative ones (e.g. different classes) apart from each other. It has also been shown that, in a supervised setting, this kind of optimization can sometimes yield better results than standard cross-entropy [5], and is also more robust against label corruption [8] which can be seen as an instance of collateral features. However, a lot remains to be done about this matter, and research should focus on how to provide reliable guarantees for avoiding collateral features learning. Furthermore, another relevant line of research is addressing this issue from an unsupervised perspective (i.e. automatically recognizing and excluding all bias and collateral information without any prior knowledge).

In summary, there is a need for a reliable way to learn robust representations which are free of biased, private and collateral information.

2. Metric framework for contrastive learning

In our research activities, we explore representation learning from a theoretical perspective. We propose a metric-learning based framework for supervised representation learning, which allows us to derive and formalize a more robust set of debiasing constraints, along with novel contrastive losses that show increased robustness compared to the current literature [9]. We provide a unified framework to analyze and compare existing formulations of contrastive losses, such as the InfoNCE loss [4, 6], the InfoL1O loss [7], and the SupCon loss [5]. Using our proposed metric learning approach, we can reformulate each loss as a set of highly explainable metric conditions. Our analysis provides a comprehensive understanding of the different loss functions, explaining their behavior from a metric point of view. Furthermore, leveraging our metric learning approach, we investigate the issue of biased learning. We point out the limitations of the studied contrastive loss functions when dealing with biased data, especially when the loss on the training set is apparently minimized. By analyzing such cases, we provide a more formal characterization of bias, which eventually allows us to derive a new set of general regularization constraints for debiasing that can be added to any contrastive or non-contrastive loss.

Foundamentals Let $x \in X$ be an original sample (i.e., anchor), x_i^+ a similar (positive) sample, x_j^- a dissimilar (negative) sample and P and N the number of positive and negative samples respectively. Contrastive learning methods look for a parametric mapping function $f : X \rightarrow \mathbb{S}^{d-1}$ that maps "semantically" similar samples close together in the representation space, a (d-1)-sphere, and dissimilar samples far away from each other. Once pre-trained, f is fixed and its representation is evaluated on a downstream task, such as classification, through linear evaluation on a test set. In general, positive samples x_i^+ can be defined in different ways depending on the problem: using transformations of x (unsupervised setting), samples belonging to the same class as x (supervised) or with similar image attributes of x (weakly-supervised). The definition of negative samples x_j^- varies accordingly. Here, we focus on the supervised case, thus samples belonging to the same/different class, but the proposed framework could be easily applied to the other cases. We define $s(f(a), f(b))$ as a similarity measure (e.g., cosine similarity) between the representation of two samples a and b. Please note that since $||f(a)||_2 = ||f(b)||_2 = 1$, using a cosine similarity is equivalent to using a L2-distance ($d(f(a), f(b)) = ||f(a) - f(b)||_2^2$). Similarly to [10, 11, 12, 13, 14], we propose to use a metric learning approach which allows us to better formalize recent contrastive losses, such as InfoNCE [4, 6], InfoL1O [7] and SupCon [5], and derive new losses that better approximate the mutual information and can take into account data biases.

Derivation of ϵ-SupInfoNCE Using an ϵ-margin metric learning point of view, probably the simplest con-

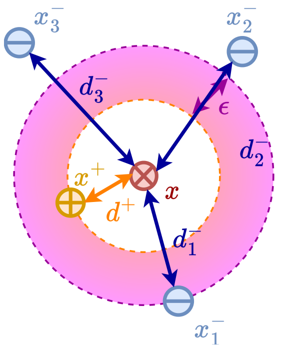

Figure 1: With ϵ-SupInfoNCE (a) we aim at increasing the minimal margin ϵ, between the distance d^+ of a positive sample x^+ (+ symbol inside) from an anchor x and the distance d^- of the closest negative sample x^- (− symbol inside). By increasing the margin, we can achieve a better separation between positive and negative samples.

