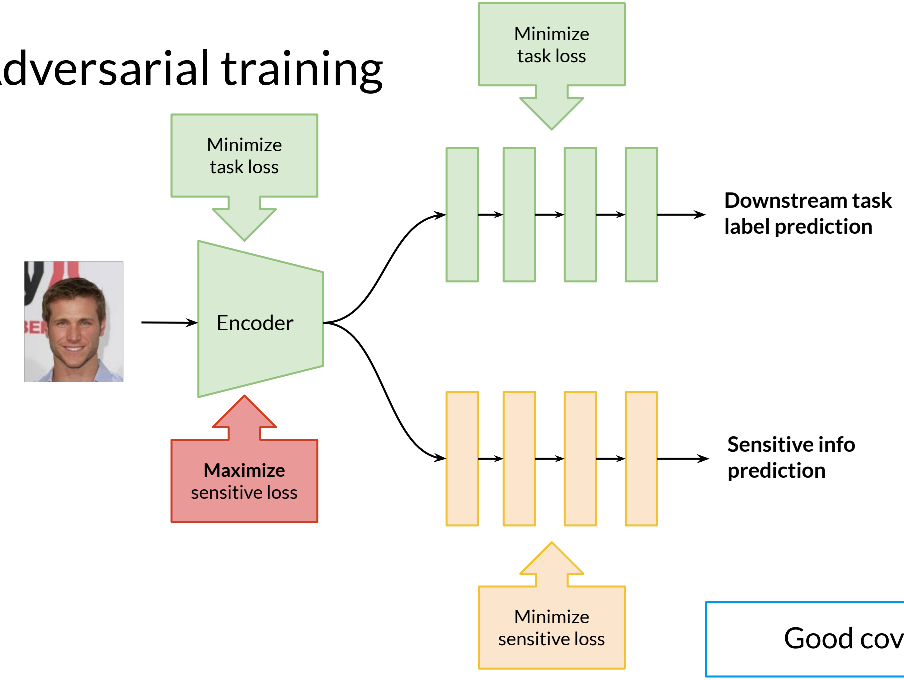
Model Debiasing via Gradient-based Explanation on Representation

Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, Lei Chen

Previous Model Debiasing Schemes

Adversarial training



- **CON: Lack of Flexibility, hard to train**
  Model is fixed with the downstream task and sensitive attribute. Cannot reuse the encoder part if the downstream task or sensitive attribute is changed.

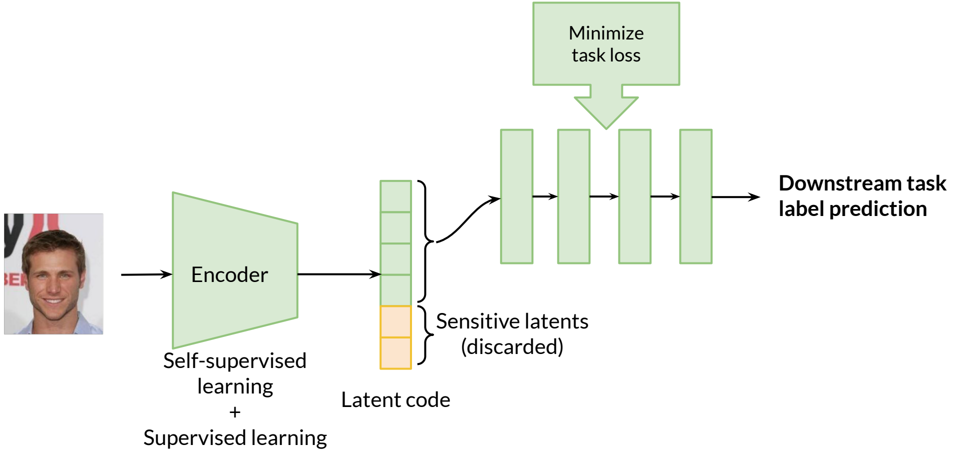
- **PRO: Good coverage of sensitive info in feature/latent code**
  Maximizing sensitive loss works for all weights of the encoder, so that every dimension of the learnt feature is affected.

Good coverage of sensitive info in the feature

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, Domain-adversarial training of neural networks, The journal of machine learning research, 17 (2016), pp. 2096–2030.

