Accuracy-compatible Fairness Computing in Multimedia

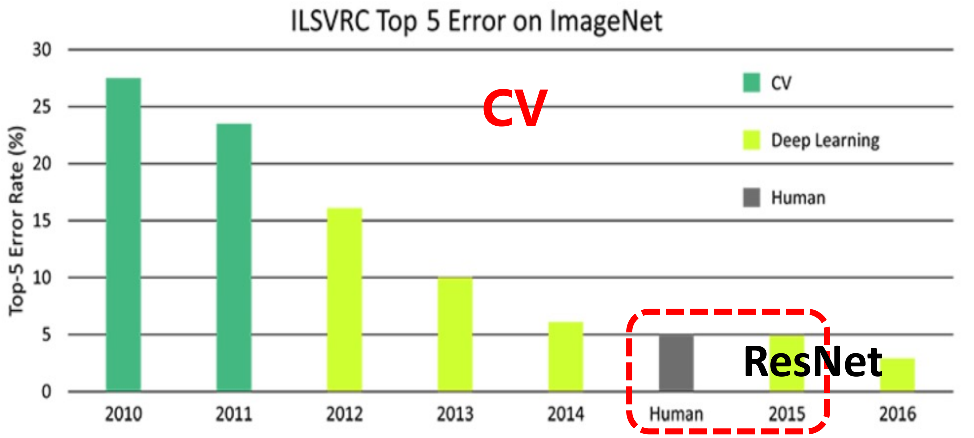
Yi Zhang

Co-Contributors: Jitao Sang, Kexin Wang, Yuxuan Yi

Beijing Jiaotong University

yi.zhang@bjtu.edu.cn

Progress for AI and challenge to fairness

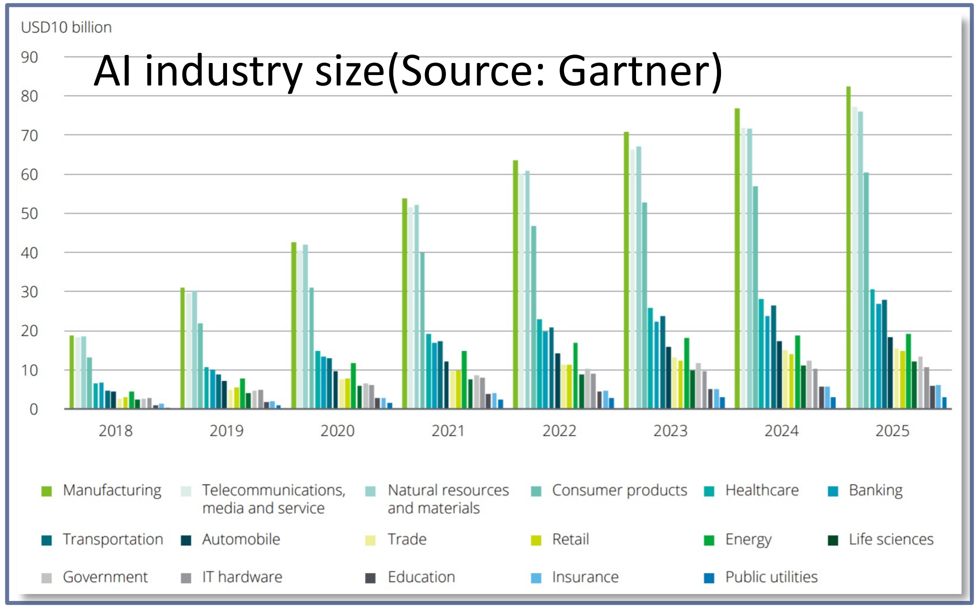


source: https://www.dsiac.org/resources/journals/dsiac/winter-2017-volume-4-number-1/real-time-situ-intelligent-video-analytics

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google A.I.	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) Google A.I.	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
2	nlnet (ensemble)	85.954	91.677

NLP

AI has achieved great success in CV and NLP



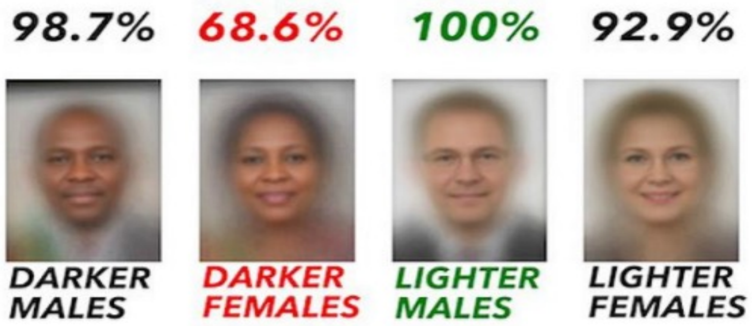
Challenge to fairness

What is fairness problem?

Accuracy is not the only one criterion, AI needs criteria in fairness

- The same algorithm produces result of discrimination for sensitive people in decision-making:
  - **(1)**Different decision-making accuracy for different groups of people;
  - **(2)**In decision-making, algorithm uses population information that are not related to the task.

Dark face cannot be detected by the face detection model



Accuracy gap between darker females and lighter males

Judicial algorithm predicts that blacks have a higher recidivism rate

Related work and challenges

To mitigate bias associated with protected attributes, many methods have been proposed.

Related work

1. Data manipulation
2. Adversarial training
3. Post processing

Challenges

difficult to debiasing, accuracy cost

Our goals

Fairness V.S. Accuracy

How to better debiasing and keep accuracy of model?

