MaxEnt-ARL: Mitigating Information Leakage in Image Representations

Proteek Chandan Roy and Vishnu Naresh Boddeti
Michigan State University, East Lansing MI 48824
{royprote, vishnu}@msu.edu

Introduction: As we witness the wide spread adoption of representation learning systems, it is imperative to consider the problem of unintended leakage of information from such systems. *Adversarial Representation Learning* (ARL) is a promising approach for learning image representations that minimizes such leakage of user information. These approaches couple together (i) an adversarial network that seeks to classify and extract sensitive information from a given representation, and (ii) an embedding network that is tasked with extracting a compact representation of data while preventing the adversarial network from succeeding at leaking sensitive information. To achieve their respective goals, the adversary is optimized to maximize the likelihood of the sensitive information, while the encoder is optimized to minimize the same likelihood i.e., adversary's likelihood of the sensitive information, thereby leading to a zero-sum game. This approach referred to as ML-ARL has been leveraged for learning censored [2], fair [3], or invariant [4] representations of data.

The zero-sum game formulation of optimizing the likelihood, however, is practically sub-optimal from the perspective of preventing information leakage. Moreover, the potential of this formulation to prevent information leakage is predicated upon: (i) the existence of an equilibrium, and (ii) the ability of practical optimization procedures to converge to such an equilibrium. when the optimization does not reach the equilibrium, a probability distribution with the minimum likelihood is the distribution that is most certain with the potential to leak the most amount of information.

Building on the observations above, we propose a framework, dubbed *Maximum Entropy Adversarial Representation Learning* (MaxEnt-ARL), which optimizes an image representation with two major objectives, (i) maximally retain information pertinent to a given target attribute, and (ii) minimize information leakage about a given sensitive attribute. We pose the learning problem in an adversarial setting as a non-zero sum three player game between an encoder, a predictor and a discriminator (proxy adversary) where the encoder tries to maximize the entropy of the discriminator on the sensitive attribute and maximizes the likelihood of the predictor on the target attribute.

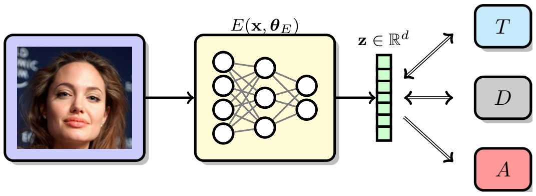

$E(\mathbf{x}, \boldsymbol{\theta}_E)$
$\mathbf{z} \in \mathbb{R}^d$
T
D
A

Figure 1: **Adversarial Representation Learning:** We consider the problem of learning an embedding function E that maps a high-dimensional image to a low-dimensional representation $\mathbf{z} \in \mathbb{R}^d$ while satisfying two competing goals: retain as much image information necessary to accurately predict a target attribute while simultaneously minimizing information leakage about a sensitive attribute by an unknown adversary A. The learning problem is formulated as a game between $\{E, T\}$ and a proxy adversary D.

Main Contribution: In the MaxEnt-ARL formulation the goal of the encoder is to maximize the likelihood of the target attribute, as measured by the *target predictor*, while maximizing the uncertainty in the sensitive attribute, as measured by the entropy of the *discriminator's* prediction. The *encoder* is modeled as a deterministic function, $\boldsymbol{z} = E(\boldsymbol{x}; \boldsymbol{\theta}_E)$, the *target predictor* models the conditional distribution $p(t|\boldsymbol{x})$ via $q_T(t|\boldsymbol{z}; \boldsymbol{\theta}_T)$ and the *discriminator* models the conditional distribution $p(s|\boldsymbol{x})$ via $q_D(s|\boldsymbol{z}; \boldsymbol{\theta}_D)$, where $p(t|\boldsymbol{x})$ and $p(s|\boldsymbol{x})$ are the ground truth labels for a given target and sensitive labels t and s, respectively. Formally, we define the MaxEnt-ARL optimization problem as a three player non-zero sum game:

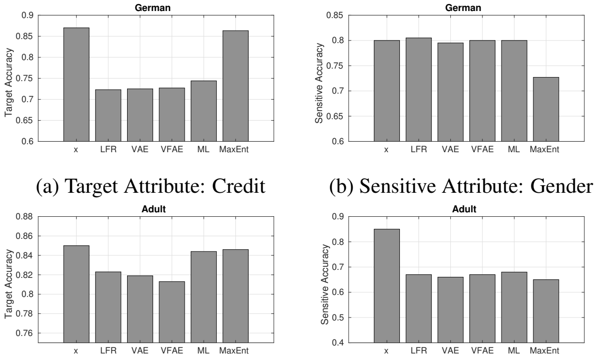
$$\begin{aligned} &\min_{\boldsymbol{\theta}_D} V_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) \\ &\min_{\boldsymbol{\theta}_E, \boldsymbol{\theta}_T} V_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) + \alpha V_3(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) \end{aligned} \tag{1}$$

