Fair-VPT: Fair Visual Prompt Tuning for Image Classification

Sungho Park * Hyeran Byun *

Department of Computer Science and Engineering, Yonsei University

Presented by Sungho Park

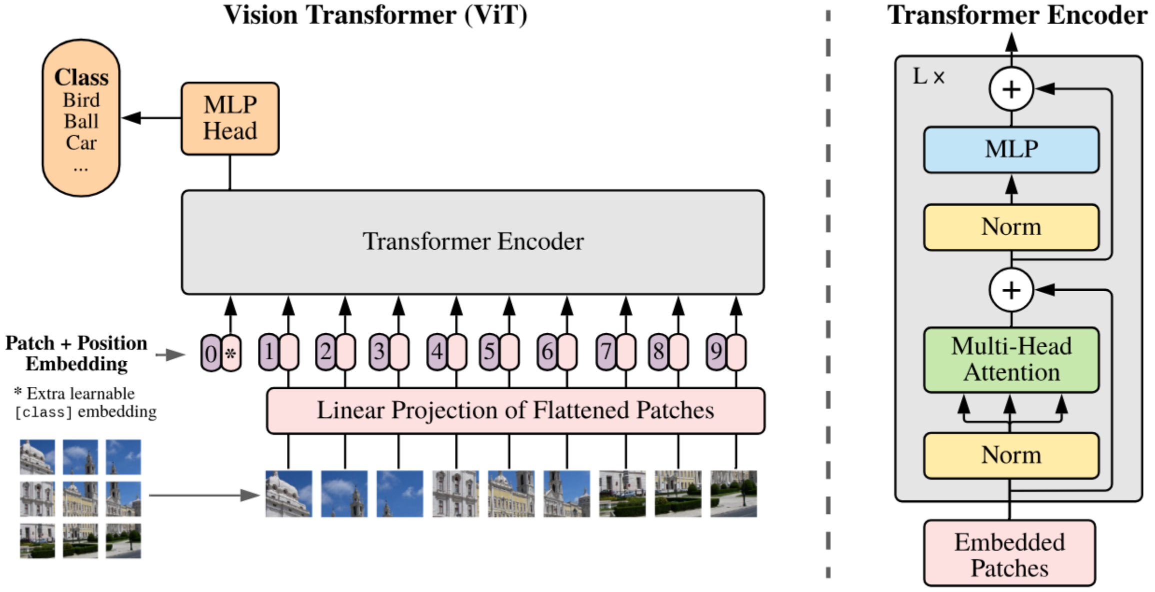
Contact: yunisomi@naver.com

ViT, ICLR 2021, Dosovitskiy et al.

Introduction

Two limitations of Vision Transformer (ViT)

1. High adaptation cost for downstream tasks
2. Severe unfairness with respect to sensitive attributes



Fairness

Method	TL	SA s=0	SA s=1	BAcc. (↑)	EO (↓)
ViT [14]	t=0	99.1	54.3	74.8	$\boxed{48.9}$
	t=1	46.3	99.5		
VPT [24]	t=0	98.9	48.3	76.0	46.3
	t=1	57.5	99.5		
VPT [24]+AT [45]	t=0	99.5	58.7	77.5	43.1
	t=1	53.1	98.7		
Fair-VPT (Ours)	t=0	99.1	62.3	**80.7**	**37.1**
	t=1	61.9	99.5		

Classification results on bFFHQ

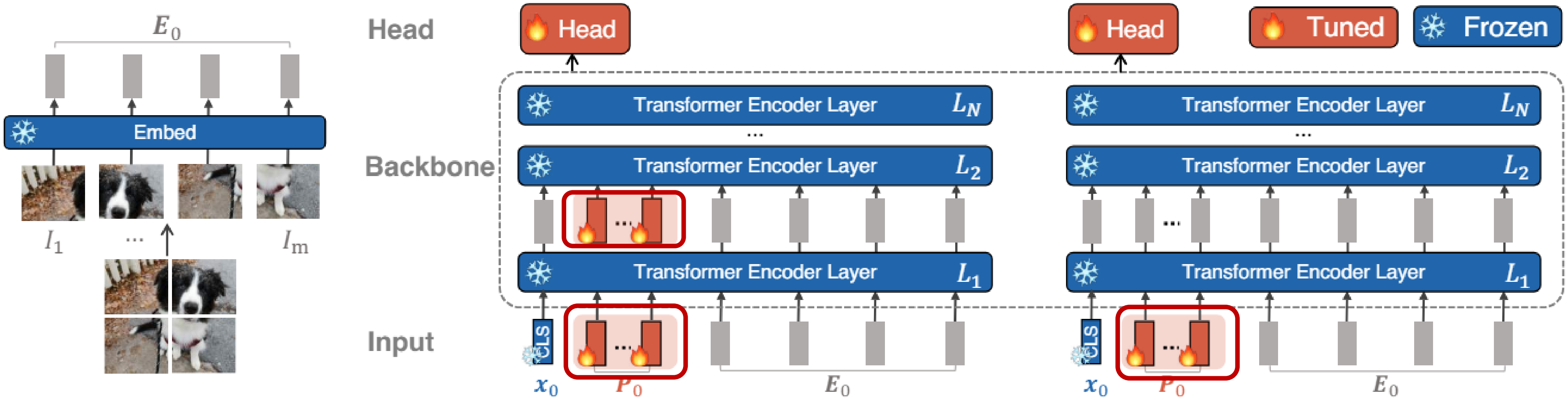
ViT, ICLR 2021, Dosovitskiy et al.