trastive learning formulation is looking for a mapping function f such that the following ϵ-condition is always satisfied:

$$\underbrace{s(f(x), f(x_j^-))}_{s_j^-} - \underbrace{s(f(x), f(x_i^+)}_{s_i^+} \leq -\epsilon \quad \forall i, j \tag{1}$$

where $\epsilon \geq 0$ represents a margin between positive and negative samples, as shown in Fig. 1. The constraint of Eq. 1 can be transformed into an optimization problem using, as it is common in contrastive learning, the max operator and its smooth approximation *LogSumExp*. The can lead to the derivation of different loss functions. Some of them can be found in [9]. We propose to use the following one, that we call ϵ-SupInfoNCE:

$$\sum_i \max(-\epsilon, \{s_j^- - s_i^+\}_{j=1,\dots,N}) \approx$$

$$\approx \sum_i \log\left(\exp(-\epsilon) + \sum_j \exp(s_j^- - s_i^+)\right) \tag{2}$$

$$= \underbrace{-\sum_i \log\left(\frac{\exp(s_i^+)}{\exp(s_i^+ - \epsilon) + \sum_j \exp(s_j^-)}\right)}_{\epsilon-SupInfoNCE}$$

Here, we can notice that when $\epsilon = 0$, we retrieve a generalization of InfoNCE loss, whereas when $\epsilon \to \infty$ we obtain a generalization of InfoL1O loss. It has been shown in [7] that these two losses are the lower and upper bound of the Mutual Information $I(X^+, X)$ respectively:

$$\text{InfoNCE} \leq I(X^+, X) \leq \text{InfoL1O} \tag{3}$$

By using a value of $\epsilon \in [0, \infty)$, one might find a tighter approximation of $I(X^+, X)$ since the exponential function at the denominator $\exp(-\epsilon)$ monotonically decreases as ϵ increases.

Experiments and Results Results on general computer vision datasets are presented in Tab. 1, in terms of top-1 accuracy. We report the performance for the best value of ϵ; the complete results can be found in [9]. The results are averaged across 3 trials for every configuration, and we also report the standard deviation. We obtain significant improvement with respect to all baselines and, most importantly, SupCon, on all benchmarks: on CIFAR-10 (+0.5%), on CIFAR-100 (+0.63%), and on ImageNet-100 (+1.31%). For the experiments, we use the original setup from SupCon [5], employing a ResNet-50. The complete experimental setup is provided in [9].

3. Debiasing with FairKL

Satisfying the ϵ-condition (1) can generally guarantee good downstream performance, however, it does not take into account the presence of biases (e.g. selection biases). To tackle this issue, we propose FairKL, a set of debiasing constraints that prevent the use of the bias features within the proposed metric learning approach. In order to give a more in-depth explanation of the ϵ-InfoNCE failure case, we employ the notion of *bias-aligned* and *bias-conflicting* samples as in Nam et al. [15]. In our context, a bias-aligned sample shares the same bias attribute of the anchor, while a bias-conflicting sample does not. In this work, we assume that the bias attributes are either known *a priori* or that they can be estimated using a bias-capturing model, such as in [16].

Characterization of bias We denote bias-aligned samples with $x^{\cdot,b}$ and bias-conflicting samples with $x^{\cdot,b'}$. Given an anchor x, if the bias is "strong" and easy-to-learn, a *positive bias-aligned* sample $x^{+,b}$ will probably be closer to the anchor x in the representation space than a *positive bias-conflicting* sample (of course, the same reasoning can be applied for the negative samples). This is why even in the case in which the ϵ-condition is satisfied and the ϵ-SupInfoNCE is minimized, we could still be able to distinguish between bias-aligned and bias-conflicting samples. Hence, we say that there is a bias if we can identify an ordering on the learned representations, e.g.:

$$s_j^- - \epsilon \leq s_k^{+,b'} < s_i^{+,b} \quad \forall i, k, j \tag{4}$$

This represents the worst-case scenario, where the ordering is total (i.e., $\forall i, k, t, j$). Of course, there can also be cases in which the bias is not as strong, and the ordering may be partial.

