Fair Contrastive Learning for Facial Attribute Classification

Sungho Park[1] Jewook Lee[1] Pilhyeon Lee[1] Sunhee Hwang[2] Dohyung Kim[3] Hyeran Byun[1]

[1]Yonsei University [2]LG Uplus [3]SK Inc.

Introduction

- Contrastive learning has achieved preeminent success in classification tasks
- However, fairness in terms of sensitive attribute (e.g., gender and race) has been underexplored in previous works
- We prove that the state-of-the art contrastive learning method (*SupCon) has two factors causing unfairness*
  - Learning unwanted sensitive attribute information
  - Data imbalance between demographic groups
- Fair Supervised Contrastive Loss (FSCL) prevents encoding networks from learning sensitive attribute information
- Group-wise normalization mitigates the group-wise disparities due to data imbalance

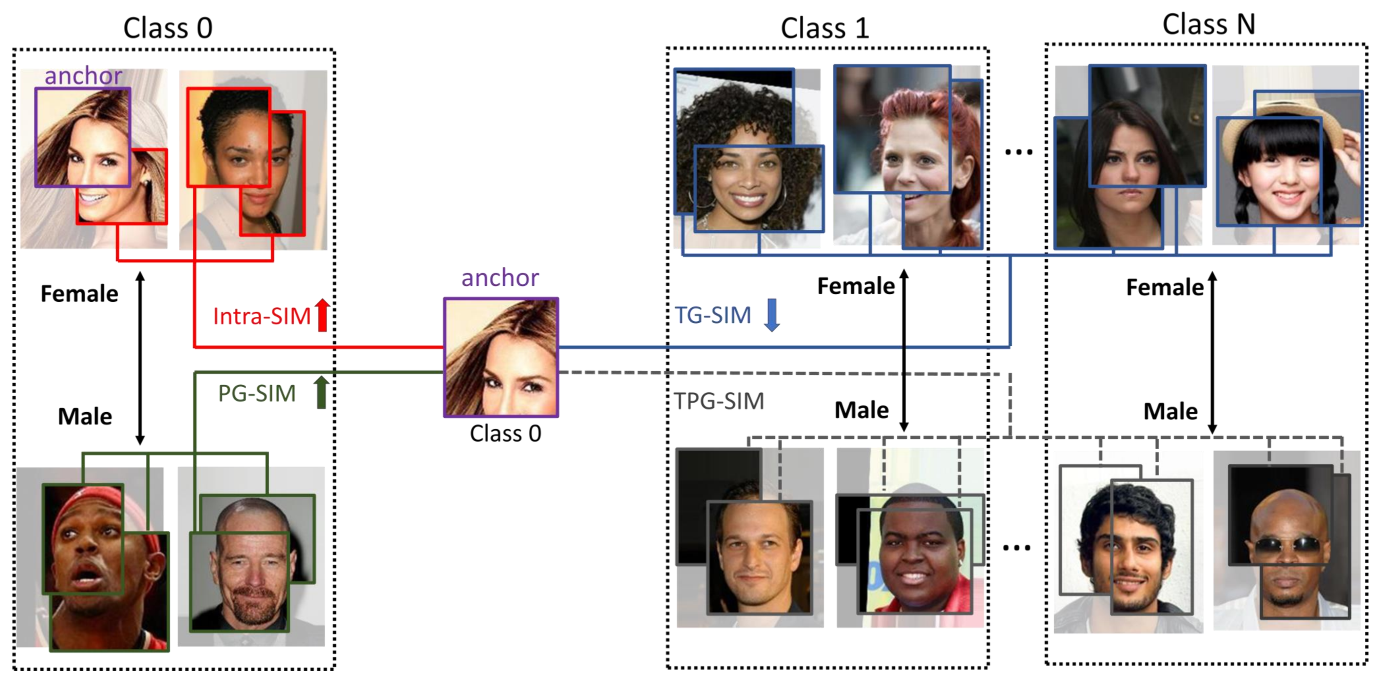
Method

- Contrastive learning draws positive samples and pushes away negative samples from an anchor

SupCon	FSCL
Positive samples - Same target classes with the anchor **Negative samples** - Different target classes from the anchor - Same/different sensitive attributes from the anchor	**Positive samples** - Same target classes with the anchor **Negative samples** - Different target classes from the anchor - Same sensitive attributes from the anchor
⇨ Learning not only target class but sensitive attribute information	⇨ Learning only target class information

- Data imbalance incurs the group-wise disparities of intra-group compactness and inter-class separability
- Group-wise normalization solves them by normalizing the number of anchors and positive samples between groups

