Unbiased Supervised Contrastive Learning

ICLR 2023 - Kigali, Rwanda

Carlo Alberto Barbano[1,2], Benoit Dufumier[2], Enzo Tartaglione[2], Marco Grangetto[1], Pietro Gori[2]

[1]University of Turin, [2]LTCI, Telécom Paris, IP Paris

Table of Contents

Our contributions

- **Aim of this work:** learn representations that are invariant to biases in the data
- We study deep representation learning with a metric approach, proposing a novel contrastive loss named ϵ-**SupInfoNCE**
- We formalize how biases can affect the representations, and we propose **FairKL**, a regularization technique for learning bias-invariant representations.

Table of Contents

Contrastive Learning - Notation

- Let $x \in X$ be a sample (*anchor*)
- Let x_i^+ be a positive sample (i.e. same class)
- Let x_j^- be a negative sample (i..e different class)

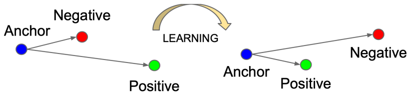


Figure: From Schroff *et al.* [5]

Contrastive Learning - Notation

- Let $x \in X$ be a sample (*anchor*)
- Let x_i^+ be a positive sample (i.e. same class)
- Let x_j^- be a negative sample (i..e different class)

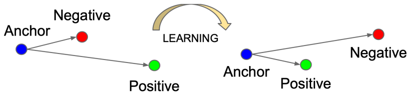


Figure: From Schroff *et al.* [5]

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

Contrastive Learning - Notation

- $d : S^{d-1} \times S^{d-1} \rightarrow R$ is a distance function, eg. Euclidean
- d_i^+ and d_j^- shorthand notations for $d(f(x), f(x_i^+))$ and $d(f(x), f(x_j^-))$
- s denotes the [cosine] similarity, with s_i^+ and s_j^- shorthand for $s(f(x), f(x_i^+))$ and $s(f(x), f(x_j^-))$

Note

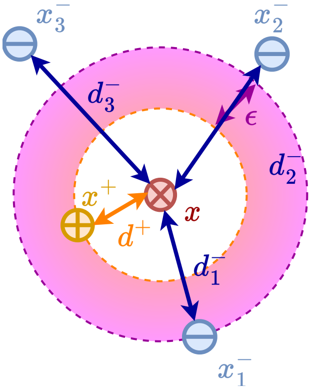
Given that $||f(x)||_2 = 1$, if we choose $d(x, y) = \frac{1}{2}||x - y||_2^2$, then we have $s(x, y) = 1 - d(x, y)$

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

- The condition $s_j^- - s_i^+ \leq -\epsilon \quad \forall i, j$ is equivalent to $\max\{s_j^- - s_i^+\} \leq -\epsilon$;

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;

Derivation of ϵ-SupInfoNCE

- 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. In order to obtain a derivable loss function, we employ *LogSumExp* (LSE), which is a smooth approximation of the max operator:

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

Derivation of ϵ-SupInfoNCE

Using the LSE approximation, we obtain the following loss function, which we call ϵ-SupInfoNCE:

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

Alternative derivations

Please note that other derivations are possibile; some of them are shown in the full paper.

Table: Accuracy on vision datasets. SimCLR and Max-Margin results from [2]. 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).

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 [4]:

The Issue of Biases

3 Debiasing with FairKL

- 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 [4]:
  1. *bias-aligned*: shares the same bias attribute of the anchor. We denote it as $x^{+,b}$

The Issue of Biases

3 Debiasing with FairKL

- 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 [4]:
  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

- 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;

Biases and Failure of ϵ-SupInfoNCE

- 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).

- 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;

FairKL

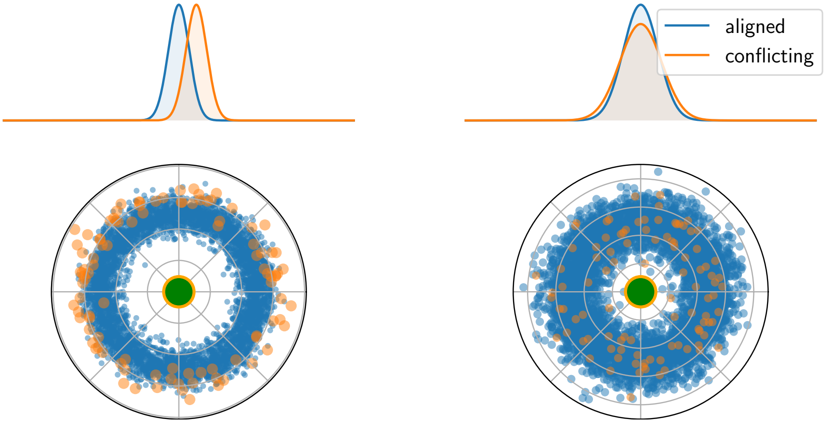
- Assuming that the similarities follow a normal distribution, we denote as $B_{+,b} \sim \mathcal{N}(\mu_{+,b}, \sigma^2_{+,b})$ and $B_{+,b'} \sim \mathcal{N}(\mu_{+,b'}, \sigma^2_{+,b'})$ the **distributions of similarities** of the bias-aligned and bias-conflicting samples respectively;

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

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

FairKL Visualized

3 Debiasing with FairKL



Final Objective

The final obective function $\mathcal{J}$ we minimize becomes:

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

where α and λ are positive hyperparameters.

Results

3 Debiasing with FairKL

Table: Accuracy (%) on Biased-MNIST. Additional experiments available in the paper.

	Correlation (%)			
Method	99.9	99.7	99.5	99
CE [1]	11.8±0.7	62.5±2.9	79.5±0.1	90.8±0.3
LNL [3]	18.2±1.2	57.2±2.2	72.5±0.9	86.0±0.2
EnD [6]	<u>59.5</u>±2.3	82.70±0.3	94.0±0.6	94.8±0.3
BC+BB* [1]	30.26±11.08	82.83±4.17	88.20±2.27	95.04±0.86
BB [1]	76.8±1.6	91.2±0.2	93.9±0.1	96.3±0.2
BC+CE* [1]	15.06±2.22	<u>90.48</u>±5.26	<u>95.95</u>±0.11	<u>97.67</u>±0.09
FairKL	**90.51**±1.55	**96.19**±0.23	**97.00**±0.06	**97.86**±0.02

Table of Contents

Conclusions

1. We test our method on standard debiasing benchmarks, achieving state-of-the-art results

2. Our metric approach allows for a clear and interpretable way of describing the behavior of different loss functions and regularizations

3. Furthermore, the usage of FairKL is not limited to ϵ-SupInfoNCE or contrastive losses

Thanks

- Thank you for listening!
- The code is available on github at `https://github.com/EIDOSLAB/unbiased-contrastive-learning`
- The full-text is available on OpenReview at `https://openreview.net/pdf?id=Ph5cJSfD2XN`

Table of Contents

5 References

References

[1] Youngkyu Hong and Eunho 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.

[2] 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.

References

[3] Byungju Kim et al. "Learning Not to Learn: Training Deep Neural Networks With Biased Data". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

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

References

[5] 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).

[6] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. "EnD: Entangling and Disentangling Deep Representations for Bias Correction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 13508–13517.