Fairness, Debiasing and Privacy in Computer Vision and Medical Imaging

Ital-IA 2023 - Pisa

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

[1]Computer Science Dept., University of Turin, Italy,

[2]LTCI, Telécom Paris, IP Paris

[3]NeuroSpin, CEA, Université Paris-Saclay, France

Table of Contents

EIDOSLAB was founded in 1985 and is the **computer vision and image processing group** of the Computer Science department of the **University of Turin**

Home: `https://eidos.di.unito.it/`

Introduction

1 Introduction

- Trustworthiness, fairness and ethics have become increasingly important topics in deep learning. It has become increasingly evident how deep learning models are often affected by biases.
- 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.
- By studing how representations are learned by Deep Neural Networks (DNNs), one might be able to avoid this issue

Our Contributions

- Recently, for the task of representation learning, Contrastive Learning (CL) has become the predominant approach

- We study the problem of learning *fair* and *robust* representations in deep neural networks (DNNs)

- For this purpose, we try to formalize how representations are learned with a theoretical metric learning approach

Table of Contents

2 A Metric Approach for Contrastive Learning

Contrastive Learning - Notation

Aim of contrastive learning methods: look for a parametric mapping function $f_\theta : X \to S^{d-1}$ that:

1. Maps similar samples close together in the representation space
2. Dissimilar samples further away

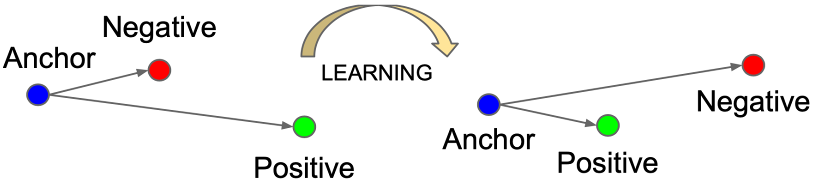


Figure: From Schroff *et al.* [7]

Contrastive Learning - Notation

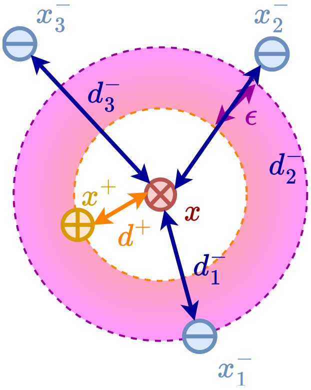
- Let $x \in X$ be a sample (*anchor*)
- x_i^+ is a positive sample (i.e. same class), and x_j^- is a negative one (i..e different class)
- s denotes the [cosine] similarity, with s_i^+ and s_j^- shorthand for $s(f(x), f(x_i^+))$ and $s(f(x), f(x_j^-))$

ϵ-margin

Using an ϵ-margin metric learning point of view, probably the simplest formulation is looking for a mapping function f that satisfies the following condition:

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

Here, $\epsilon \geq 0$ is the minimal margin between a positive sample and a negative sample (purple area)



Derivation of ϵ-SupInfoNCE

2 A Metric Approach for Contrastive Learning

- The condition $s_j^- - s_i^+ \leq -\epsilon \quad \forall i, j$ is equivalent to $\max\{s_j^- - s_i^+\} \leq -\epsilon$;
- In other words, we want to **maximize the minimal margin** between a positive and a negative sample;
- However, max is not differentiable. We employ *LogSumExp* (LSE) as a smooth approximation of the max operator, obtaining the ϵ-SupInfoNCE loss:

$$\arg\min_f \sum_i \max(-\epsilon, \{s_j^- - s_i^+\}) \approx -\sum_i \log\left(\frac{\exp(s_i^+)}{\exp(s_i^+ - \epsilon)\sum_j \exp(s_j^-)}\right)$$

Alternative derivations

Note that other derivations are possibile: some of them are shown in the full paper [2]. We can also retrieve the SupCon loss [5] or the InfoNCE loss used in SimCLR [3].

Table: 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

Table of Contents

3 Debiasing with FairKL

The Issue of Biases

- Satisfying the ϵ-condition can generally guarantee good **downstream performance**. However, it does not take into account the presence of **biases** (e.g. selection biases).