Previous Model Debiasing Schemes

Disentangled Representation learning



Decoupling the representation learning process and the downstream task

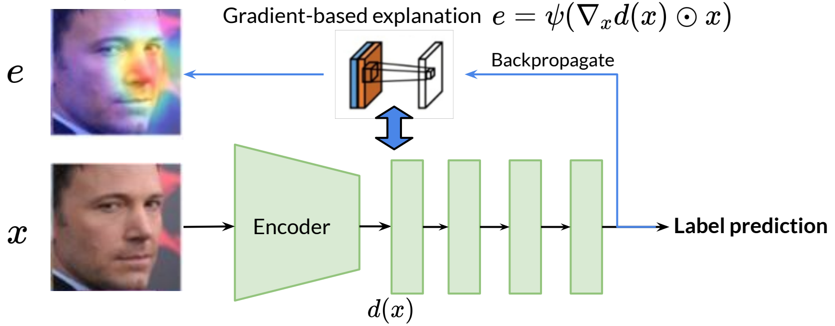
- **PRO: Flexible**
  The representation learning process and the downstream task label prediction are decoupled. If downstream task is changed, encoder does not need retraining. **But if sensitive attribute is changed, it still needs to be retrained.**
- **CON: Poor coverage of sensitive info in feature/latent code; losing downstream task info**
  The method is to disentangle sensitive info from non-sensitive info in the latent code, which cannot be done perfectly. Some sensitive info still remains in non-sensitive dimensions, and some useful info in sensitive dimensions.

E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, Flexibly fair representation learning by disentanglement, in ICML, PMLR, 2019, pp. 1436–1445.

Our Idea: Leveraging Gradient-based Explanation

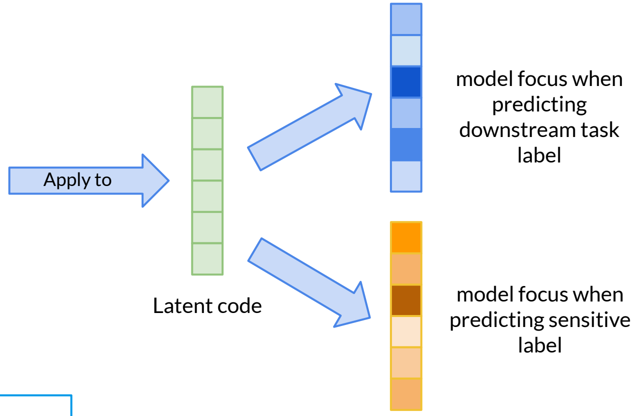
Gradient-base XAI

Highlighted area indicates where the model focuses for prediction



Why Gradient-base XAI? Decoupling from the feature learning process and Good coverage of sensitive info in feature/latent code

- **Leverage gradient-based XAI to obtain the model focus on latent code when predicting sensitive label and downstream task label**



DVGE: Debiasing via Gradient-based Explanation

Latent code $z = f(x)$

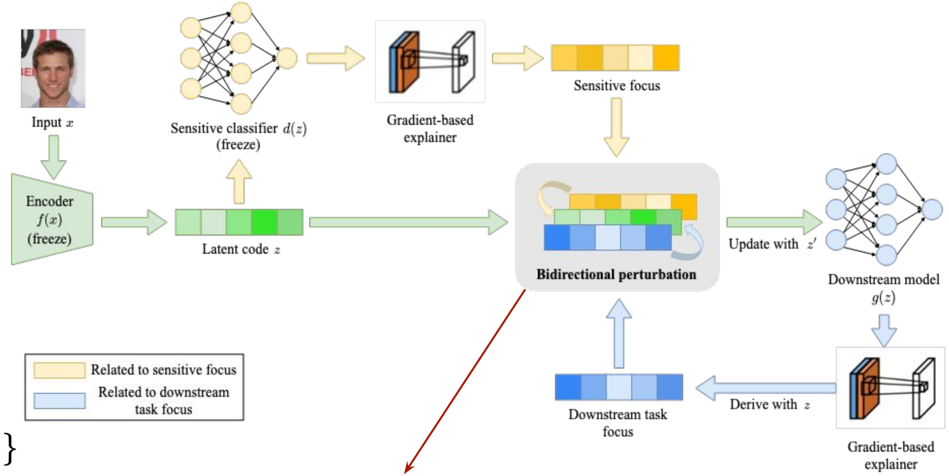
Sensitive focus $F_{sens} = \nabla_z d(z)$

Downstream task focus $F_{task} = \nabla_z g(z)$

Bidirectional perturbation

$$z' = z + \underbrace{Clip_\epsilon}_{\text{Prevent introducing too much information distortion}}\{\overbrace{\eta_1 \times F_{sens}}^{\text{Guide model to focus \textbf{less} on sensitive info, debiasing model}} - \underbrace{\eta_2 \times F_{task}}_{\text{Guide model to focus \textbf{more} on downstream task info, boosting model performance}}\}$$

$$Clip_\epsilon\{v\} = \begin{cases} v, & \text{if } v > \epsilon \\ \max(v, -\epsilon), & \text{otherwise} \end{cases}$$