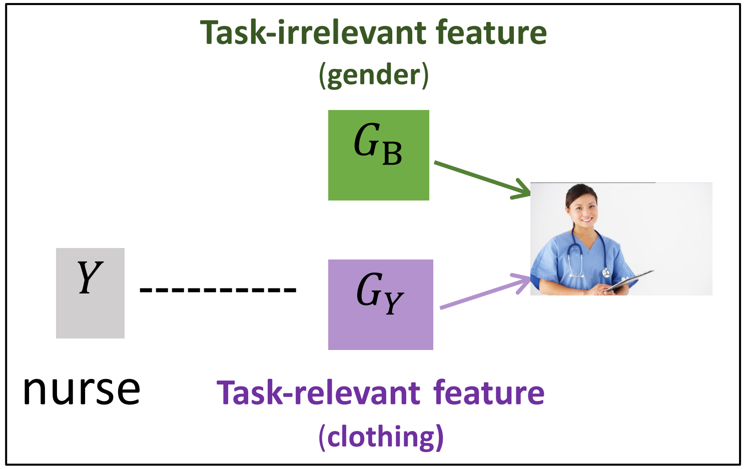
Source of bias: spurious correlations

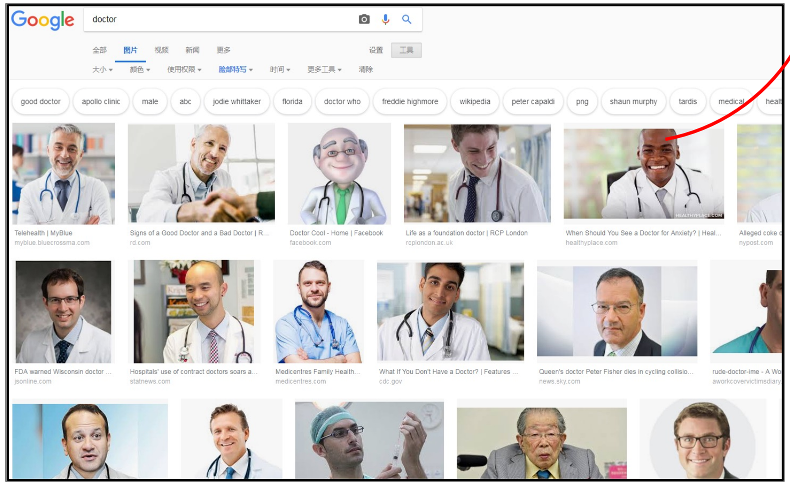
The algorithm learns spurious correlations from training data :

- ☐ The data samples contain task-irrelevant features（In fairness issues, gender）
- ☐ The task-irrelevant features are strongly related to task-label in the training set
- ☐ The model uses spurious correlations to reduce training loss and turns out model bias

Running example:

Doctor/nurse
Occupation recognition

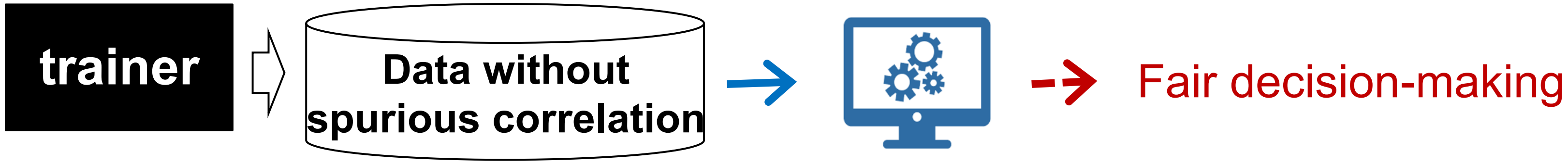




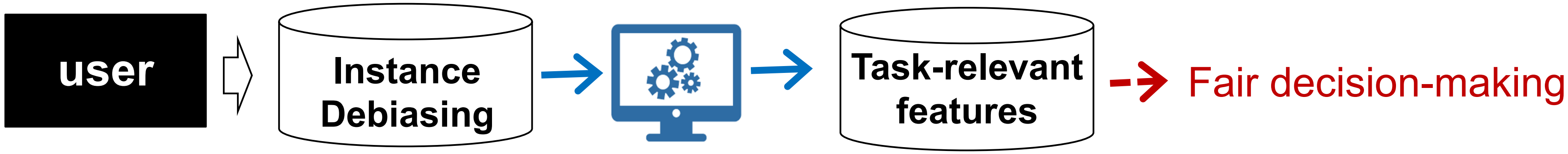
Retrieved “doctor” images from Google

Eliminate spurious correlations

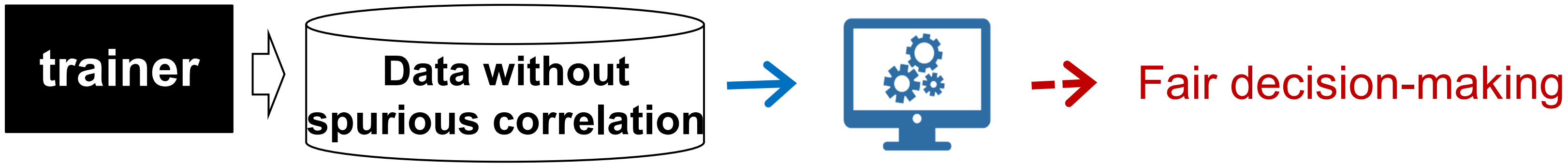
1 Train a fair model
A trainer with the target of fairness to only use task-relevant features, which makes the learned model fair



2 Get fair results based on a unfair model (more easy-to-use)
A user with the target of fairness to remove the task-irrelevant features of test samples and keep task-irrelevant features, which obtains fair output on unfair model



Part1 : Train a fair model



Adversarial Example-based Data Augmentation for Eliminating Spurious Correlations in Dataset

Yi Zhang, JitaoSang: Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. *ACM Multimedia 2020*.

Preliminary

Target variable： Task label to be predicted (occupation: doctor/nurse)

Bias variable： The task-irrelevant attributes that may affect the prediction result (social attributes: gender/skin color)

Female nurse ✓

Male nurse ✗

$$bias(\theta, t) = |P(\hat{t} = t | b = 0, t^* = t) - P(\hat{t} = t | b = 1, t^* = t)|$$

Model Bias: **True positive rate (TPR) gap between groups with different Bias variable for each Target variable**

