Self-Supervised Fair Representation Learning without Demographics

- Junyi Chai and Xiaoqian Wang

이종진

Seoul National University

ga0408@snu.ac.kr

Jan 30, 2023

Self-Supervised Fair Representation Learning without Demographics

- Learning fair representation without sensitive information and even without labels in the classification task.
  - Absence of sensitive information in real scenarios - (privacy, regulation)
- The proposed method is built on fully unsupervised training data and only a small labeled validation set.
  - Unsupervised training data / Contrastive Learning
  - A small labeled validation set / Max-Min Problem

A general fair classification task

- $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{a}_i), 1 \leq i \leq N\}$
  - $\boldsymbol{x}_i$: input data
  - $\boldsymbol{y}_i \in \{0, 1\}^c$: one-hot encoding label
  - $\boldsymbol{a}_i \in \{0, 1\}^s$: the senstivie attribute
- Learning the classifier h with the fairness constraint $\phi(x)$

$$\arg\min_h \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{cls}\left(h\left(\boldsymbol{x}_i\right), \boldsymbol{y}_i\right), \text{ s.t. } \phi(h) \leq \epsilon$$

Max-min Problem

Definition [Rawlsian Max-Min Fairness, (2001, Rawls)]

Suppose H is a set of hypotheses, and $U_{\mathcal{D}_{a'}}(h)$ is the expected utility of the hypothesis h in group $a' \in A'$, then a hypothesis h^* is said to satisfy Rawlsian Max-Min fairness principle if it maximizes the utility of the worst-off group, i.e., the group with the lowest utility.

$$h^* = \arg\max_{h \in H} \min_{a' \in A'} U_{\mathcal{D}_{a'}}(h)$$

- If we choose accuracy as the utility metric and relaxation of error based fairness constraints can be formulated with cross-entropy loss,

$$\arg\min_{h} \max_{\boldsymbol{a}'} \frac{1}{|\{i \mid \boldsymbol{a}_i = \boldsymbol{a}'\}|} \sum_{i \in \{i \mid \boldsymbol{a}_i = \boldsymbol{a}'\}} \mathcal{L}_{cls}\left(h\left(\boldsymbol{x}_i\right), y_i\right)$$

Contrastive Loss

- A Loss for learning a representation on the unit hypersphere based on the similarity of input features
- With each mini-batch of size n, $\{\boldsymbol{x}_i, 1 \leq i \leq n\}$
- Apply random augmentation on each sample twice resulting $\{\tilde{\boldsymbol{x}}_i, 1 \leq i \leq 2n\}$
- Denote $\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_i^{\mathrm{pos}}$ as samples with applying different augmentation to $\boldsymbol{x}_i$
- The contrastive loss with temperature τ with encoder f_θ

$$\mathcal{L}_{ctr}\left(\tilde{\boldsymbol{x}}_i; \theta\right) = -\log \frac{\exp\left(\operatorname{sim}\left(f_\theta\left(\tilde{\boldsymbol{x}}_i\right), f_\theta\left(\tilde{\boldsymbol{x}}_i^{\mathrm{pos}}\right)\right)/\tau\right)}{\sum_{j \neq i} \exp\left(\operatorname{sim}\left(f_\theta\left(\tilde{\boldsymbol{x}}_i\right), f_\theta\left(\tilde{\boldsymbol{x}}_j\right)\right)/\tau\right)}$$

Problem Formulation

- $\{(\boldsymbol{x}_i), 1 \leq i \leq N\}$: unlabeled data
- $\left\{\left(\boldsymbol{x}_j^{\mathrm{lbl}}, \boldsymbol{y}_j^{\mathrm{lbl}}\right), 1 \leq j \leq M\right\}$ with $M \ll N$: labeled data
- f_θ: contrastive encoder
- g_w: linear classifier with learned representation as input.

Proposed Method

- Train f_θ with the weighted contrastive loss with unlabeled data

$$\theta^*(v) = \arg\min_\theta \frac{1}{2N} \left[\sum_{i=1}^{2N} v_i \mathcal{L}_{ctr} \left(\tilde{\boldsymbol{x}}_i; \theta \right) \right]$$

- Train g_w and assign weights with the average top-k labeled loss

$$l^{\mathrm{lbl}}(k, \theta, \omega) = \left[\frac{1}{k} \sum_{j=1}^{M} \left[\mathcal{L}_{cls} \left(g_\omega \left(f_\theta \left(\boldsymbol{x}_j \right) \right), \boldsymbol{y}_j \right) - \lambda^{\mathrm{lbl}}(k, \theta, \omega) \right]_+ + \lambda^{\mathrm{lbl}}(k, \theta, \omega) \right]$$

where

$\lambda(k, \theta, \omega)$ is the k-th largest cross-entropy loss among $\left\{ \mathcal{L}_{cls} \left(g_\omega \left(f_\theta \left(\boldsymbol{x}_j \right) \right) \right) \right\}_{j=1}^{M}$