VPT, ECCV 2022, Jia et al.

Introduction

Visual Prompt Tuning (VPT)

- Effectively reducing adaptation cost in transfer learning
- Not addressing the unfairness problem with respect to sensitive attributes



(a) Visual-Prompt Tuning: Deep (b) Visual-Prompt Tuning: Shallow

Introduction

ViT, ICLR 2021, Dosovitskiy et al.

VPT, ECCV 2022, Jia et al.

Exploring primary factors of unfairness

- A pre-trained ViT, prompts, and a classification head
- ***The pre-trained ViT** stands out as the primary factor of unfairness*

Fairness

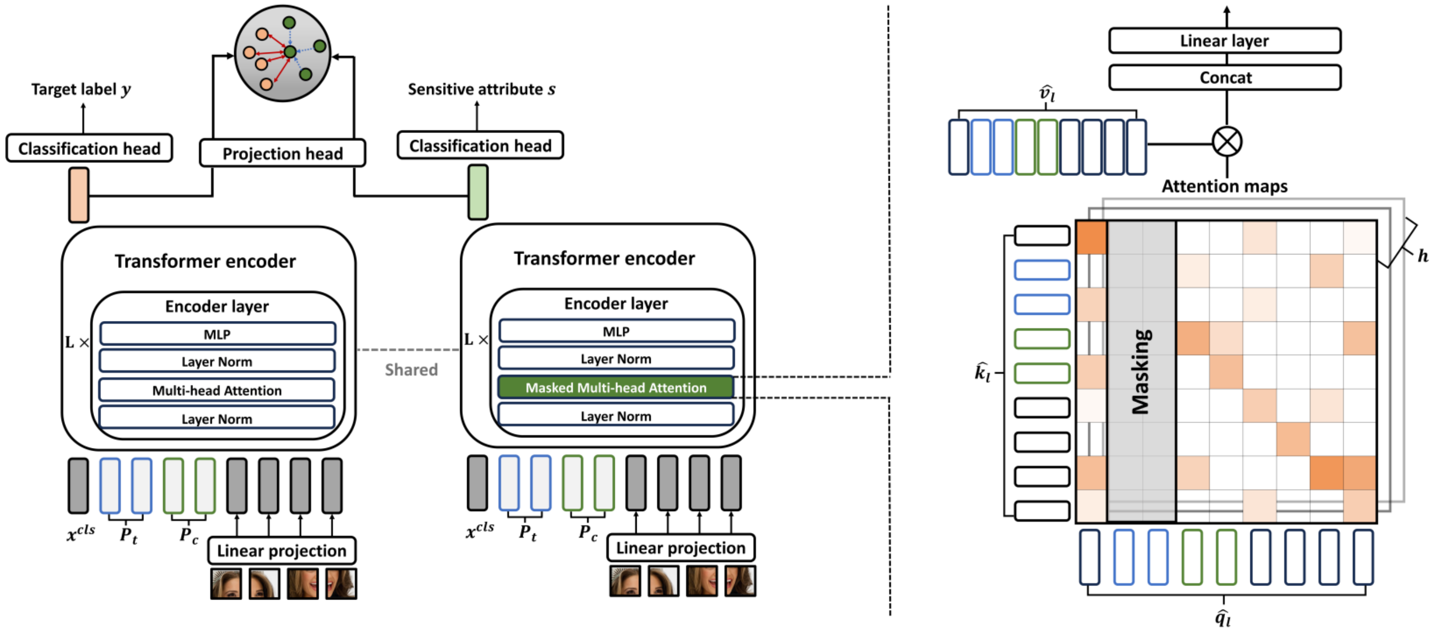
Method	TA	SA *M*	SA *F*	Acc. (↑)	EO (↓)
VPT [24]	*A*	52.7	93.1	**81.7**	**32.1**
	NA	89.1	65.2		
VPT [24]-Head+NCM [39]	*A*	68.1	84.6	76.1	47.4
	NA	99.4	21.0		
ViT [14]+NCM [39]	*A*	20.6	92.3	69.4	82.3
	NA	99.6	6.7		