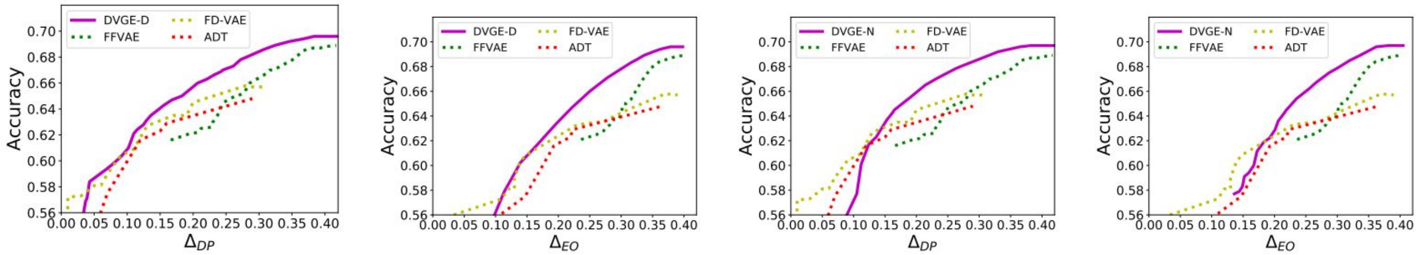
Works similarly to **adversarial training** while keeping the framework decoupled

Experimental Setups

- Metric
  - Fairness-accuracy trade-off
  - Demographic Parity (DP) $\Delta_{DP} = |P(\hat{y} = 1 \mid s = s_1) - P(\hat{y} = 1 \mid s = s_2)|$
  - Equal Opportunity (EO) $\Delta_{EO} = |P(\hat{y} = 1 \mid s = s_1, y = 1) - P(\hat{y} = 1 \mid s = s_2, y = 1)|$
- Datasets
  - **CelebA**: 202,599 facial images, each of which is associated with 40 attributes, such as "Attractive", "Male", "Young". And all attributes are in binary form.
  - **South German Credit**: 1,000 entries with 21 attributes. The first 20 attributes are about the loan applicants (gender, age, income, etc.), and the last one is the application result.
- Baselines
  - Adversarial training (ADT)
  - FFVAE
  - FD-VAE: separates the latent code into three portions, i.e., sensitive dimensions, downstream-task-related dimensions, and mutual-information dimensions
- Our framework
  - DVGE-D: \w disentangled VAE
  - DVGE-N: \w non-disentangled VAE

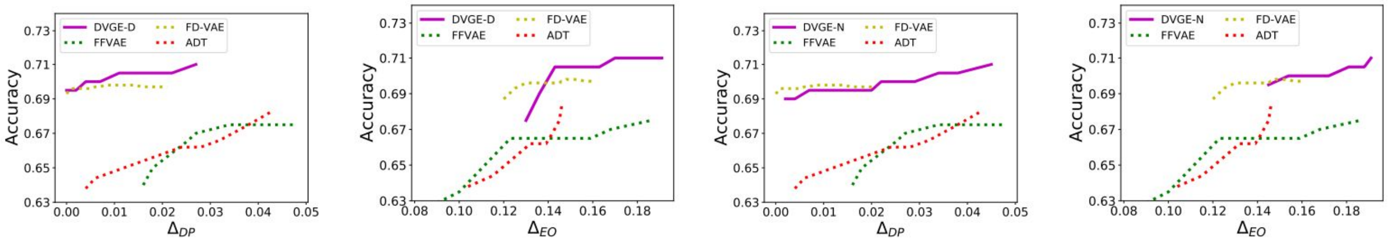
1. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, Domain-adversarial training of neural networks, The journal of machine learning research, 17 (2016), pp. 2096–2030.
2. E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, Flexibly fair representation learning by disentanglement, in ICML, PMLR, 2019, pp. 1436–1445.
3. S. Park, S. Hwang, D. Kim, and H. Byun, Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment, in Proceedings of AAAI, vol. 35, 2021, pp. 2403–2411.

Experiment: Single Sensitive Attribute



(a) DVGE-D, Δ_{DP} (b) DVGE-D, Δ_{EO} (c) DVGE-N, Δ_{DP} (d) DVGE-N, Δ_{EO}

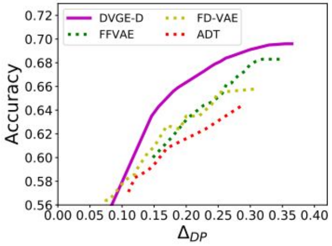
CelebA dataset, sensitive attribute = “Male”, task label = “Oval Face”.