The concept of FSCL



Conclusion

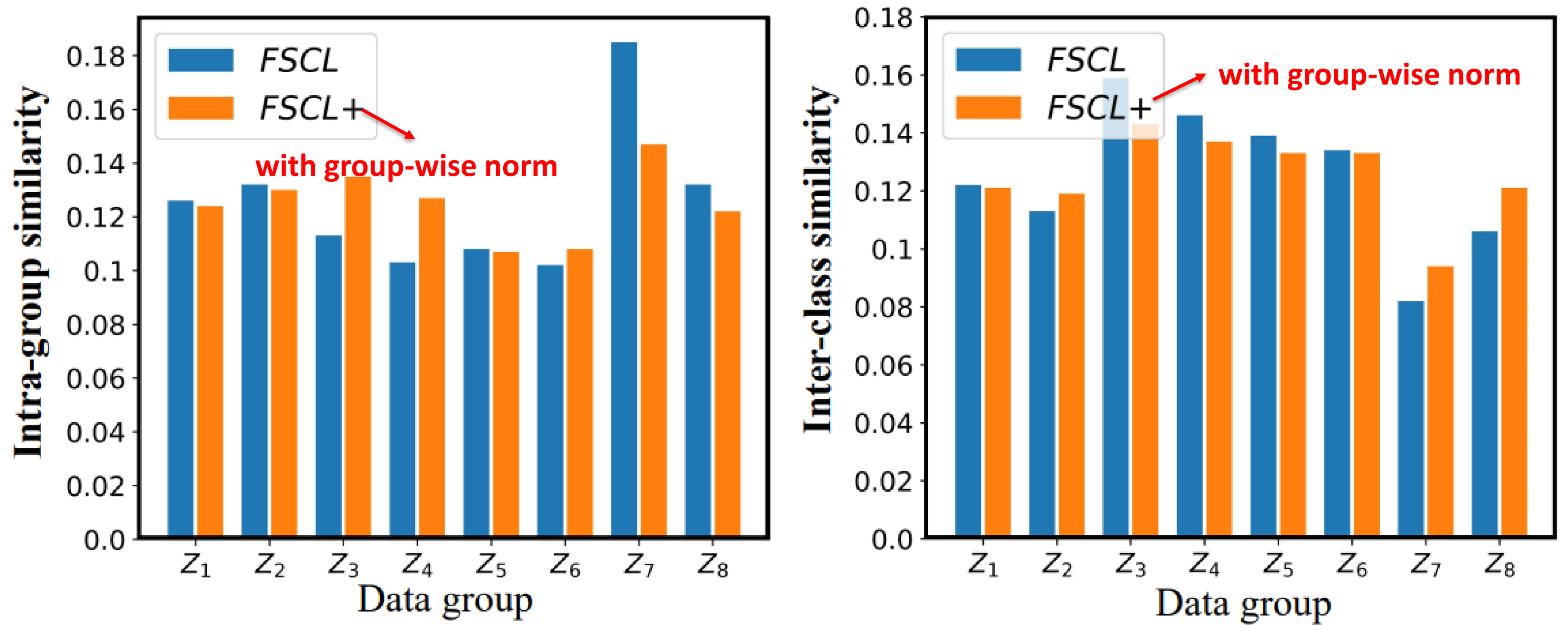
- We analyzed the causative factors of unfairness in contrastive learning
- We proposed a fair contrastive loss (FSCL) and group-wise normalization solving the causes
- Our method achieves the best trade-off performance and works efficiently in various challenging environments

Experiment

Classification results on CelebA

Method	T=*a* / S=*m*		T=*a* / S=*y*		T=*b* / S=*m*		T=*b* / S=*y*	
	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.
CE [15]	27.8	79.6	16.8	79.8	17.6	84.0	14.7	84.5
GRL [38]	24.9	77.2	14.7	74.6	14.0	82.5	10.0	83.3
LNL [26]	21.8	79.9	13.7	74.3	10.7	82.3	6.8	82.3
FD-VAE [37]	15.1	76.9	14.8	77.5	11.2	81.6	6.7	81.7
MFD [22]	7.4	78.0	14.9	80.0	7.3	78.0	5.4	78.0
SupCon [25]	30.5	80.5	21.7	80.1	20.7	84.6	16.9	84.4
FSCL	11.5	79.1	13.0	79.1	7.0	82.1	6.4	83.8
FSCL+	**6.5**	79.1	**12.4**	79.1	**4.7**	82.9	**4.8**	84.1