Classification results on CelebA

Introduction

Fair Visual Prompt Tuning (Fair-VPT)

- Goal: enhancing both fairness and efficiency of ViT in transfer learning
- Removing bias information in the pre-trained ViT and adapting it to downstream tasks



Method

Fairness definition

- **Equalized Odds (EO): ensuring the equality of TPR and FPR between sensitive groups**

- $EO = \frac{|TPR_{s=0} - TPR_{s=1}| + |FPR_{s=0} - FPR_{s=1}|}{2}$

Preliminaries

- Input space $\hat{z}_0 = [x^{cls}, \underbrace{P^{(1)}, ..., P^{(M)}}_{\text{Prompts}}, \underbrace{E(x_p^{(1)}), ..., E(x_p^{(N)})}_{\text{Embedded patches}}]$

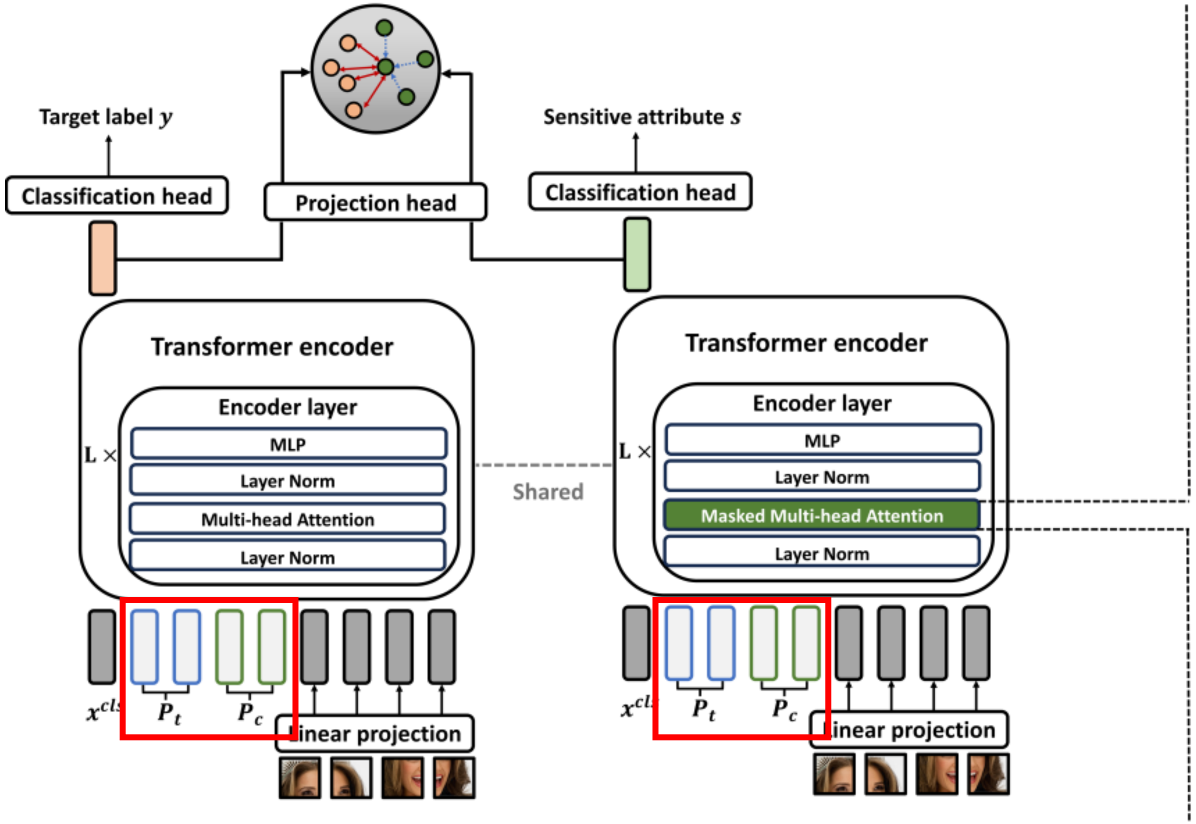
- Encoded patches (outputs) $z_l = \underbrace{T_l}_{\text{Transformer layers}}(z_{l-1}), l = 1, 2, ...L$