Proposed Method

- We want to learn a weight assignment for training samples s.t. minimizing the weighted contrastive loss.

$$\theta^*(v) = \arg\min_{\theta} \frac{1}{2N} \left[\sum_{i=1}^{2N} v_i \mathcal{L}_{ctr} \left(\tilde{\boldsymbol{x}}_i ; \theta \right) \right],$$

$$v^*, \omega^* = \arg\min_{v \geq 0, \omega} l^{\mathrm{lbl}} \left(k, \theta^*(v), \omega \right).$$

Weight Approximation

- At iteration t, $l_{t,i} = \mathcal{L}_{ctr}\left(\tilde{x}_i;\theta\right)$ and denote labeled loss l_t^v
- We use a simple approximation of the optimal weight based on the inner product between gradients.

$$u_{t,i} = \left(\nabla_\theta l_t^{\mathrm{lbl}}\right)^\top \nabla_\theta l_{t,i}$$

- Assign weights at iteration t with

$$v_{t,i} = \frac{2n\hat{v}_{t,i}}{\sum_{i'=1}^{2n} \hat{v}_{t,i'} + \delta\left(\sum_{i'=1}^{2n} \hat{v}_{t,i'}\right)}$$

where $\delta(r) = 1 \iff r = 0$ and $\hat{v}_{t,i} = \max\left(u_{t,i}, 0\right)$

Algorithm 1: Optimization Algorithm

Pre-train the encoder f_θ on the labeled set $\left\{\left(\boldsymbol{x}_j^{\mathrm{lbl}}, \boldsymbol{y}_j^{\mathrm{lbl}}\right), 1 \le j \le M\right\}$

for *for* $t = 0, 1 \cdots, T-1$ *do* **do**

1. Sample a mini-batch of training samples of size n, apply random augmentation on each sample twice and get a unlabled set $\{\tilde{x}_i, 1 \le i \le 2n\}$;
2. Calculate contrastive loss $\{\mathcal{L}_{ctr}(\tilde{x}_i; \theta)\}_{i=1}^{2n}$ denote it as $\{l_{t,i}\}_{i=1}^{2n}$;
3. Freeze f_θ and fine-tune the linear layer g_ω on labeled set;
4. Calculate labeled loss $l^{\mathrm{val}}(k, \theta, \omega)$ denote it as $l_t^{\mathrm{lbl.}}$
5. Update $\hat{v}_{t,i} = \max\left(\left(\nabla_\theta l_t^v\right)^\top \nabla_\theta l_{t,i}, 0\right)$
6. Update $v_{t,i} = \frac{2n\hat{v}_{t,i}}{\sum_{i'=1}^{2n} \hat{v}_{t,i'} + \delta\left(\sum_{t'=1}^{2n} \hat{v}_{t,i'}\right)}$, where $\delta(r) = 1 \iff r = 0$;
7. Update $\theta_{t+1} = \theta_t - \frac{1}{2n}\nabla_\theta \sum_{i=1}^{2n} v_{t,i} l_{t,i}$;

end

Convergence Proof

Assumption 3.1

1. *The partial derivative of labeled loss l^{lbl} with respect to θ is Lipschitz continuous with constant L, i.e., $\nabla^2_{\omega\theta} l^{\mathrm{val}}$ and $\nabla^2_{\theta\theta} l^{\mathrm{val}}$ are upper-bounded by L.*
2. *The contrastive loss l has σ-bounded gradients w.r.t. θ.*

Theorem 3.2

Under Assumption 3.1 at iteration t, let the learning rate of contrastive encoder f satisfies $\alpha_{1,t} \leqslant \frac{4\sigma^2 L \sum_t \beta_{t,i}^2}{n \sum_t \left(\beta_{t,i}^2 - 2\gamma_{t,i}\beta_{t,i}\right)}$, and the learning rate of linear classifier satisfies $\alpha_{2,t} \leq \min\left(\frac{2}{L}, \frac{\sum_t \beta_{t,i}^2}{L \sum_t \gamma_{t,i}\beta_{t,i}}\right)$, where

$$\gamma_{t,i} = \left\| \nabla_\omega l_t^{lbl} \right\| \left\| \nabla_\theta l_{t,i} \right\|, \quad \beta_{t,i} = \left(\left(\nabla_\theta l_{t,i}\right)^\top \nabla_\theta l_t^{lbl} \right),$$

then the labeled loss will monotonically decrease until convergence.

Reference

- J. Rawls. 2001., Justice as fairness: A restatement. Harvard University Press.