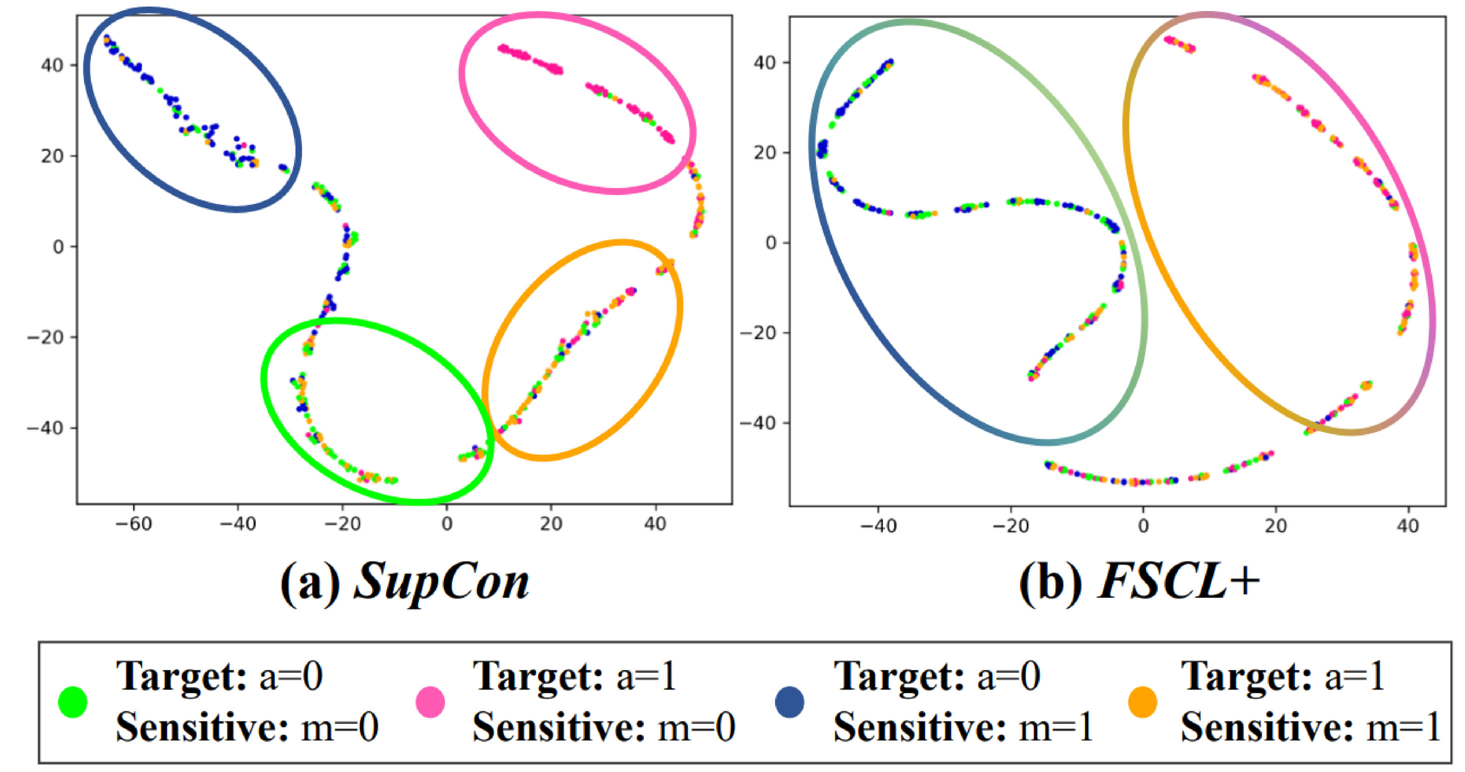
Effectiveness of group-wise normalization



Efficiency in semi-supervised settings

Method	# of Sensitive	Pseudo-labeling	EO (↓)	Acc. (↑)
SupCon [25]	0	-	$30.5_{\pm 1.3}$	$80.5_{\pm 0.7}$
SupCon [25] + *GRL* [38]	1	-	$21.0_{\pm 0.5}$	$76.6_{\pm 0.3}$
FSCL+	1	-	$\mathbf{6.5}_{\pm 0.4}$	$79.1_{\pm 0.1}$
	1/2	✗	$13.4_{\pm 0.1}$	$79.3_{\pm 0.3}$
		✓	$12.8_{\pm 1.2}$	$79.4_{\pm 0.3}$
	1/4	✗	$18.7_{\pm 0.3}$	$80.0_{\pm 0.3}$
		✓	$13.4_{\pm 0.1}$	$79.5_{\pm 0.5}$
	1/10	✗	$20.7_{\pm 0.5}$	$80.2_{\pm 0.1}$
		✓	$16.5_{\pm 0.5}$	$79.6_{\pm 0.4}$
	1/20	✗	$23.4_{\pm 0.0}$	$80.6_{\pm 0.1}$
		✓	$18.8_{\pm 1.1}$	$78.5_{\pm 0.2}$

t-SNE visualization



(a) *SupCon* (b) *FSCL+*