- Prediction $y' = \overbrace{C}^{\text{Classifier}}\underbrace{(x_L^{cls})}_{\text{Encoded class token}}$

Method

1. Categorizing prompts into "Target prompts" and "Cleaner Prompts"

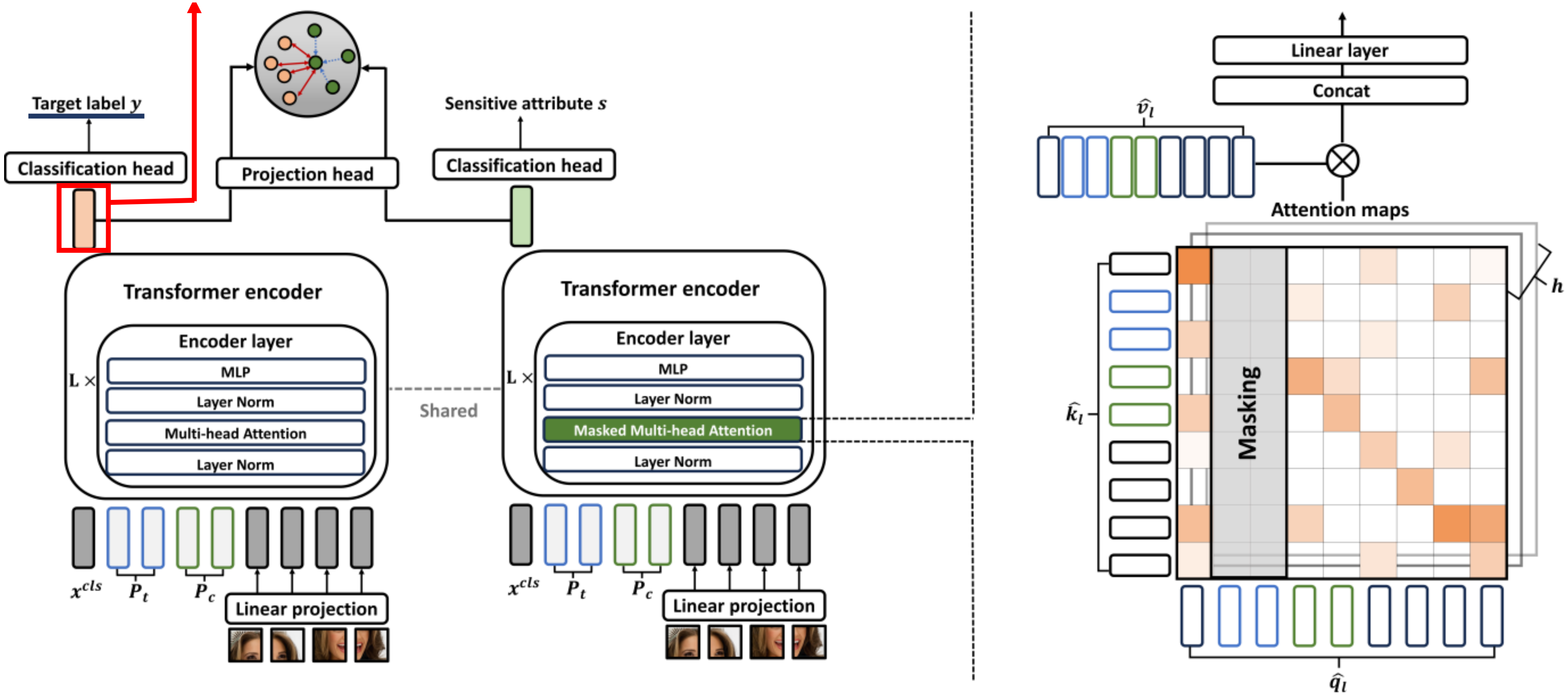
- $\hat{z}_0 = [x_{cls}, \underbrace{P_t^{(1)}, ..., P_t^{(\alpha)}}_{\text{Target Prompts}}, \underbrace{P_c^{(1)}, ..., P_c^{(M-\alpha)}}_{\text{Cleaner Prompts}}, E(x_p)]$