The larger the model bias value, the more unfair the algorithm

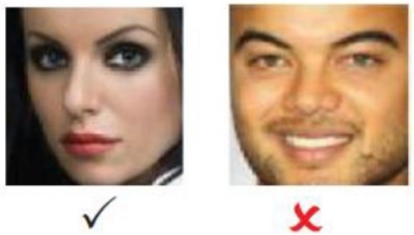
Dataset imbalance V.S. Model bias

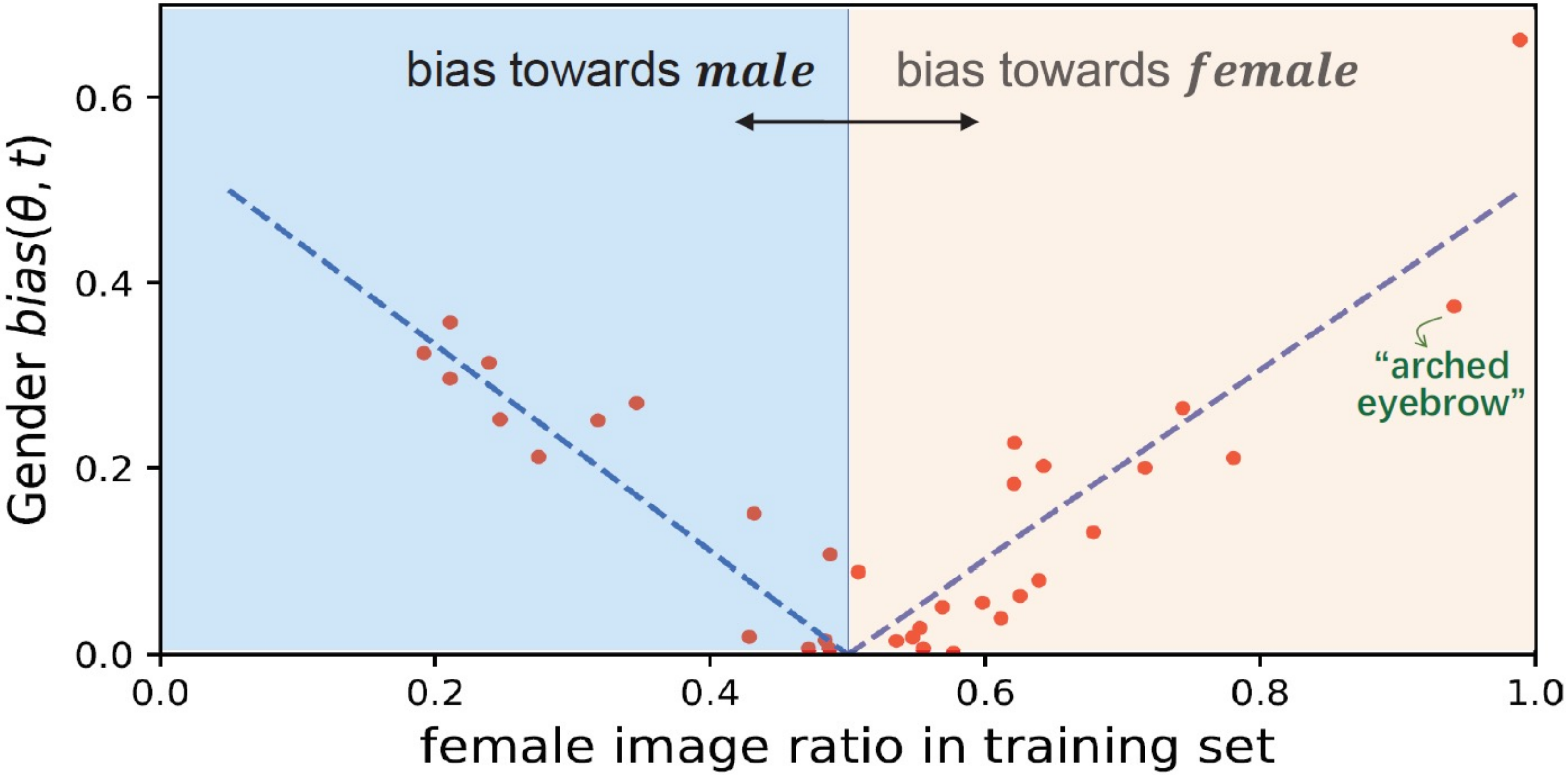
CelebA Dataset

Target variable： facial attributes

Bias variable： gender

"Arched eyebrows" Recognition





Goal: To balance the data distribution

<u>Up-sampling</u>: assigning different sampling rate to samples, **only guarantees superficial data balance** and can not completely balance

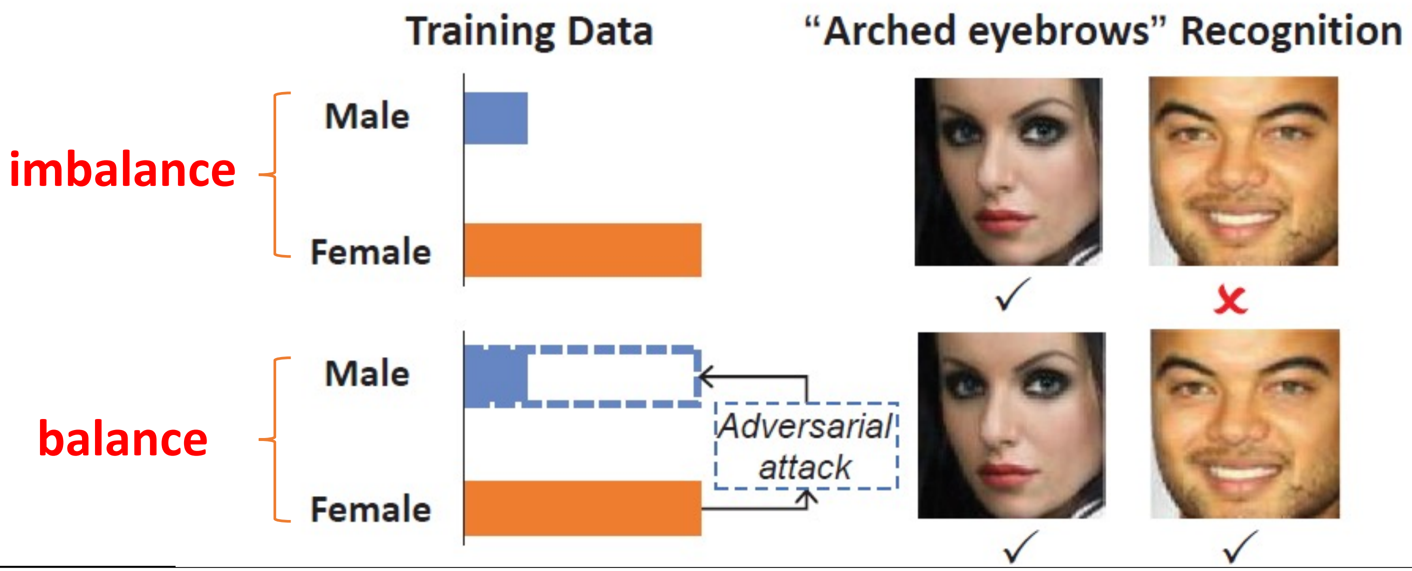
<u>Down-sampling</u>: discarding samples with majority bias variable, **fails to make full use of the training data**

Fairness V.S. Accuracy

Two goals:

1 For fairness, <u>superficial data balance →completely balance</u>

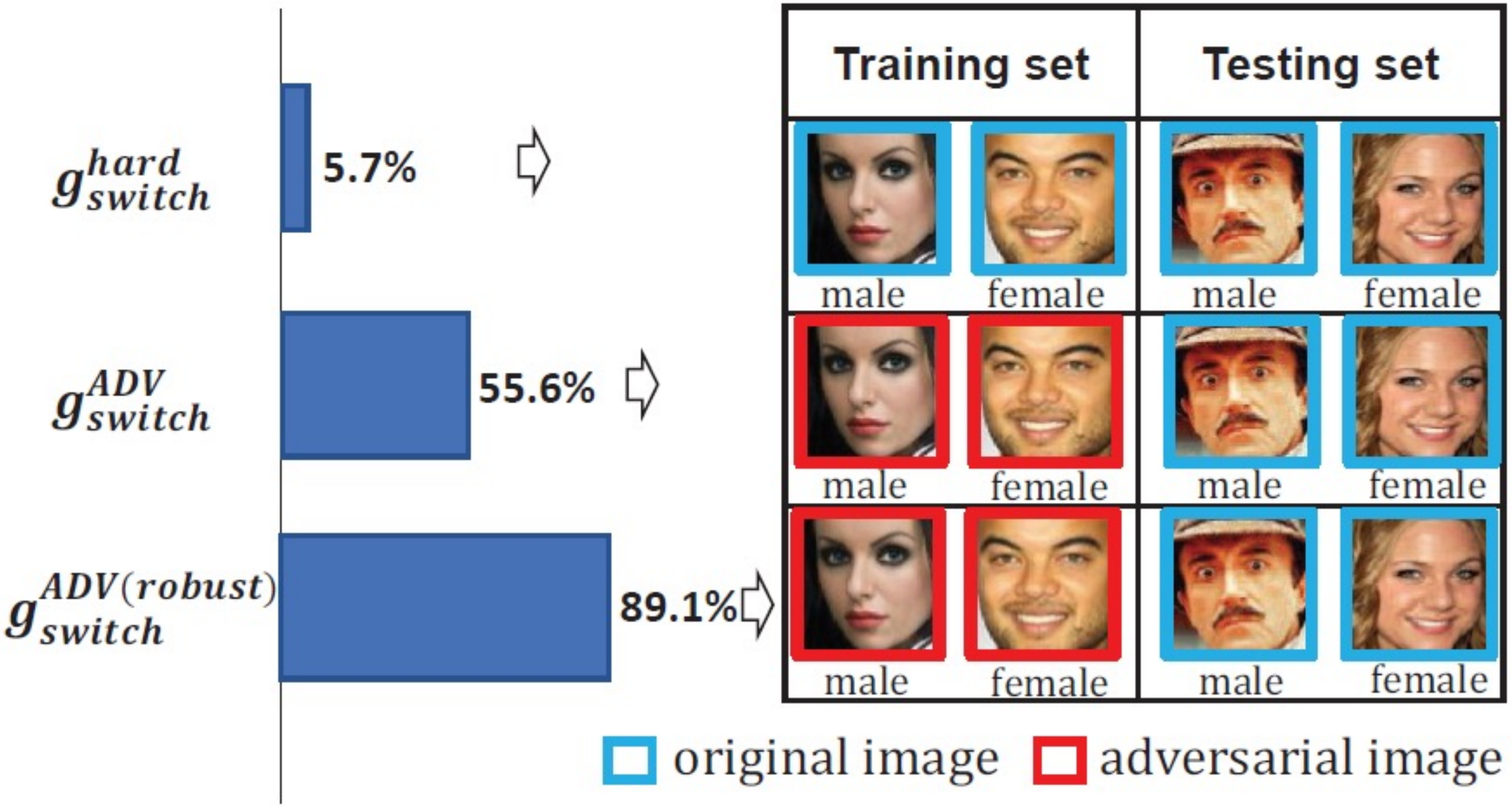
2 For accuracy, <u>fails to make full use of the training data→make full use</u>



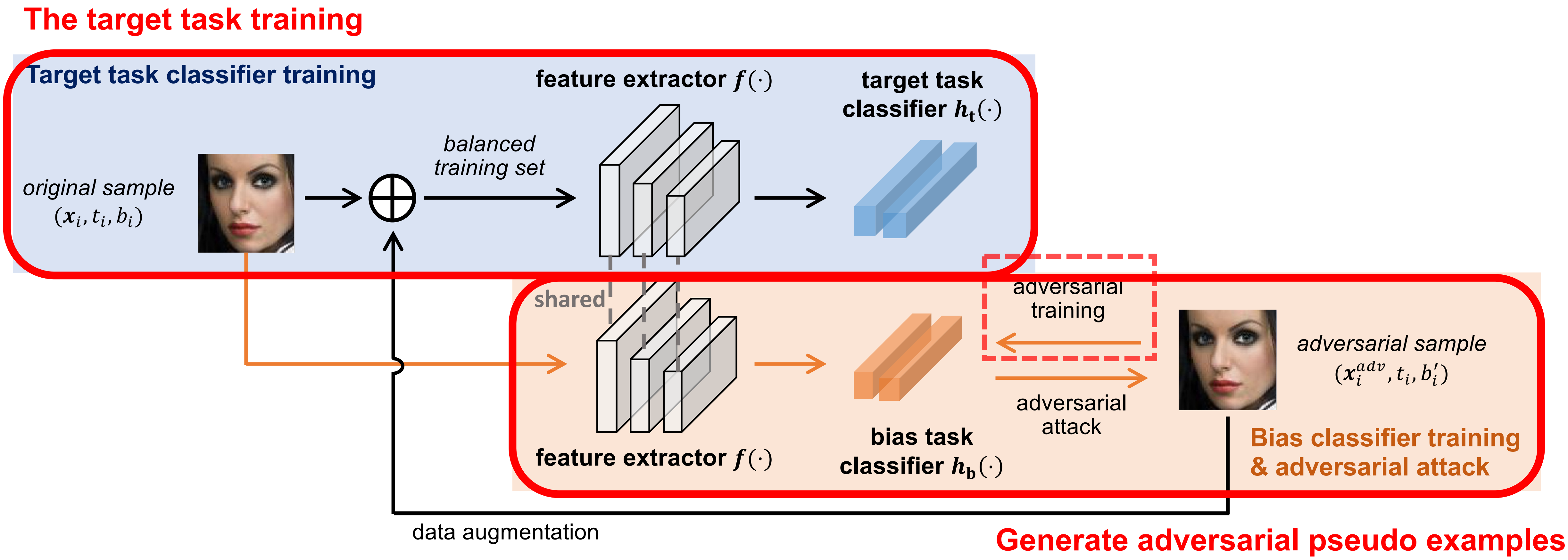
generate pseudo samples

The Potential of Adversarial Example in Balancing

- Adversarial examples **contain useful information about the attack target class** and have potential to **generalize** to original real data