where α allows us to trade-off between the two competing objectives for the encoder and,

$$\begin{aligned} V_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) &= KL\left(p\left(s|\boldsymbol{x}\right) \| q_D\left(s|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_D\right)\right) \\ V_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) &= KL\left(p\left(t|\boldsymbol{x}\right) \| q_T\left(t|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_T\right)\right) \\ V_3(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) &= KL\left(q_D\left(s|E(\boldsymbol{x}; \boldsymbol{\theta}_E)\,; \boldsymbol{\theta}_D\right) \| U\right) \end{aligned}$$

where U is the uniform distribution.

Experimental Results: We evaluate MaxEnt-ARL and compare to ML-ARL on two tasks, fair classification on the UCI and on the CIFAR-100 datasets.



(a) Target Attribute: Credit

(b) Sensitive Attribute: Gender

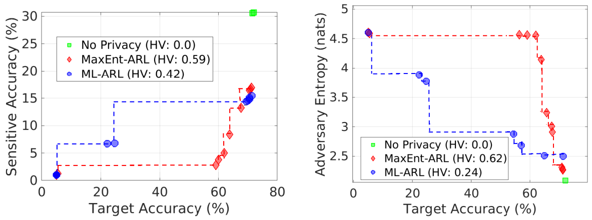
(c) Target Attribute: Income

(d) Sensitive Attribute: Gender

Figure 2: Representation Learning for Fair Classification

Fair Classification: We consider the setting of fair classification on two datasets from the UCI ML-repository [1], (a) The German credit dataset with 20 attributes for 1000 instances with target label being classifying bank account holders with good or bad credit and gender being the sensitive attribute, (b) The Adult income dataset has 45,222 instances with 14 attributes. The target is a binary label of annual income more or less than $\$50,000$, while gender is the sensitive attribute. For both ML-ARL and MaxEnt-ARL, the encoder is a NN with one hidden layer, discriminator is a NN with 2 hidden layers, and target predictor is linear logistic regression. Following ML-ARL [4] we choose 64 units in each hidden layer. We compare both ARL formulations with state-of-the-art baselines LFR (Learning Fair Representations), VAE (Variational Auto-Encoder) and VFAE (Variational Fair Auto-Encoder). For MaxEnt-ARL, after learning the embedding, we again learn an adversary to extract the sensitive attribute.

Figure 2 show the results for the German and Adult datasets, for both the target and sensitive attributes. For German data, MaxEnt-ARL's prediction accuracy is 86.33% which is close to that of the original data (87%). Other models such as, LFR, VAE, VFAE and ML-ARL have target accuracies of 72.3%, 72.5%, 72.7% and 74.4% respectively. On the other hand, for the sensitive attribute, the MaxEnt-ARL adversary's accuracy is 72.7%. Other models reveal much more information with adversary accuracies of 80%, 80.5%, 79.5%, 79.7% and 80.2% for the original data, LFR, VAE, VFAE and ML-ARL, respectively. For the adult income dataset, the target accuracy for original data, ML-ARL and MaxEnt-ARL is 85%, 84.4% and 84.6%, respectively, while the adversary's performance on the sensitive attribute is 67.7% and 65.5% for ML-ARL and MaxEnt-ARL, respectively.



(a)

(b)

Figure 3: **CIFAR-100:** Trade-off fronts for two different ARL approaches, ML-ARL and MaxEnt-ARL, in comparison to standard no privacy representation learning. In (a) the ideal desired solution is the bottom right corner, while in (b) it is the top right corner. HV in the legend corresponds to normalized hyper-volume.

CIFAR-100: We formulate a new privacy problem on the CIFAR-100 dataset. The dataset consists of 100 classes and are grouped into 20 superclasses. Each image has a "fine" (the class to which it belongs) and a "coarse" (the superclass to which it belongs) label. We treat the "coarse" (superclass) and "fine" (class) labels as the target and sensitive attribute, respectively. So the encoder is tasked to learn features of the super-classes while not revealing the information of the underlying classes. We adopt ResNet-18 as the encoder while the predictor, discriminator and adversary are all 2-layered fully connected networks. We report the trade-off achieved between predictor and adversary along with the corresponding normalized hyper-volume (HV) in Fig. 3. Our results indicate that, representation learning without privacy considerations leaks significant amount of information. MaxEnt-ARL is able to significantly outperform ML-ARL on this task, achieving trade-off solutions that are far better, both in terms of adversary accuracy and entropy of adversary.

References

[1] D. Dua and C. Graff. UCI machine learning repository, 2017. 2

[2] H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[3] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[4] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2