CENSORING REPRESENTATIONS WITH MULTIPLE-ADVERSARIES OVER RANDOM SUBSPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

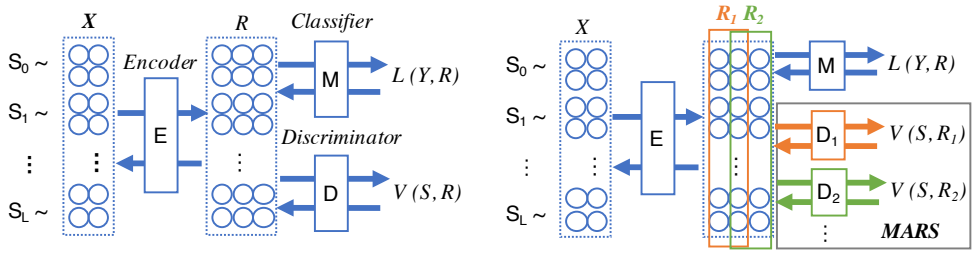
In practice, often there are explicit constraints on the representations that are acceptable in a machine learning application in the real world; for example, representations of data must not contain identifying information so as to avoid privacy issues. This paper improves the performance of the recently proposed adversarial feature learning approach (AFL) for incorporating such explicit constrains, by introducing the concept of the *vulnerableness* of the adversary. In AFL, the censoring representation is done by training the networks to deceive the adversary that try to predict the sensitive information from the network, and the success of the AFL relies on the choice of the adversary. This motivate the use of high capacity networks as an adversary for improving the performance; however, this approach does not work well in practice as reported in this paper. Instead of the capacity of networks, this paper proposes to consider the *vulnerableness* in design of the adversary, i.e., the adversary should be designed not to be easily fooled. We also propose a method *multiple adversaries over random subspaces* (MARS) that instantiate the concept, and provides empirical validations on the efficacy of the proposed method compared to the various baselines, indicating the importance of the proposed concept. This is significant because it gives new implications about designing the adversary, which is important to improve the performance of AFL framework.

1 INTRODUCTION

Since its invention over 10 years ago (Hinton et al., 2006), deep neural networks (DNN) have shown significant performance improvements in various fields. When we apply DNN or more generally machine learning techniques to real-world data, one of the key challenges is how to systematically incorporate the desired constraints into the learned representations in a controllable manner. For example, when these techniques are applied to the data that contain a lot of user information (such as images with user name (Edwards & Storkey, 2016) or data of wearables (Iwasawa et al., 2017)), the desired representations should not contain user information that may result in privacy issues. Moreover, for legal and ethical reasons, machine learning algorithms have to make fair decisions, which do not rely on sensitive variables such as gender, age, or race (Louizos et al., 2016; Edwards & Storkey, 2016). This requires removal of information related to specific factors (such as user ID, race etc.) from the representation; this is called *censoring representations* in this paper.

One of the recently proposed approaches for censoring representation is *adversarial feature learning* (AFL) (Edwards & Storkey, 2016; Iwasawa et al., 2017; Xie et al., 2017), which employs the adversarial training framework to constrain the representations (Figure 1-a). Specifically, AFL considers an adversarial classifier who attempts to predict sensitive variables from the representations of a DNN, and simultaneously trains the DNN to deceive the classifier. By alternatively or jointly (using gradient reversal layer proposed by Ganin & Lempitsky (2015)) training the adversary and DNN in such a manner, AFL ensures that there is little or no information about the sensitive variables in the representations.

Although some previous studies report significant performance improvements of the AFL in the context of censoring representations, the success of the AFL depends on the choice of the adversarial classifier. In other words, AFL can only eliminate the information that the adversary can correctly predict. For example, if we use a logistic regression as the adversarial classifier, AFL can only



(a) AFL with an adversary

(b) Multiple Adversaries over Random Subspaces

Figure 1: Multiple Adversaries over Random Subspaces (MARS) for censoring representations: (a)AFL with an adversary, (b) AFL with MARS. MARS is designed to have diverse-view and to be less vulnerable to meaningless update of the encoder E.

eliminate the information that is linearly separated in the representation spaces and cannot eliminate any non-linear dependency. This motivates the use of high-capacity networks as an adversary, such as DNNs for removing complex information more correctly; however, this approach does *not* work well in practice as reported in (Iwasawa et al., 2017) and this paper. This implies that only the *capacity* of the adversary is not a determinant factor of the success of adversarial feature learning; it is still unclear what is required for improving the quality of AFL.