Method

2. Encoding prompts in different manners (*i.e.,* Standard MSA and Masked MSA)

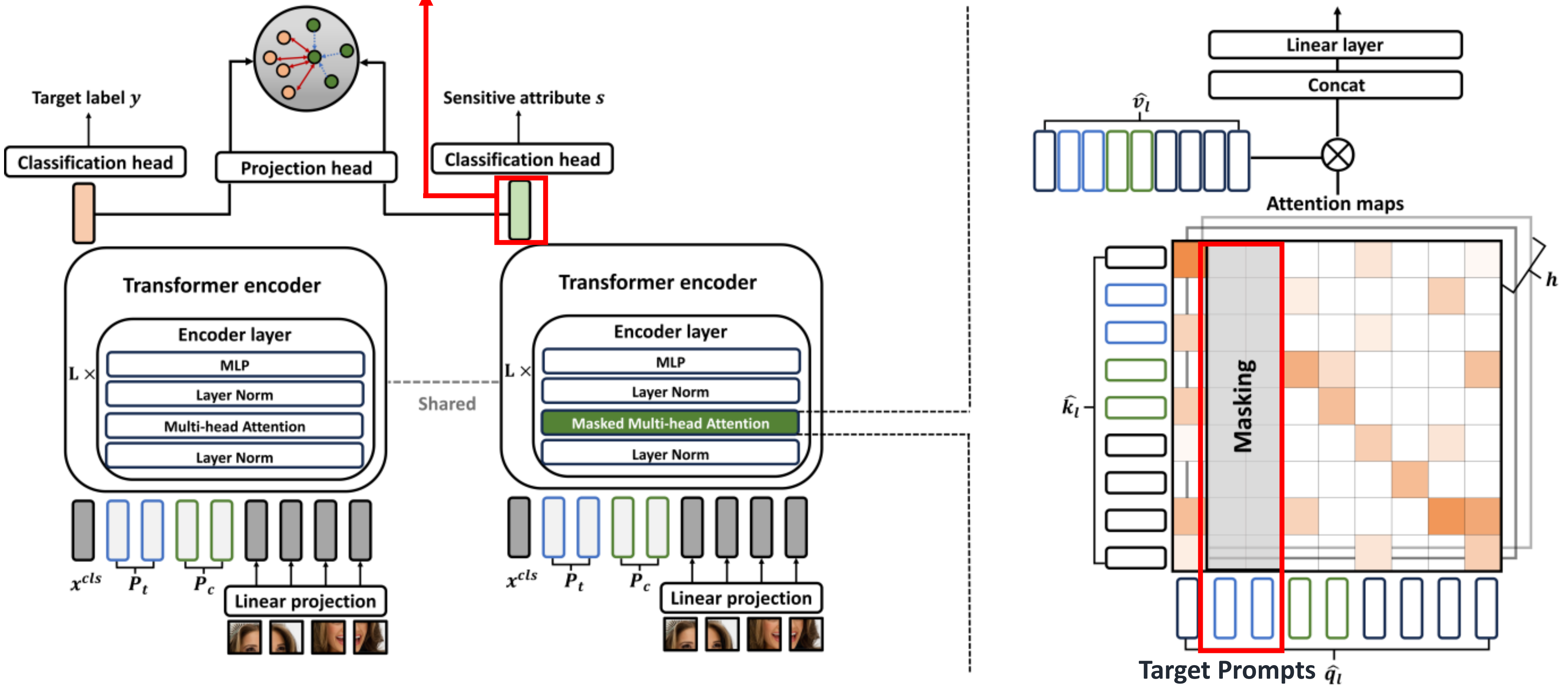
- Standard self-attention $SA(\hat{z}_l) = softmax(\frac{\hat{q}_l \hat{k}_l^T}{\sqrt{d_k}})\hat{v}_l$
- $\hat{z}_l = \hat{T}_l(\hat{z}_{l-1}), l = 1, ...L$
- $L_{cls} = l(\boxed{\hat{C}(\hat{z}_L^{(0)})}, y) + l(\tilde{C}(\hat{z}_L^{*(0)}), s)$



Method

2. Encoding prompts in different manners (*i.e.,* Standard MSA and Masked MSA)

- Masked self-attention $SA^*(\hat{z}_l) = softmax(\frac{\hat{q}_l \hat{k}_l^T + Mask}{\sqrt{d_k}})\hat{v}_l$, $Mask_{i,j} = \begin{cases} -inf & \text{if} \quad 1 \leq j \leq \alpha \\ 0 & \text{else} \end{cases}$
- $\hat{z^*}_l = \hat{T}_l^*(\hat{z^*}_{l-1}), l = 1, \ldots L$
- $L_{cls} = l(\hat{C}(\hat{z}_L^{(0)}), y) + l(\boxed{\tilde{C}(\hat{z}_L^{*(0)})}, s)$