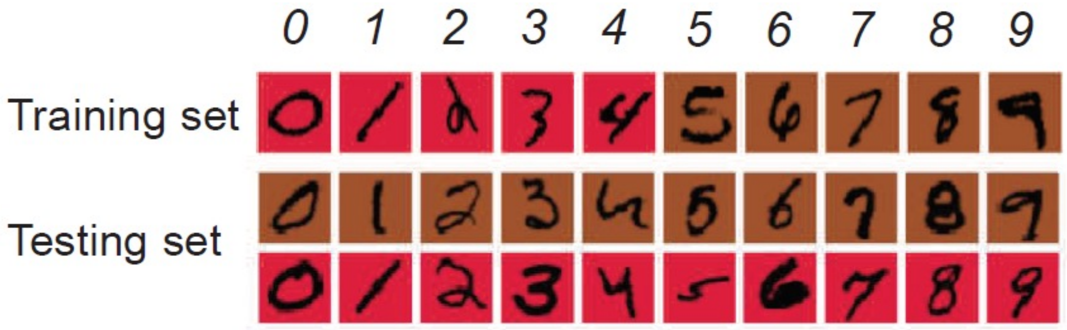
Method



The target task training
Target task classifier training
feature extractor $f(\cdot)$
target task classifier $h_t(\cdot)$
balanced training set
original sample (x_i, t_i, b_i)
shared
adversarial training
adversarial sample (x_i^{adv}, t_i, b_i')
adversarial attack
bias task classifier $h_b(\cdot)$
feature extractor $f(\cdot)$
Bias classifier training & adversarial attack
data augmentation
Generate adversarial pseudo examples

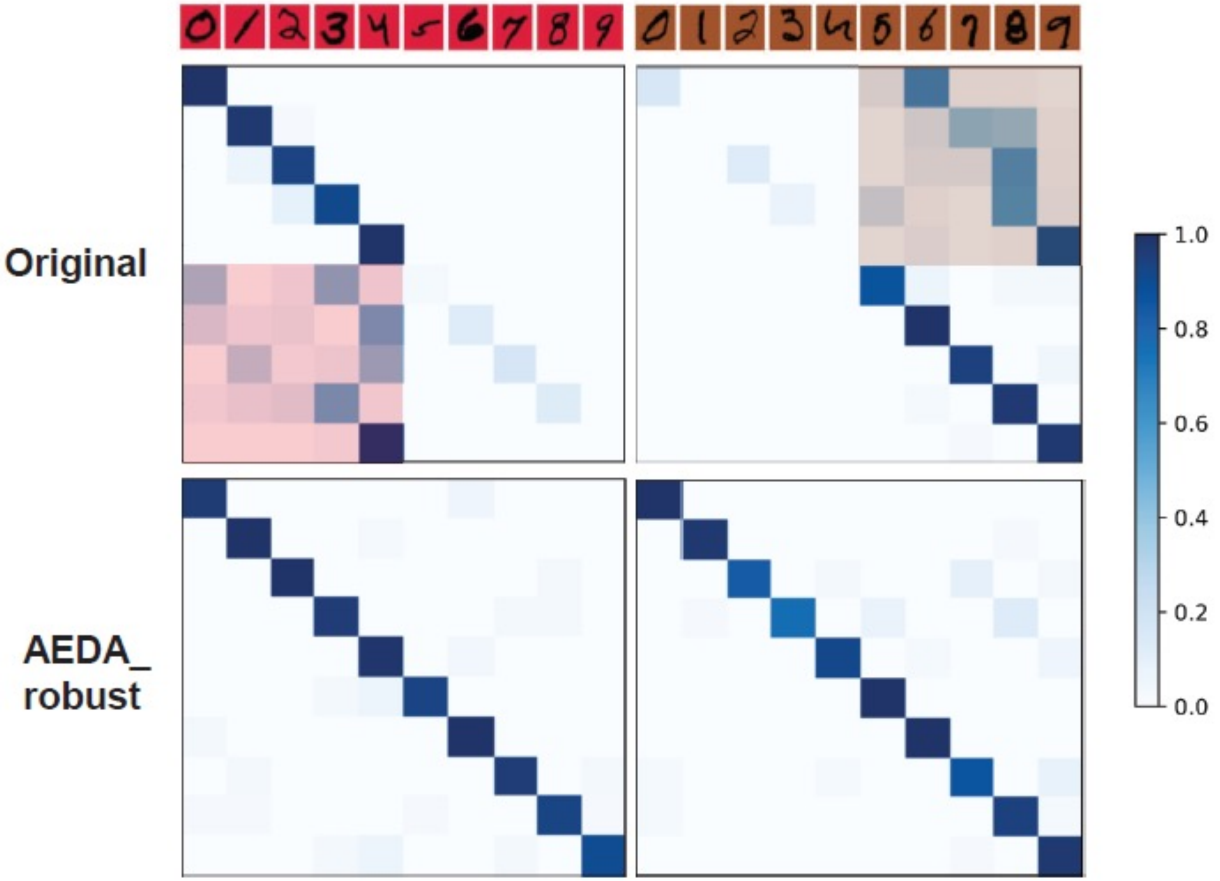
Simulated Debiasing Evaluation

- ***target var***: **digit//** ***bias var***: **background color**



Methods	bACC (%)	Model bias
Original	55.62	7.84
Under-sampling [7, 30]	–	–
Reweighting [15]	–	–
Adv debiasing [4, 11, 26]	89.93	1.37
CycleGAN [31]	65.23	5.42
AEDA_pre	64.53	5.89
AEDA_online	80.57	3.20
AEDA_robust	**91.80**	**0.53**