In this paper, we propose a method for improving the quality of AFL, *multiple-adversaries over random subspace* (MARS), which consider the *vulnerableness* of the adversary. The term vulnerableness indicates how the adversarial classifier is easy to be fooled when updating the target networks (encoder in Figure 1-a) to deceive the adversary. We assume that an adversary is too easy to be fooled as it relies on single view of a network, resulting in the failure of censoring representations. In contrast, the proposed method, MARS, is designed to have diverse view by incorporating multiple adversaries where each adversary tries to predict sensitive variables from a different *subset* of the representations. This makes adversary less vulnerable to the update as the target networks need to deceive a set of diverse adversaries. In this paper, we empirically validate that (1) MARS archives better performance compared with a single adversary and multiple adversaries over the entire representation spaces, (2) using the subsets of features improve the vulnerableness compared to using the entire spaces, which implying that introducing diversity for adversaries improves the vulnerableness and the performance of AFL.

The primary contributions of this paper are as follows:

- This is the first study that introduces the concept of vulnerableness in AFL, and empirically validates the importance of the concept by showing the superior performance of the proposed method that instantiate the concept of vulnerableness. This is significant because it gives implications about designing the adversary in AFL, which is important in practice.
- The proposed method achieved significant performance gain in the task of censoring representations. The empirical validation using three user-anonymization tasks shows that the proposed method allows the learning of significantly more anonymized representations with negligible performance degradation. Specifically, the probability of correctly predicting the user ID from learned representations is 0.124 and 0.112 points better on average than that of a single adversary and multiple adversaries over entire representation spaces, respectively.

2 Problem Definition and Related Works

2.1 Problem Definition: Censoring Representations

Censoring representation means obtaining unbiased features. Here, *unbiased features* are features that are less affected by S, where S is a random variable that we want to remove from the data for some reason. One typical reason is related to fairness or privacy, which requires the output of neural networks not to be affected by unfair information or not contain user information.

It should be noted that poor design of the censoring procedure significantly reduces the utility of data. For example, the output of random mapping f_{rand} apparently has no information about S,

but it also gives no information about target Y. Alternatively, as a more realistic example, a neural network with limited capacity possibly acquires less information about S, but it may also result in poorer performance. Therefore, the primary goal of censoring representation is to obtain an encoder E that reduces information about S, while maintaining information about Y. Formally, the task can be written as a joint optimization problem of the loss:

$$J = L(E(X), Y) - \lambda V(E(X), S),$$

where X indicates the input random variable, E is an encoder that transforms X to representation R, λ is the weighting parameter, and V and L are loss functions that represent how much information about S and Y is present, respectively. Note that S can be any form of variable such as binary variable, categorical variable, or continuous variable. In this paper, we especially consider a special variant of censoring representation tasks, where we learn E with deep neural networks and S is the user ID (anonymization tasks).

2.2 Related Works: Adversarial Feature Learning

A recently proposed approach for the censoring representation task is *adversarial feature learning* (AFL). In AFL, $V(E(X), S)$ is measured by an external neural network D, i.e., if the external networks can accurately predict S from R, AFL regards that R has too much information about S and if it is difficult to accurately predict S, AFL regards that R has little or no information about S. The external network D is called discriminator or adversary in this context. The information is used to update the weights of the encoder E so that the updated representations have less information about S. Formally, AFL solves the joint optimization problems:

$$\min_{E,M} \max_{D} \mathbb{E}[-\log q_M(y|h = E(x)) + \lambda \log q_D(s|h = E(x))],$$

where $M : R \to Y$ is the label classifier, $D : R \to S$ is the adversarial classifier (see Figure 1), and q_M and q_D are conditional distributions parameterized by M and D, respectively. The optimization procedure is called adversarial training because E and D are in an adversarial relationship (E tries to minimize of log-likelihood q_D and D tries to maximize it).