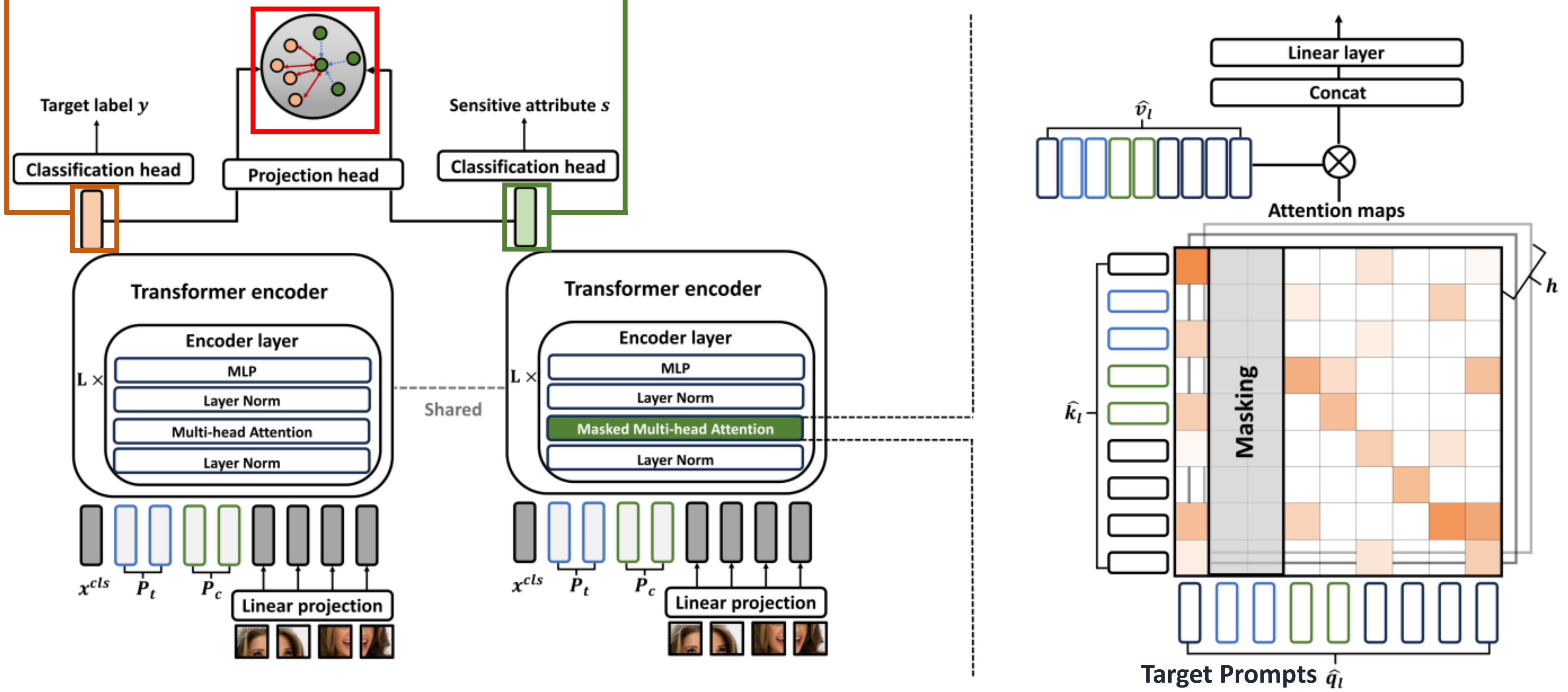


Method

3. Disentangling the encoded class tokens based on contrastive learning

$$- L^{dis} = - \sum_{\forall r^*(i)} \frac{1}{|P(i)|} \sum_{r^*(j) \in P(i)} \log \frac{exp(r^*(j) \cdot r^*(i)/\tau)}{\sum_{r(k) \in N(i)} exp(r(k) \cdot r^*(i)/\tau)}$$