FairKL regularization for debiasing Ideally, we would enforce the conditions $s_k^{+,b'} - s_i^{+,b} = 0 \quad \forall i, k$ and, meaning that every positive bias-conflicting sample should have the same distance from the anchor as any

Table 1
Accuracy on vision datasets. SimCLR and Max-Margin results from [5]. Results denoted with * are (re)implemented with mixed precision due to memory constraints.

Dataset	Network	SimCLR	Max-Margin	SimCLR*	CE*	SupCon*	ϵ-SupInfoNCE*
CIFAR-10	ResNet-50	93.6	92.4	91.74±0.05	94.73±0.18	95.64±0.02	**96.14**±0.01
CIFAR-100	ResNet-50	70.7	70.5	68.94±0.12	73.43±0.08	75.41±0.19	**76.04**±0.01
ImageNet-100	ResNet-50	-	-	66.14±0.08	82.1±0.59	81.99±0.08	**83.3**±0.06

other positive bias-aligned sample. However, in practice, this condition is very strict, as it would enforce uniform distance among all positive samples. A more relaxed condition would instead force the distributions of distances, $\{s_k^{\cdot,b'}\}$ and $\{s_i^{\cdot,b}\}$, to be similar. Here, we propose two new debiasing constraints using either the first moment (mean) of the distributions or the first two moments (mean and variance). Using only the average of the distributions, we obtain:

$$\frac{1}{P_c}\sum_k |s_k^{+,b'}| - \frac{1}{P_a}\sum_i |s_i^{+,b}| = 0 \tag{5}$$

where P_a and P_c are the number of positive bias-aligned and bias-conflicting samples, respectively[2]. Coincidentally, this constraint is also known as EnD [17], which we proposed in 2021. Denoting the first moments with $\mu_{+,b} = \frac{1}{P_a}\sum_i s_i^{+,b}$, $\mu_{+,b'} = \frac{1}{P_c}\sum_k s_k^{+,b'}$, and the second moments of the distance distributions with $\sigma^2_{+,b} = \frac{1}{P_a}\sum_i (s_i^{+,b} - \mu_{+,b})^2$, $\sigma^2_{+,b'} = \frac{1}{P_c}\sum_k (s_k^{+,b'} - \mu_{+,b'})^2$, and making the hypothesis that the distance distributions follow a normal distribution, we can define a new debiasing constraint $\mathcal{R}^{FairKL}$ using, for example, the Kullback–Leibler divergence:

$$\frac{1}{2}\left[\frac{\sigma^2_{+,b} + (\mu_{+,b} - \mu_{+,b'})^2}{\sigma^2_{+,b'}} - \log\frac{\sigma^2_{+,b}}{\sigma^2_{+,b'}} - 1\right] = 0 \tag{6}$$

The proposed debiasing constraint can be easily added to any contrastive loss using the method of the Lagrange multipliers, as a regularization term. Thus, our final loss function is:

$$\mathcal{L} = \alpha\mathcal{L}^{\epsilon-SupInfoNCE} + \lambda\mathcal{R}^{FairKL} \tag{7}$$

where α and λ are positive hyperparameters.

Experiments and results We perform experiments on our proposed loss on five biased datasets: Biased-MNIST, Corrupted-CIFAR10, bFFHQ, and 9-Class ImageNet along with ImageNet-A. For brevity, in this presentation we report Biased-MNIST only, the results are reported in Tab. 2. The complete results and experimental details are provided in [9]. On this dataset, where colors are injected into the background of the MNIST digits with a varying degree of correlation, we achieve state-of-the-art results.