To the best of the authors knowledge, the adversarial training framework was first introduced by (Schmidhuber, 1992), and later re-invented by (Goodfellow et al., 2014) in the context of an image-generation task. In the context of censoring representations, Edwards & Storkey (2016) first proposed the use of adversarial training to remove sensitive information from the representations. They show its efficacy for fair classification tasks (S is binary) compared to an LFR (Learned Fair Representation) proposed by Zemel et al. (2013) that regularizes the l_1 distance between the distributions for data with different S. Iwasawa et al. (2017) first applied the adversarial training to learn anonymized representations on the data of wearables, where S is categorical. More recently, Xie et al. (2017) first introduced the notion of *adversarial feature learning* and showed superior performances compared to variational fair auto-encoder (Louizos et al., 2016).

Although AFL already shows state-of-the-art performance to learn unbiased representations, how to improve its performance is still a challenge, which is tackled in this paper. This work is motivated by previous studies conducted by Iwasawa et al. (2017). They reported that simply using high-capacity networks as an adversary (such as deep neural networks) sometimes *fails* to improve the performance; on the other hand, the leaned representations are still highly biased if the adversary does not have enough capacity. Our work can be regarded as introduction of the novel design consideration (i.e., vulnerableness) of adversary for improving the quality of AFL and validation of it by proposing a method that instantiate the concept, which is not done in the previous works.

It is worth mentioning that the concept of multiple adversaries itself is already proposed and verified in the context of the image generation with adversarial training (Durugkar et al., 2017), and this paper does not insist that using multiple adversaries is the primal novelty. From the methodological perspective, the proposed method (MARS) can be seen as an extension of the concept of multiple adversaries by introducing the concept of diversity, and we verify that this extension is important in the context of censoring representations. Though the primary focus of this paper is limited to the context of adversarial feature learning, we further discuss the applicability of the proposed method on the other application of adversarial training, such as image/text generation (Goodfellow et al., 2014; Zhang et al., 2017) or domain adaptation(Ganin & Lempitsky, 2015); this is done at the end of this paper.

3 MULTIPLE-ADVERSARIES OVER RANDOM SUBSPACES

The proposed method, *multiple-adversaries over random subspaces (MARS)*, considers multiple adversaries where each adversary is in charge of different subsets of features. The development of MARS is motivated by the assumption that the adversary should not be vulnerable and it could be achieved by enforcing the adversary to have diverse view. Figure 1 shows the comparison of overall architectures of the proposed method between AFL with an adversary, and with multiple adversaries over random subspaces (i.e., proposed method). Since the primary difference between naive AFL and MARS is how to build the adversary (training adversaries) and how to use the information of adversary (training encoders), the explanation is focused on the part.

3.1 TRAINING ADVERSARIES

There are many possible options to select a subspace; however, this paper specifically considers the case where each subspace is randomly selected. Suppose that n is the number of dimensions of representation $R \in \mathbb{R}^n$, and K is the number of adversaries. In MARS, each adversary D_k is trained to predict S over a randomly selected subset of features R^k, whose dimension is m_k and $m_k < n$. Each adversary is trained to maximize the expected log-likelihood. Formally, the optimization problem for D_k is as follows

$$\max_{\theta_{D_k}} \mathbb{E}[\log q_{D_k}(s|h_k = Sub_k(E(x)))], \tag{1}$$

where Sub_k is a function that return the subset of R, which is fixed before the training. Specifically, each Sub_k determine whether to remove the i-th dimension of R with the probability of α. The θ_{D_k} is the parameter of D_k.

3.2 TRAINING ENCODER

The adversaries are then integrated and a single prediction is made to train the encoder. Formally, the probability distribution of $q_D(s = s_i|h = E(x))$ is parameterized by an ensembled adversary D:

$$q_D(s = s_i|h = E(x)) = \frac{1}{K}\sum_{k}^{K} q_{D_k}(s|h_k), \tag{2}$$

which means, the predictions of adversaries D_k are integrated by averaging the predictions. Note that integration operator F is not essential to be the averaging, it may be the max operator or the soften version of the max operator proposed in (Durugkar et al., 2017).

The encoder is trained for minimizing the log-likelihood of Eq.2 and negative log-likelihood $q_M(y|h)$. Formally,

$$\min_{\theta_E, \theta_M} \mathbb{E}[-\log q_M(y|h) + \lambda \log \frac{1}{K}\sum_{k}^{K} q_{D_k}(s|h_k)], \tag{3}$$

where θ_E and θ_M are the parameter of E and M respectively. The overall algorithm is shown in Algorithm 1.