- We employ the notion of *bias-aligned* and *bias-conflicting* samples as in [6]:
  1. *bias-aligned*: shares the same bias attribute of the anchor. We denote it as $x^{+,b}$
  2. *bias-conflicting*: has a different bias attribute. We denote it as $x^{+,b'}$

Biases and Failure of ϵ-SupInfoNCE

3 Debiasing with FairKL

- 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;

- Thus, we say that there is a bias if we can identify an **ordering** on the learned representations, e.g.:

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

Note

This represents the worst-case scenario, where the ordering is total (i.e., $\forall i, k, j$). Of course, there can also be cases in which the bias is not as strong, and the ordering may be partial. Furthermore, the same reasoning can be applied to negative samples (omitted for brevity).

FairKL

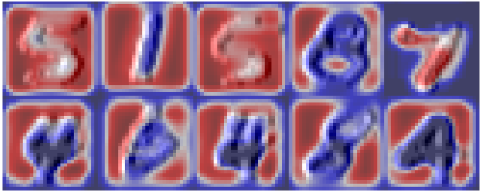
- Assuming that the similarities follow a normal distribution, we denote as $B_{+,b} \sim \mathcal{N}(\mu_{+,b}, \sigma^2_{+,b})$ and $B_{+,b'} \sim \mathcal{N}(\mu_{+,b'}, \sigma^2_{+,b'})$ the **distributions of similarities** of the bias-aligned and bias-conflicting samples respectively;

- We minimize the Kullback-Leibler divergence of the two distributions with the FairKL regularization term:

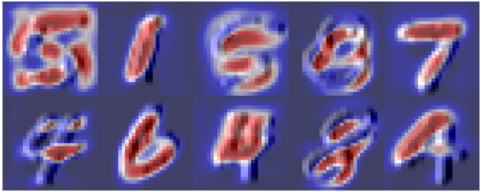
$$R^{FairKL} = D_{KL}(B_{+,b}||B_{+,b'}) = \frac{1}{2}\left[\frac{\sigma^2_{+,b} + (\mu_{+,b} - \mu_{+,b'})^2}{\sigma^2_{+,b'}} - \log\frac{\sigma^2_{+,b}}{\sigma^2_{+,b'}} - 1\right]$$

Results

3 Debiasing with FairKL



(a)



(b)

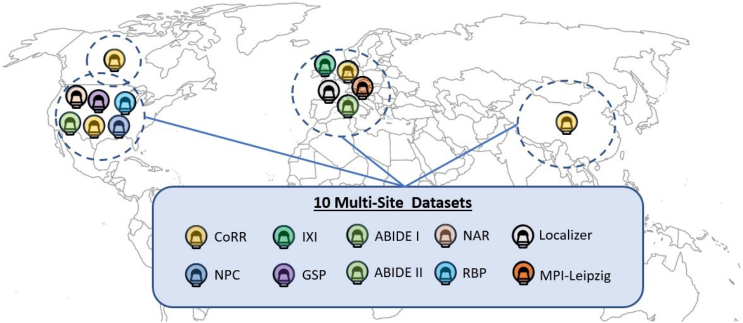
Figure: Grad-CAM [8] on Biased-MNIST: vanilla model (a) and regularized model (b).

Table of Contents

The OpenBHB Challenge

4 Multi-site Brain Age Prediction

- Multi-site medical datasets often present issues for DNNs generalization
- For this purpose, the OpenBHB challenge was created [4]. It gathers brain MRIs from 64 different acquisition sites.



Contrastive Learning for Regression

- The prediction target is age (regression). We no longer have a hard boundary between positive and negative samples
- We propose a novel contrastive loss for regression ($\mathcal{L}^{\text{exp}}$), using a *kernel* to define a *degree of positiveness* between two samples, based on the age difference:

$$\mathcal{L}^{exp} = -\frac{1}{\sum_j w_j} \sum_{k \in A(i)} w_k \log \frac{\exp(s_k)}{\sum_{t \neq k} \exp(s_t(1 - w_t))}$$

- We achieve the best results on the challenge in terms of error and robustness to site noise. More details in [1].
- We are currently working on extending FairKL also to regression tasks