Table 2
Top-1 accuracy (%) on Biased-MNIST. Reference results from [16]. Results denoted with * are re-implemented without color-jittering and bias-conflicting oversampling.

	Correlation (%)			
Method	99.9	99.7	99.5	99
CE [16]	11.8±0.7	62.5±2.9	79.5±0.1	90.8±0.3
LNL [18]	18.2±1.2	57.2±2.2	72.5±0.9	86.0±0.2
EnD [17]	59.5±2.3	82.70±0.3	94.0±0.6	94.8±0.3
BC+BB* [16]	30.26±11.08	82.83±4.17	88.20±2.27	95.04±0.86
BB [16]	76.8±1.6	91.2±0.2	93.9±0.1	96.3±0.2
BC+CE* [16]	15.06±2.22	90.48±5.26	95.95±0.11	97.67±0.09
FairKL	**90.51**±1.55	**96.19**±0.23	**97.00**±0.06	**97.86**±0.02

4. Multi-site acquisition noise in brain age prediction

In this section, we present our recent work in the field of neuroimaging, focusing on brain age prediction from MRI. This is a challenging task that requires robust and accurate models capable of generalizing across different imaging sites. Dealing with multi-site dataset is a delicate matter in biomedical imaging in general, as the collateral noise related to the different acquisition sites often limits the generalization capability of DL models. In this context, together with our partners at Télécom Paris (IP Paris) and NeuroSpin (CEA), we have developed a novel contrastive learning loss for regression of brain age from MRI [19], which is based on our metric learning framework. We validated it on the OpenBHB challenge [20], a recently released[3] public challenge, which provides one of the largest datasets of healthy brain MRIs. Based on the framework presented in Sec. 2, we propose a novel contrastive learning regression loss for brain age prediction, achieving state-of-the-art performance on the OpenBHB challenge.

[2] The same reasoning can be applied to negative samples (omitted for brevity.)

[3] https://baobablab.github.io/bhb/

Contrastive Learning Regression Loss The notion of negative and positive samples is rooted in the contrastive learning framework. The loss formulation of Sec. 2 is thus not adapted for regression (i.e. continuous labels), as it is not possible to determine a hard boundary between positive and negative samples. All samples are somehow positive and negative at the same time. Given the continuous label y_i for the anchor and y_k for a sample k, one could threshold the difference Δ between y_i and y_k at a certain value τ in order to create positive and negative samples (i.e. k is positive if $\Delta(y_i, y_k) < \tau$). The problem would then be how to choose τ. Differently, we propose to define a degree of "positiveness" between samples using a kernel function $w_k = K(y_i - y_k)$, where $0 \leq w_k \leq 1$. Our goal is thus to learn a parametric function $f : \mathcal{X} \rightarrow \mathbb{S}^{d-1}$ that maps samples with a high degree of positiveness ($w_k \sim 1$) close in the latent space and samples with a low degree ($w_k \sim 0$) far away from each other. To adapt such a framework to continuous labels, we propose to use a kernel function w_k, and we develop multiple formulations. A first approach would be to consider as "positive" only the samples that have a degree of positiveness greater than 0, and align them with a strength proportional to the degree:

$$\frac{w_k}{\sum_j w_j}(s_t - s_k) \leq 0 \quad \forall j, k, t \neq k \in A(i) \tag{8}$$

where we have normalized the kernel so that the sum over all samples is equal to 1 and we denote with $A(i)$ the indices of samples in the minibatch distinct from x_i. From Eq. 8 we can derive the following loss:

$$\mathcal{L}^{y-aware} = -\sum_k \frac{w_k}{\sum_t w_t} \log \left(\frac{\exp(s_k)}{\sum_{t=1}^N \exp(s_t)} \right) \tag{9}$$