Algorithm 1 Optimization of the proposed model

Require: parameter λ, α
  Initialize Sub_k for all $k \in 1 \cdots K$ based on α
  Initialize neural networks $\{E, M, D_1, \cdots D_K\}$
  while training() **do**
    Update weights of D_k for all $k \in 1 \cdots K$ (eq. 1)
    Update weights of E and M (eq. 3)
  end while

4 EXPERIMENTS

4.1 DATASETS

We used two datasets to demonstrate the efficacy of MARS. Both datasets are related to human activity recognition using the data of wearables, where privacy issues have recently been pointed out by (Iwasawa et al., 2017). In both datasets, the task is to learn anonymized representations (R that does not contain user-identifiable information), while maintaining classification performance.

The opportunity recognition dataset (Sagha et al., 2011) consists of sensory data regarding human activity in a breakfast scenario, and the task is to predict the activity performed by a user. A wide variety of body-worn, object-based, and ambient sensors are used (see Figure 1 in Sagha et al. (2011) for more details). Each record consists of 113 real-value sensory readings, excluding time information. We considered two recognition tasks: gesture recognition (Opp-G) and locomotion recognition (Opp-L). Opp-G requires the recognition of 18 class activities [1], while Opp-L requires the recognition of 4 class locomotion-activities, i.e., stand, walk, sit, and lie. Given a sampling frequency of 30 Hz, the sliding window procedure with 30 frames and a 50% overlap produced 57,790 samples. We used data of subjects 2–4 as training/validation sets (90% for training and 10% for validation) and that of subject 1 as test sets.

The USC-HAD dataset (Zhang & Sawchuk, 2012) is a relatively new benchmark dataset, which contains a relatively large number of subjects (14 subjects: 7 males and 7 females). There are 12 activity classes corresponding to the most basic and common activities of people's daily lives [2]. MotionNode, which is a 6-DOF inertial measurement unit specially designed for human motion sensing applications, is used to record the outputs from accelerometers that record 6 real sensory values. The sliding window procedure, using 30 frames and a 50% overlap, produced 172,169 samples. We used data of subjects 1–10 as training/validation sets (90% for training and 10% for validation) and that of subjects 11–14 as test sets.

4.2 EXPERIMENTAL SETTING

In all experiments, we parameterized the encoder E by convolutional neural networks (CNN) with three convolution-ReLU-pooling repeats followed by one fully connected layer and M by a logistic regression, in accordance with a previous study (see Figure 2 in Iwasawa et al. (2017) for more details). Every component of the network was trained with the Adam algorithm (learning rate is set to 0.0001) (Kingma & Ba, 2015) (150 epochs). In the optimization, we used the annealing heuristics to weighting parameter λ in accordance with (Iwasawa et al., 2017); we linearly increased λ during 15 epochs to 135 epochs of training.

To demonstrate the efficacy of the proposed method, we compare the following four methods and its variants: (1) None: w/o adversary (correspond to normal CNN), (2) Adv: w/ a single adversary, (3) MA: w/ multiple adversaries where each adversary tries to predict from the entire space, and (4) MARS: w/ multiple adversaries where each adversary tries to predict from a different subspace of the entire space. Each adversary is parametrized by multi-layer perceptron (MLP) with 800 hidden units. If we need to express the type of adversary, we denote it by a suffix. For example, Adv-LR means an adversary is parametrized by a logistic regression and MA-LR means that each adversary is parametrized by a logistic regression. The hyper-parameter λ is set to 1.0, 1.0, 0.1 for Opp-G, Opp-L, and USC-HAD respectively. Without mentioning otherwise, we set the number of adversaries K as 10 for MA and MARS, and α for determine size of the subset as 0.8.

Following previous works, we evaluated the level of anonymization by training a classifier f_{eva} that tries to predict S over learned representations. This means that we regarded that a learned representation is anonymized well if we could not build an accurate classifier over the representations. Unless mentioned otherwise, we use a multi-layer perceptron with 800 hidden units as f_{eva}. We later discuss how the choice of classifier for f_{eva} and D affects the results by performing empirical evaluations.

[1] open door 1, open door 2, close door 1, close door 2, open fridge, close fridge, open dishwasher, close dishwasher, open drawer 1, close drawer 1, open drawer 2, close drawer 2, open drawer 3, close drawer 3, clean table, drink from cup, toggle switch, and Null

[2] walking forward, walking left, walking right, walking upstairs, walking downstairs, running forward, jumping, sitting, standing, sleeping, elevator up, elevator down