Deep Representation Debiasing via Mutual Information Minimization and Maximization (Student Abstract)

Ruijiang Han[1,2], **Wei Wang**[1,2], **Yuxi Long**[1,2], **Jiajie Peng**[1,2*]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, China
{hanruijiang, wangwei0206, yxlong}@mail.nwpu.edu.cn, jiajiepeng@nwpu.edu.cn

Abstract

Deep representation learning has succeeded in several fields. However, pre-trained deep representations are usually biased and make downstream models sensitive to different attributes. In this work, we propose a post-processing unsupervised deep representation debiasing algorithm, DeepMinMax, which can obtain unbiased representations directly from pre-trained representations without re-training or fine-tuning the entire model. The experimental results on synthetic and real-world datasets indicate that DeepMinMax outperforms the existing state-of-the-art algorithms on downstream tasks.

Introduction

Deep representation algorithms pre-train encoder models and directly extract representations for downstream tasks. Representation models usually suffer from population bias (Madhavan and Wadhwa 2020). The encoder learns biased relationships between latent variables during pre-training, and inherits bias from the pre-training dataset. Biased deep representation may cause domain shift problems and unfairness, thus making debiasing an urgent need.

Existing study has proven that disentangle representation and bias without any given attributes is impossible (Locatello et al. 2019). Therefore, representation debiasing aims to extract representations that are independent of given sensitive attributes. Existing debiasing algorithms can be loosely grouped as pre-, in- or post-processing methods. Pre-processing methods try to normalize the original data and fit the data distribution precisely for debiasing efficiently (Zheng et al. 2019). In-processing methods apply additional regularization or optimization objectives while re-training or fine-tuning the entire model (Ganin et al. 2016; Sanchez, Serrurier, and Ortner 2020). Post-processing methods directly debias from the output of pre-trained encoder, and are more flexible and scalable. However, most existing post-processing methods assume certain relations between the unbiased representation and bias, which limit their extensiveness and discrimination ability (Georgopoulos, Panagakis, and Pantic 2020; Madhavan and Wadhwa 2020).

In this work, we developed a post-processing method, DeepMinMax, to eliminate bias from pre-trained encoders without supervision labels and re-training procedures. DeepMinMax is based on mutual information minimization and maximization, which do not assume the distribution and relationship of variables.

Proposed Method

Problem Formulation Let X and B be a sample and its sensitive attributes in the dataset $\{\mathcal{X}, \mathcal{B}\}$. For a given pre-trained encoder $f : X \rightarrow Z$, we can extract the original biased representation Z from X. We assume that Z is controlled by B and other properties of X. Post-processing methods aims to obtain a debiasing function $h_\theta^{f,\mathcal{B}} : Z \rightarrow \tilde{Z}$ to extract unbiased $\tilde{Z}$ from Z, where $\tilde{Z} \perp\!\!\!\perp B$.

We uses a multi-layer perceptron (MLP) as h_θ and propose two objectives: debiasing via mutual information (MI) minimization and retaining via MI maximization.

Debiasing MI measures the dependence between two variables. Minimizing the MI of $\tilde{Z}$ and B can directly remove the bias. However, MI of high-dimensional continuous variables cannot be calculated exactly. We use the contrastive log-ratio upper bound of mutual information (CLUB) as a gradient estimation (Cheng et al. 2020). The debiasing objective can be described by MI as:

$$\min_\theta \mathcal{L}^d = \hat{I}_{\widetilde{CLUB}}\left(B; \tilde{Z}\right) \geq I(B; \tilde{Z}) \tag{1}$$

Retaining A shortcut to meet the above objective is to let h_θ give a constant C as output. To avoid these collapse solutions, we need additional objectives to ensure $\tilde{Z}$ is discriminative enough. We assume the encoder f to be well pre-trained so that the Z contains the required information of X. A straightforward idea is to let $\tilde{Z}$ preserve the information contained in Z by maximizing the MI of them.

Since Z is biased, forcing $\tilde{Z}$ to retain the complete information in Z conflicts with the previous objective $\min_\theta \mathcal{L}^d$. Therefore, we maximize the MI of Z and the joint of $\tilde{Z}$ and B. According to the identities of MI, we can decompose the joint MI as:

$$I(Z; \tilde{Z}, B) = H(Z) - H(Z \mid \tilde{Z}, B) \tag{2}$$

where $H(Z)$ is the information entropy and independent with θ measuring the uncertainty of Z, and $H(Z \mid \tilde{Z}, B)$

[*]Corresponding author.