Interestingly, this is exactly the *y-aware* loss proposed in [21] for classification with weak continuous attributes. Due to the non-hard boundary between positive and negative samples, both s_t and s_k are defined over the entire minibatch. The kernel w_k is used to avoid aligning samples not similar to the anchor (i.e. $w_k \approx 0$). It can be noted that, while the numerator aligns x_k, in the denominator, the uniformity term (as defined in [22]) focuses more on the closest samples in the representation space: this could be undesirable, as these samples might have a greater degree of positiveness than the considered x_k. To avoid that, we formulate a first extension ($\mathcal{L}^{thr}$) of (8), which limits the uniformity term (i.e., denominator) to the samples that are at least more distant from the anchor than the considered x_k in the kernel space (omitting the normalization in the starting condition):

$$w_k(s_t - s_k) \leq 0 \quad \text{if } w_t - w_k \leq 0 \quad \forall k, t \neq k \in A(i)$$
$$\mathcal{L}^{thr} = -\sum_k \frac{w_k}{\sum_t \delta_{w_t < w_k} w_t} \log \left(\frac{\exp(s_k)}{\sum_{t \neq k} \delta_{w_t < w_k} \exp(s_t)} \right) \tag{10}$$

Table 3
Final scores on the OpenBHB leaderboard.

Method	Model	Int. MAE	BAcc	Ext. MAE	$\mathcal{L}_c$
Baseline ($L1$)	ResNet-18	2.67±0.05	6.7±0.1	4.18±0.01	1.86
ComBat	ResNet-18	4.15±0.01	**4.5**±0.0	4.76±0.03	1.88
$\mathcal{L}^{exp}$	ResNet-18	**2.55**±0.00	5.1±0.1	**3.76**±0.01	**1.54**

Ideally, $\mathcal{L}^{thr}$ avoids repelling samples more similar than x_k. However, it still focuses more on the closest sample "less positive" than x_k, i.e. x_t s.t $w_t > w_x$ and $w_t \leq w_j \; \forall j \neq k$. As noted in [9, 5], increasing the margin with respect to the closest "negative" sample works well for classification; however we argue it might not be best suited for regression. For this reason, we propose a second formulation ($\mathcal{L}^{exp}$) that takes an opposite approach. Instead of focusing on repelling the closest "less positive" sample, we increase the repulsion strength for samples proportionally to their distance from the anchor in the kernel space:

$$w_k[s_t(1 - w_t) - s_k] \leq 0 \quad \forall k, t \neq k \in A(i)$$
$$\mathcal{L}^{exp} = -\frac{1}{\sum_j w_j} \sum_{k \in A(i)} w_k \log \frac{\exp(s_k)}{\sum_{t \neq k} \exp(s_t(1 - w_t))} \tag{11}$$

In the resulting $\mathcal{L}^{exp}$ formulation, the weighting factor $1 - w_t$ acts like a temperature value, by giving more weight to the samples which are farther away from the anchor in the kernel space. Also, for a proper kernel choice, samples closer than x_k will be repelled with very low strength (∼0). We argue that this approach is more suited for continuous attributes (i.e., regression task), as it enforces that samples close in the kernel space will be close in the representation space.

Results With our proposed loss, we achieve the best results (at this time) [9] on the OpenBHB leaderboard, as shown in Tab. 3 ($\mathcal{L}_c$). Compared to the L1 and ComBat baselines [19], we achieve a lower generalization error to unseen sites (Ext. MAE), meaning that our method is more robust to the collateral information related to the site noise. We are currently carrying out further research to gain further insights on the reasons of this behavior.

5. Privacy in deep learning

We investigated the possibility of utilizing debiasing technique also to prevent privacy leakage. In this context, we are interested in recovering some private attribute of the data, starting from the model outputs or embeddings. These kind of private attributes can be, in the example of natural or facial images, age, gender, race, etc. We observed that, under certain conditions, some of the debiasing approaches are also suitable for privacy preservation. We discovered the determining condition to be the capability of effectively suppressing the bias related information inside of the model, rather than simply re-weighting it. We show in [23] that debiasing techniques can be used for privacy preservation purposes when they allow to retain a high accuracy on the target class, while making it harder to determine the private attributes. In our work, we successfully remove collateral private information, e.g. gender or age, from the latent representation of the DL models on a variety of datasets, including medical images; thus ensuring that they cannot leak from the model outputs.