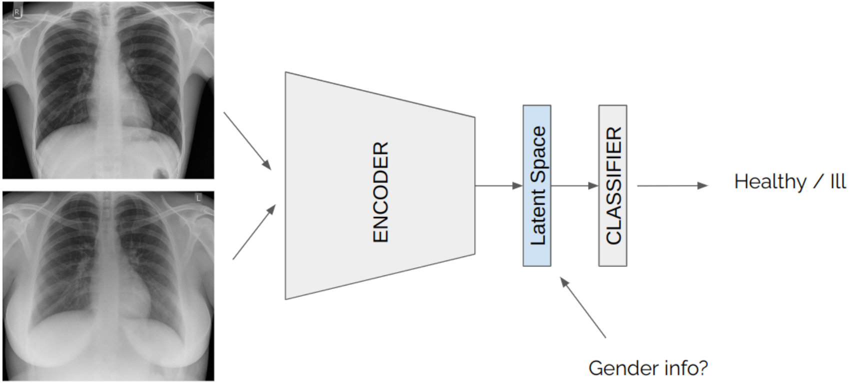
Table of Contents

5 Connection to Privacy

What about privacy?

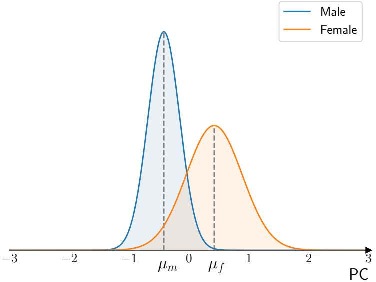
5 Connection to Privacy

Not only biases are a problem..

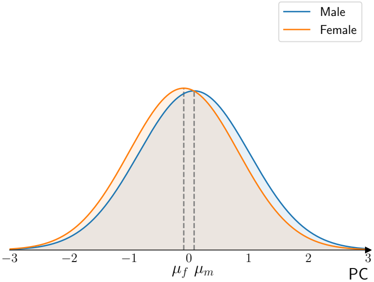


Debiasing to preserve privacy

5 Connection to Privacy



(a)



(b)

Figure: **Gaussian fit on the principal component** (PC) of the IMDB embeddings using a vanilla model (a) and a *EnD*-regularized model (b).

Table of Contents

6 Conclusions

Relevant publications

1. C.A. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, and P. Gori. Contrastive learning for regression in multi-site brain age prediction. *ISBI, 2023.*
2. C.A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, and P. Gori. Unbiased supervised contrastive learning. *ICLR, 2023.*
3. C. A. Barbano, E. Tartaglione, and M. Grangetto. Unsupervised learning of unbiased visual representations. *Preprint, 2022.*
4. C.A. Barbano, E. Tartaglione, and M. Grangetto. Bridging the gap between debiasing and privacy for deep learning. In *ICCVW, 2022.*
5. E. Tartaglione, C.A. Barbano, and M. Grangetto. End: Entangling and disentangling deep representations for bias correction. *CVPR, 2021.*

Thanks

Thank you for listening! Questions?

Table of Contents

7 References

References

[1] Carlo Alberto Barbano et al. "Contrastive learning for regression in multi-site brain age prediction". In: *International Symposium on Biomedical Imaging (ISBI)*. 2023.

[2] Carlo Alberto Barbano et al. "Unbiased Supervised Contrastive Learning". In: *The Eleventh International Conference on Learning Representations (ICLR)*. 2023. URL: https://openreview.net/forum?id=Ph5cJSfD2XN.

References

[3] Ting Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. en. In: *International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, Nov. 2020, pp. 1597–1607. URL: `http://proceedings.mlr.press/v119/chen20j.html` (visited on 02/24/2021).

[4] Benoit Dufumier et al. “OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing”. In: *NeuroImage* (2022).

References

[5] Prannay Khosla et al. “Supervised Contrastive Learning”. In: ed. by H Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[6] Junhyun Nam et al. “Learning from Failure: Training Debiased Classifier from Biased Classifier”. In: *Advances in Neural Information Processing Systems*. 2020.

References

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015). arXiv: 1503.03832, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. URL: http://arxiv.org/abs/1503.03832 (visited on 09/23/2021).

[8] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.