Accuracy & Model bias in C-MNIST



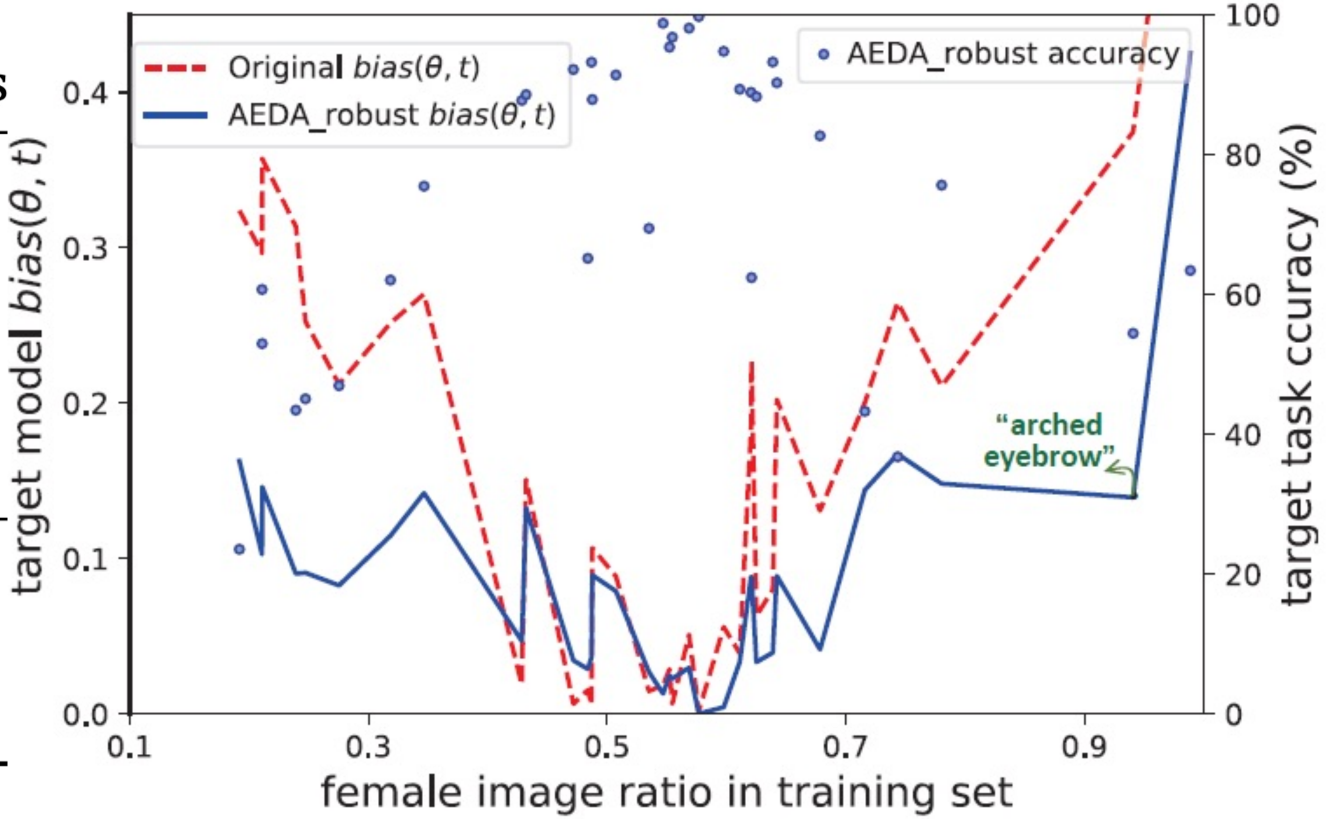
Confusion matrix for testing subset of ⟨0 ~ 9, red⟩ (left) and ⟨0 ~ 9, brown⟩ (right)

Real-world Debiasing Evaluation

- *target var*: facial attributes // *bias var*: gender

Methods	bACC (%)	Model bias
Original	73.57	5.48
Under-sampling [7, 30]	66.35	**2.35**
Reweighting [15]	73.82	4.39
Adv debiasing [4, 11, 26]	72.82	4.23
CycleGAN [31]	73.65	4.75
AEDA_pre	73.68	5.23
AEDA_online	74.03	4.22
AEDA_robust	**74.30**	3.27

Accuracy & Model bias (gender)



Accuracy & Model bias (gender) in different facial attributes

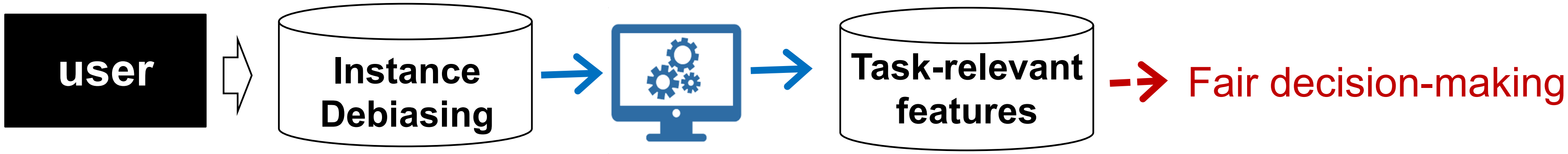
AEDA→ Accuracy-compatible Fairness

Discussion

- ***Accuracy-fairness paradox?***
- It is recognized in conventional debiasing attempts that there exists the tradeoff between accuracy and fairness, and the goal is to reduce model bias under the slightly decreased accuracy.

- ***Accuracy-compatible Fairness***
- Previous slides: *close relation between distribution balance, model bias and accuracy*
- Without discarding training data or adding constraints to affect target task learning, the data augmentation provides alternative perspective to simultaneously improve fairness and accuracy.