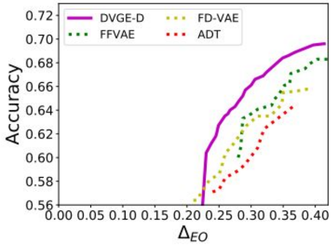
(a) DVGE-D, Δ_{DP} (b) DVGE-D, Δ_{EO} (c) DVGE-N, Δ_{DP} (d) DVGE-N, Δ_{EO}

South German Credit dataset, sensitive attribute = “age”, task label = “credit risk”.

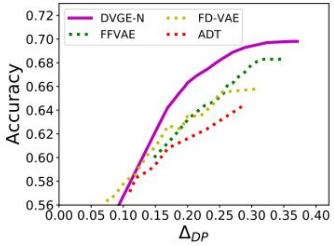
Experiment: Multiple Sensitive Attributes



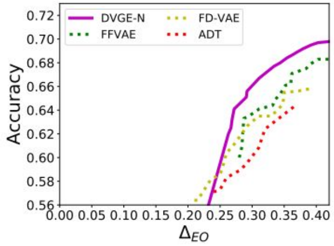
(a) DVGE-D, Δ_{DP}



(b) DVGE-D, Δ_{EO}

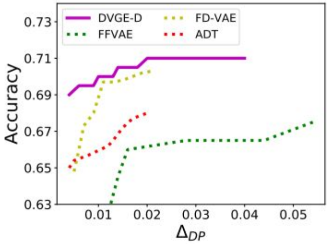


(c) DVGE-N, Δ_{DP}

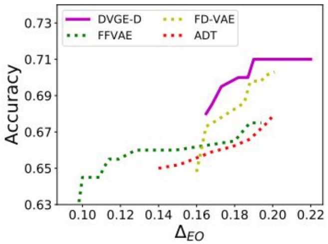


(d) DVGE-N, Δ_{EO}

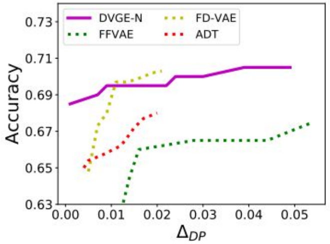
CelebA dataset, sensitive attribute = “Male” ∧ “Young”, task label = “Attractive”.



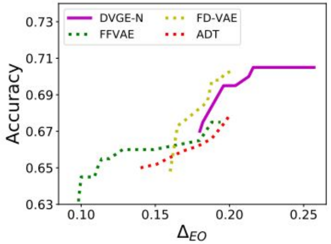
(a) DVGE-D, Δ_{DP}



(b) DVGE-D, Δ_{EO}



(c) DVGE-N, Δ_{DP}



(d) DVGE-N, Δ_{EO}

South German Credit dataset, sensitive attribute = “age” ∧ “foreign worker”, task label = “credit risk”.

Ablation

- Evaluating the coverage on sensitive information in our framework
- Metric
  - The highest accuracy of the sensitive classifiers trained with the **perturbed** latent codes.
  - **Lower accuracy indicates less sensitive information in the perturbed latent code**, and thus further indicates better coverage on sensitive information.

Table 1: Debiasing performance of DVGE in the setting of single sensitive attribute

Encoder	No removal	Sens. dim. removed	DVGE with η_1									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Disentangled	0.798	0.736	0.767	0.735	0.706	0.682	0.675	0.661	0.655	0.658	0.650	0.648
Non-disentangled	0.804	0.746	0.769	0.733	0.705	0.692	0.686	0.682	0.682	0.674	0.671	0.668

Table 2: Debiasing performance of DVGE in the setting of multiple sensitive attributes

Encoder	No removal	Sens. dim. removed	DVGE with η_1									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Disentangled	0.752	0.690	0.732	0.707	0.680	0.661	0.644	0.638	0.637	0.633	0.633	0.631
Non-disentangled	0.757	0.704	0.736	0.709	0.683	0.664	0.657	0.653	0.653	0.651	0.653	0.649

Conclusion

- We propose a fairness framework DVGE to address poor coverage on sensitive information and the loss of useful downstream task information when using representation learning to debias model.
- We introduce to exploit gradient-based explanation to obtain model focuses related to sensitive info and and downstream task info, and propose bidirectional perturbation to guide the model training for fairness purpose with the focuses.
- Experiments on two datasets demonstrate that DVGE achieves better fairness-accuracy trade-off and better coverage on sensitive information while not relying on complete disentanglement for debiasing.

Q&A

Thanks