References

[1] EidosLab, Image proecessing, computer vision and virtual reality, https://eidos.di.unito.it, 2021.

[2] CVPL, Italian Association for Computer Vision, Pattern Recognition and Machine Learning, http://www.cvpl.it, 2021.

[3] J. Dewey, Experience And Education, Free Press, 1997. URL: https://books.google.fr/books?id=UWbuAAAAMAAJ.

[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607. URL: http://proceedings.mlr.press/v119/chen20j.html, iSSN: 2640-3498.

[5] P. Khosla, et al., Supervised contrastive learning, in: NeurIPS, 2020.

[6] A. v. d. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv:1807.03748 [cs, stat] (2019). URL: http://arxiv.org/abs/1807.03748, arXiv: 1807.03748.

[7] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, G. Tucker, On Variational Bounds of Mutual Information, in: ICML, 2019.

[8] F. Graf, et al., Dissecting supervised contrastive learning, in: ICML, 2021. URL: https://proceedings.mlr.press/v139/graf21a.html.

[9] C. A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, P. Gori, Unbiased supervised contrastive learning, in: The Eleventh International Conference on Learning Representations (ICLR), 2023. URL: https://openreview.net/forum?id=Ph5cJSfD2XN.

[10] S. Chopra, R. Hadsell, Y. LeCun, Learning a Similarity Metric Discriminatively, with Application to Face Verification, in: CVPR, volume 1, IEEE, 2005, pp. 539–546. URL: http://ieeexplore.ieee.org/document/1467314/. doi:10.1109/CVPR.2005.202.

[11] K. Sohn, Improved Deep Metric Learning with Multi-class N-pair Loss Objective, in: Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://papers.nips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

[12] J. Wang, Y. song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning Fine-grained Image Similarity with Deep Ranking, in: CVPR, 2014.

[13] X. Wang, Y. Hua, E. Kodirov, N. M. Robertson, Ranked List Loss for Deep Metric Learning, in: CVPR, 2019.

[14] B. Yu, D. Tao, Deep Metric Learning With Tuplet Margin Loss, in: IEEE ICCV, 2019, pp. 6489–6498.

[15] J. Nam, H. Cha, S. Ahn, J. Lee, J. Shin, Learning from failure: Training debiased classifier from biased classifier, in: Advances in Neural Information Processing Systems, 2020.

[16] Y. Hong, E. Yang, Unbiased classification through bias-contrastive and bias-balanced learning, in: Thirty-Fifth Conference on Neural Information Processing Systems, 2021. URL: https://openreview.net/forum?id=2OqZZAqxnn.

[17] E. Tartaglione, C. A. Barbano, M. Grangetto, End: Entangling and disentangling deep representations for bias correction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 2021.

[18] B. Kim, H. Kim, K. Kim, S. Kim, J. Kim, Learning not to learn: Training deep neural networks with biased data, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] C. A. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, P. Gori, Contrastive learning for regression in multi-site brain age prediction, in: International Symposium on Biomedical Imaging (ISBI), 2023.

[20] B. Dufumier, et al., Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing, NeuroImage (2022).

[21] B. Dufumier, et al., Contrastive learning with continuous proxy meta-data for 3d mri classification, in: MICCAI, Springer, 2021.

[22] T. Wang, P. Isola, Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, ICML (2020). URL: http://arxiv.org/abs/2005.10242.

[23] C. A. Barbano, E. Tartaglione, M. Grangetto, Bridging the gap between debiasing and privacy for deep learning, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, 2021, pp. 3799–3808.