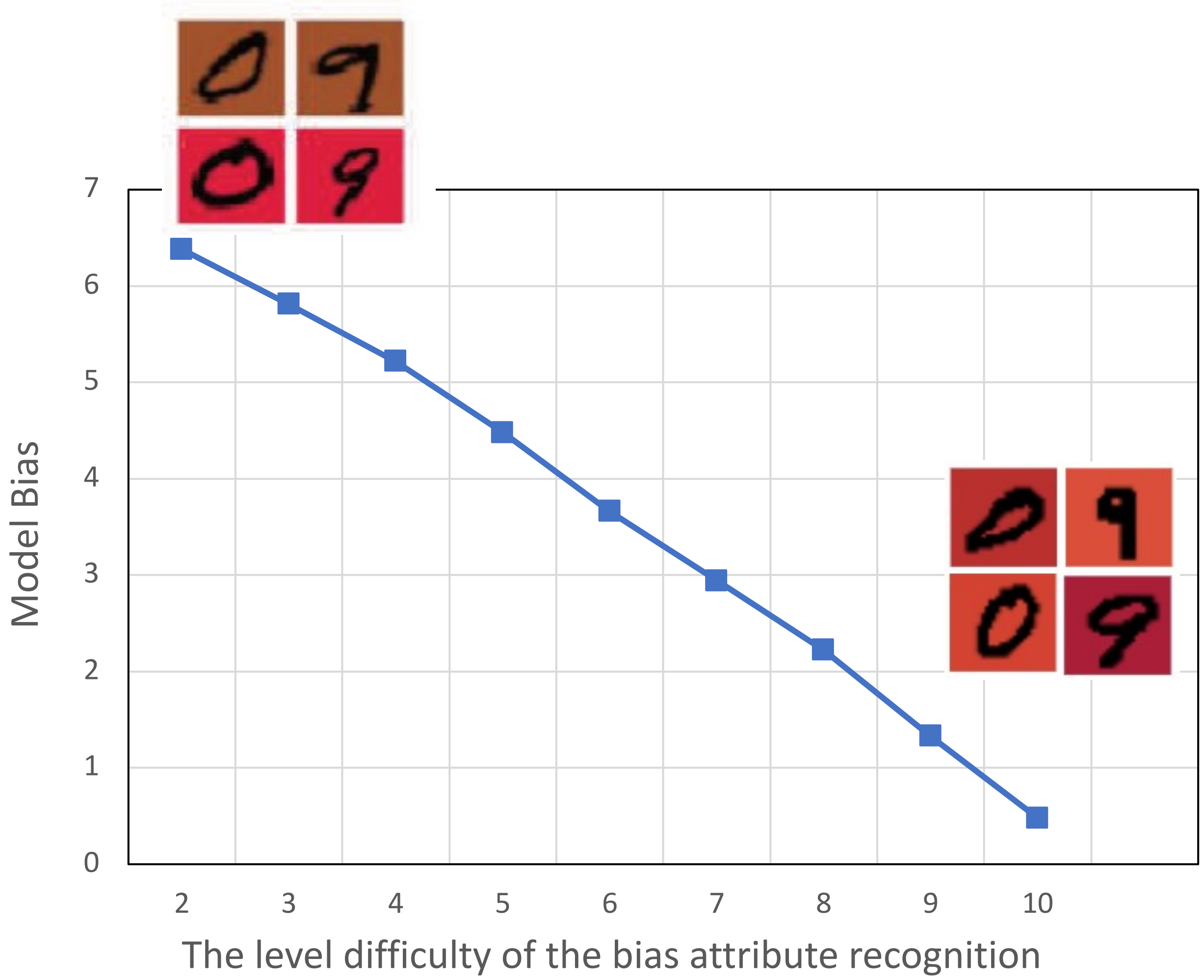
Part2: Get fair result based on unfair model



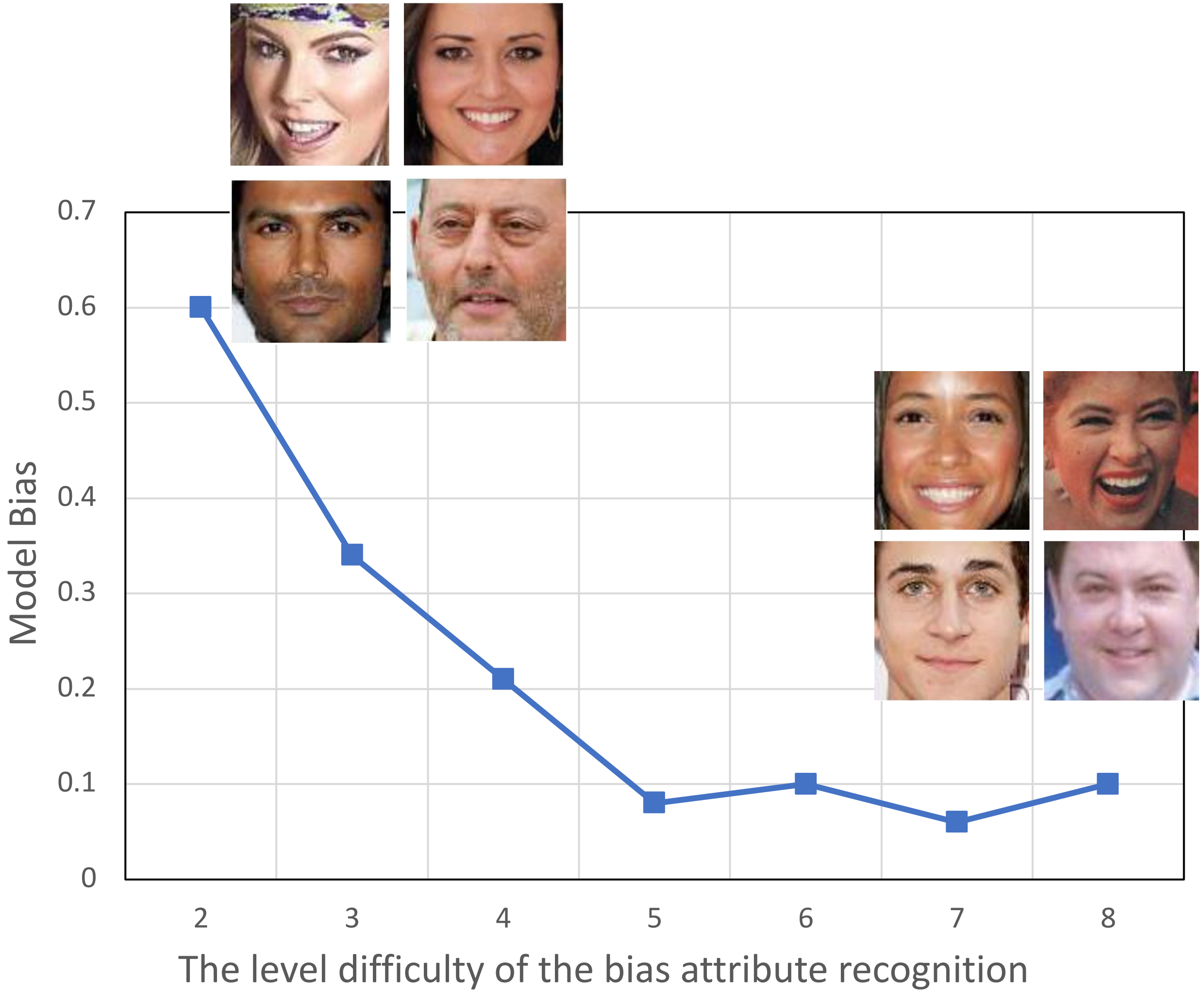
Post-fairness: Cross-task Adversarial Attack-based Instance Debiasing for Deactivating Spurious Correlations

https://adam-bjtu.org/paper/ZhangYi/Post_fairness_yi/Post_fairness.html

Not all results are based on task-irrelevant features



C-MNIST



CelebA *"arched eyebrows"* recognition

Regular adversarial attack can't cross task

- *target var*: facial attributes // *bias var*: gender

Model Bias	VGG16	VGG19	Resnet18	Restnet32	Resnet50
Original Bias	5.48	5.43	4.93	4.25	5.09
Bias(Instance Debiasing)	4.13	4.26	3.79	3.12	3.47
Bias(Natural Unbiased Instance)	1.98	2.13	1.80	1.96	1.41
$\Delta Bias_{ID}$	*1.35*	*1.17*	*1.14*	*1.13*	*1.62*
$\Delta Bias_{NUI}$	**3.50**	**3.30**	**3.13**	**2.29**	**3.68**