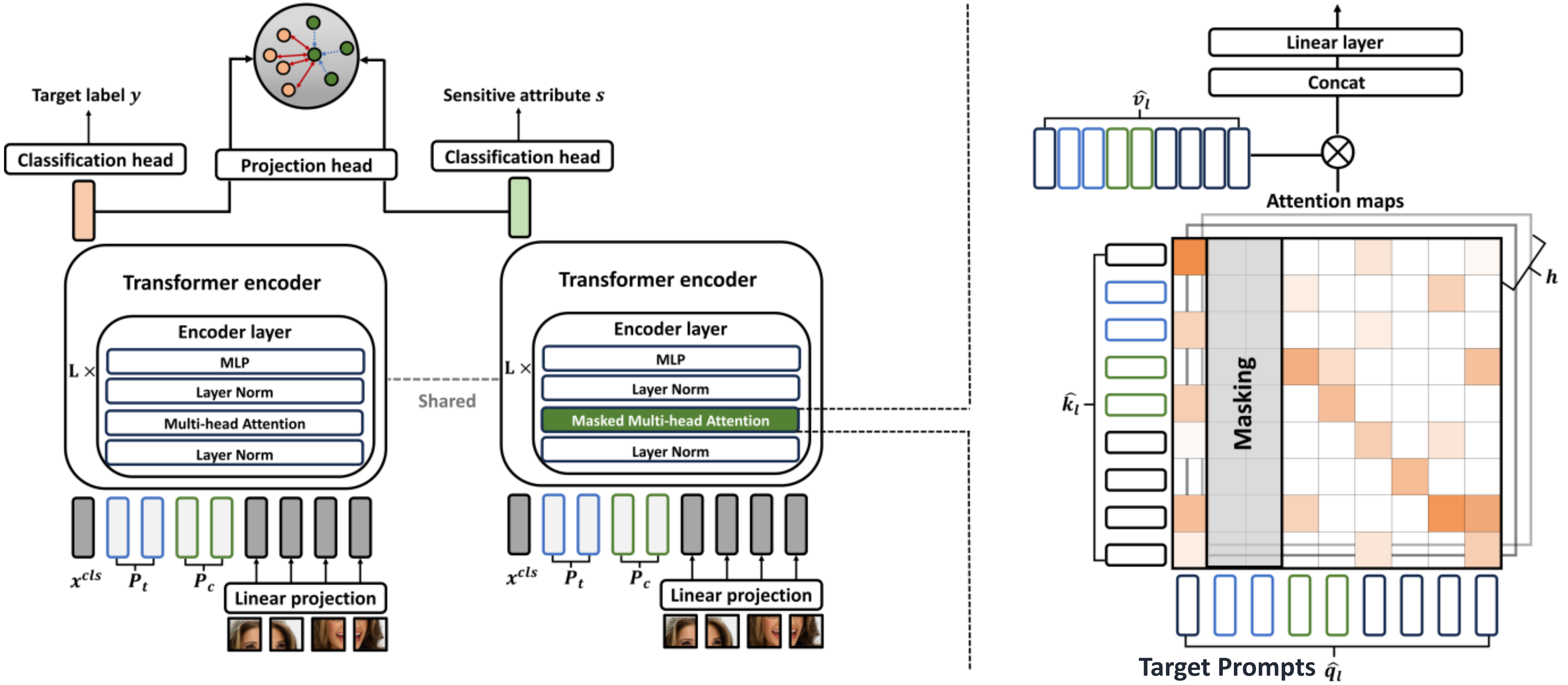
$$P(i) = \{\boxed{r^*(j)} | y(j) = y(i), s(j) = s(i)\} \qquad N(i) = \{\boxed{r(k)} | y(k) = y(i), s(k) = s(i)\}$$



Method

4. Training Downstream Classifier

- Utilizing masked self-attention for the cleaner prompts $\overline{Mask}_{i,j} = \begin{cases} -inf & \text{if} \quad \alpha < j \leq M \\ 0 & \text{else} \end{cases}$
- Excluding the sensitive attribute information in downstream classification



Experiment

Classification results on CelebA dataset

Sensitive attribute: gender

Method	Target label	Sensitive attribute s=0	Sensitive attribute s=1	Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
ViT [14]	t=0	69.1	96.4	78.4	68.7	41.6
	t=1	82.7	26.8			
VPT [24]	t=0	65.2	93.1	**81.7**	75.0	32.1
	t=1	89.1	52.7			
VPT [24]+AT [45]	t=0	38.9	63.7	67.6	63.2	24.0
	t=1	86.9	63.5			
VPT [24]+FSCL+ [42]	t=0	30.7	60.1	69.3	66.5	20.6
	t=1	93.6	81.8			
Fair-VPT (Ours)	t=0	73.8	85.8	78.6	**76.3**	**12.0**
	t=1	78.8	66.7			

Classification results for "Attractiveness"

Method	Target label	Sensitive attribute s=0	Sensitive attribute s=1	Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
ViT [14]	t=0	98.1	79.1	81.7	61.3	30.6
	t=1	12.8	55.1			
VPT [24]	t=0	98.3	81.6	82.7	62.8	28.5
	t=1	15.4	55.8			
VPT [24]+AT [45]	t=0	99.4	86.3	81.2	57.3	23.7
	t=1	4.5	38.8			
VPT [24]+FSCL+ [42]	t=0	99.3	89.9	**84.6**	63.6	25.1
	t=1	12.2	53.2			
Fair-VPT (Ours)	t=0	92.7	79.1	79.9	**65.4**	**15.9**
	t=1	35.6	53.9			

Classification results for "Big Nose"