Method	Stage	Super- vised	RU. →UM.	UM. →RU.	both →RS.	both →CW.	both →LP.
ACD	pre	no	0.974	0.935	0.810	0.820	0.885
RevGrad	in	semi	0.963	0.946	0.923	0.942	0.968
DRMIE	in	yes	**0.979**	**0.949**	0.939	0.950	0.972
KANFace	post	yes	0.936	0.928	0.912	0.929	0.931
PFR	post	yes	0.928	0.909	0.906	0.904	0.938
ours	post	no	0.973	0.942	**0.948**	**0.955**	**0.985**

Table 1: Method properties and AUROC on Camelyon17

is the conditional entropy which measures Z's uncertainty under given $\tilde{Z}$ and B. Then, we have the retaining objective:

$$\min_{\theta} \mathcal{L}^r = \mathbb{E}_{p_{Z,\tilde{Z},B}} \left[-\log m_{\phi}(Z \mid \tilde{Z}, B) \right] \approx H(Z \mid \tilde{Z}, B) \quad (3)$$

where m_{ϕ} is a MLP as the variational approximation to the unknown conditional distribution and is updated by negative log-likelihood minimization in each training iteration.

Finally, we formulate the full objective as $\min_{\theta} \lambda \mathcal{L}^d + (1-\lambda)\mathcal{L}^r$, where $\lambda \in [0, 1]$ is used to balance two objectives.

Experiments

Datasets and Comparison Methods

Synthetic MNIST is a handwritten digit dataset synthesized conditionally. We train a supervised InfoGAN to generate an image with a specified digit and two latent codes modeling rotation and width. *Camelyon* is a real-world breast pathology dataset consisting of two parts. All training procedures are conducted on Camelyon16 collected from RUMC and UMCU. Camelyon17 from five medical centers (including RUMC and UMCU) are used for testing. The task is to predict the presence of metastasis.

We compare DeepMinMax with the five state-of-the-art debiasing methods, including ACD (Zheng et al. 2019), RevGrad (Ganin et al. 2016), DRMIE (Sanchez, Serrurier, and Ortner 2020), KANFace (Georgopoulos, Panagakis, and Pantic 2020) and PFR (Madhavan and Wadhwa 2020). Their properties are listed in Table **??**.

Implementation Details

The MLP has 3 layers in h_{θ} and 2 layers in m_{ϕ}, with hidden dimensions same as input and activated by ReLU. The λ is 0.5, the batch size is 128, the initial learning rate is 0.005, and the epochs are 200. We use the logistic regression model as the downstream classifier. The encoder of *Synthetic MNIST* is the backbone of the InfoGAN discriminator. The encoder of *Camelyon* is the backbone of a SimCLR pretrained ResNet-50.

Results

Synthetic MNIST Figure 1a shows test samples encode by the pre-trained encoder, visualized by t-SNE, and colored by rotation codes. As shown in Figure 1b, DeepMinMax eliminates the effect of latent codes on the representation and makes it more discriminative. The accuracy of digit recognition has increased from 0.935 using biased representations to 0.994 using DeepMinMax representations.

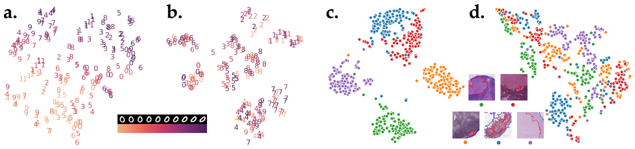

Figure 1: Visualization of DeepMinMax representations.

Camelyon Samples from different centers have significant visual differences. We regard center identities as the sensitive attributes. Camelyon17 samples represented by the pre-trained encoder are shown in Figure 1c. DeepMinMax makes samples from the same center no longer closely clustered, indicating the elimination of the domain bias (Figure 1d). The test AUROC is shown in Table **??**. The left side of the column name represents the source of supervision labels in Camelyon16, and the right side is the tested center in Camelyon17. DeepMinMax matches supervised and inprocessing methods' performance on dual-domain adaption tasks, and performs best on blind domain adaption tasks in entirely unseen medical centers.

Conclusions

We have developed a novel unsupervised post-processing method named DeepMinMax to eliminate bias in deep representation learning via mutual information minimization and maximization. DeepMinMax can obtain unbiased and discriminative representations for non-specific distributions and relationships. The results on synthetic and real-world tests show that DeepMinMax effectively eliminates sensitive attributes and improves performance in downstream tasks.

References

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *Proc. 37th Int. Conf. Mach. Learn.*

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*.

Georgopoulos, M.; Panagakis, Y.; and Pantic, M. 2020. Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and vision computing*.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. 36th Int. Conf. Mach. Learn.*

Madhavan, R.; and Wadhwa, M. 2020. Fairness-Aware Learning with Prejudice Free Representations. In *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*

Sanchez, E. H.; Serrurier, M.; and Ortner, M. 2020. Learning Disentangled Representations via Mutual Information Estimation. In *Proc. Eur. Conf. Comput. Vis.*

Zheng, Y.; Jiang, Z.; Zhang, H.; Xie, F.; Shi, J.; and Xue, C. 2019. Adaptive color deconvolution for histological WSI normalization. *Computer methods and programs in biomedicine*.