Experiment

Classification results on CelebA dataset

Sensitive attribute: gender

Method	Target label	Sensitive attribute s=0	Sensitive attribute s=1	Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
ViT [14]	t=0	69.1	96.4	78.4	68.7	41.6
	t=1	82.7	26.8			
VPT [24]	t=0	65.2	93.1	**81.7**	75.0	32.1
	t=1	89.1	52.7			
VPT [24]+AT [45]	t=0	38.9	63.7	67.6	63.2	24.0
	t=1	86.9	63.5			
VPT [24]+FSCL+ [42]	t=0	30.7	60.1	69.3	66.5	20.6
	t=1	93.6	81.8			
Fair-VPT (Ours)	t=0	73.8	85.8	78.6	**76.3**	**12.0**
	t=1	78.8	66.7			

Classification results for “Attractiveness”

Method	Target label	Sensitive attribute s=0	Sensitive attribute s=1	Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
ViT [14]	t=0	98.1	79.1	81.7	61.3	30.6
	t=1	12.8	55.1			
VPT [24]	t=0	98.3	81.6	82.7	62.8	28.5
	t=1	15.4	55.8			
VPT [24]+AT [45]	t=0	99.4	86.3	81.2	57.3	23.7
	t=1	4.5	38.8			
VPT [24]+FSCL+ [42]	t=0	99.3	89.9	**84.6**	63.6	25.1
	t=1	12.2	53.2			
Fair-VPT (Ours)	t=0	92.7	79.1	79.9	**65.4**	**15.9**
	t=1	35.6	53.9			

Classification results for “Big Nose”

Experiment

Classification results on UTK Face, bFFHQ, and Waterbirds

Sensitive attribute: gender / background

Method	TL	SA s=0	SA s=1	BAcc. (↑)	EO (↓)
ViT [14]	t=0	96.0	80.3	88.4	13.4
	t=1	83.1	94.4		
VPT [24]	t=0	95.3	82.3	89.0	12.6
	t=1	83.6	94.9		
VPT [24]+AT [45]	t=0	95.5	81.5	88.9	11.6
	t=1	84.8	94.1		
VPT [24]+FSCL+ [42]	t=0	96.1	85.8	89.0	9.9
	t=1	82.3	91.9		
Fair-VPT (Ours)	t=0	95.1	89.3	**90.9**	**4.9**
	t=1	87.5	91.6		

Classification results on UTK Face

Method	TL	SA s=0	SA s=1	BAcc. (↑)	EO (↓)
ViT [14]	t=0	99.1	54.3	74.8	48.9
	t=1	46.3	99.5		
VPT [24]	t=0	98.9	48.3	76.0	46.3
	t=1	57.5	99.5		
VPT [24]+AT [45]	t=0	99.5	58.7	77.5	43.1
	t=1	53.1	98.7		
Fair-VPT (Ours)	t=0	99.1	62.3	**80.7**	**37.1**
	t=1	61.9	99.5		

Classification results on bFFHQ

Method	TL	SA s=0	SA s=1	Acc.	BAcc.	EO
ViT [14]	t=0	99.7	77.8	85.1	80.5	31.3
	t=1	52.0	92.6			
VPT [24]	t=0	99.6	82.9	**86.8**	81.2	29.2
	t=1	50.3	92.0			
VPT +AT [45]	t=0	98.7	81.3	86.3	81.6	27.0
	t=1	54.8	91.5			
Fair-VPT (Ours)	t=0	93.9	70.9	83.3	**84.3**	**18.7**
	t=1	78.9	93.6			

Classification results on Waterbirds

Experiment

Ablation study

Demonstrating the effectiveness of each proposed component

Categorized Prompts	L_{cls}			L_{dis}	CelebA			UTK Face	
	$\hat{z}_L^{(0)}$	$\hat{z}_L^{*(0)}$	$\overline{z}_L^{(0)}$		Acc. (↑)	BAcc. (↑)	EO (↓)	BAcc. (↑)	EO (↓)
	✓				**81.7**	75.0	32.1	89.0	12.6
✓	✓	✓			77.9	75.9	15.0	89.4	8.1
✓	✓	✓		✓	78.6	**76.3**	**12.0**	**90.9**	**4.9**
✓		✓		✓	78.0	74.0	25.2	88.0	10.9
✓		✓	✓		77.3	72.9	29.1	89.2	12.2
✓		✓	✓	✓	78.4	73.9	24.4	89.6	9.4

Conclusion

- We demonstrated that there exists two key factors causing unfairness in supervised contrastive loss (SupCon)
- To suppress them, we proposed Fair Supervised Contrastive Loss (FSCL) and Group-wise Normalization
- Our method achieves the best trade-off performances on benchmark datasets and works efficiently in various challenging environments