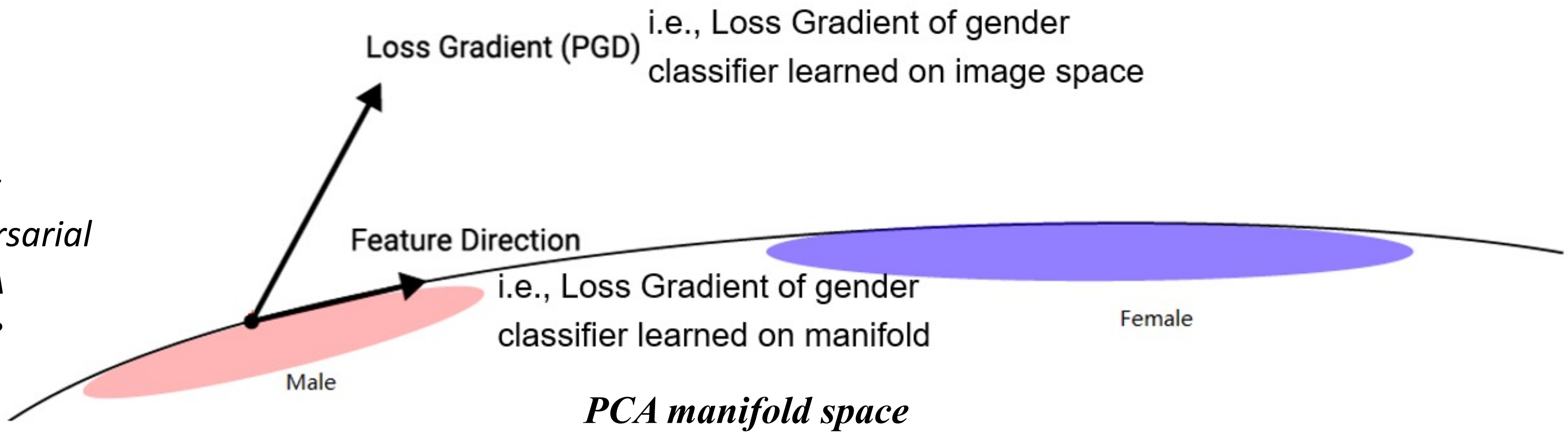
- Instance Debiasing: using pre-trained gender classifier to remove bias information by adversarial attack.
- Natural Unbiased Instance: searching samples that are difficult for gender classification from the original test set.

Why does the adversarial attack underperform?

The direction of adversarial attack isn't feature direction

- Train a gender classifier ***M*** in PCA manifold space.
- Train a gender classifier ***I*** in image space.
- In PCA manifold space, the mean Cosine Similarity between "Loss Gradient of ***I*** " and "Loss Gradient of ***M***" is **0.36**

Analysis Tool :
Encoding adversarial attack into PCA manifold space



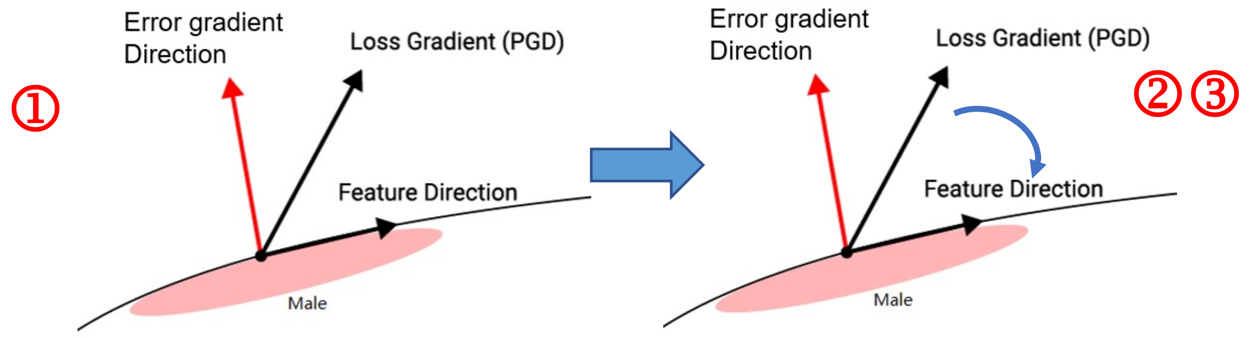
PCA manifold space

The direction of adversarial attack isn't feature direction

Control the gradient of bias classifier

How to control the gradient of gender classifier to be close to the feature direction?

1. Find the error gradient direction that is not in the feature direction.
2. Only use error gradient direction to generate augmented sample.
3. Use the original sample and the augmented sample to train the bias classifier, and the bias classifier can understand corresponding features of error gradient direction are not the bias features.



Now, the direction of adversarial attack is feature direction

Debiasing Evaluation

Comparison with Post-processing debiasing methods

Dataset	Metric	Vanilla	Ours		Other post-processings		
			Ours†	Ours	ROC	EqOdds	CalEqOdds
C-MNIST	Accuracy	55.62	78.90	**82.52**	N/A	N/A	N/A
	Bias	7.84	1.06	**0.74**			
CelebA	Accuracy	73.57	73.88	**74.32**	73.57	68.73	68.63
	Bias	5.48	4.15	**2.39**	3.52	2.63	5.19

Comparison with Pre and In-processing debiasing methods

Dataset	Metric	Vanilla	Pre-processing		In-processing		Ours
			Re-sampling	Down-sampling	Adv debiasing	AEDA	
C-MNIST	Accuracy	55.62	N/A	N/A	89.93	**91.80**	82.52
					89.96	92.32	
	Bias	7.84	N/A	N/A	1.37	**0.53**	0.74
					0.57	0.29	
CelebA	Accuracy	73.57	74.19	64.35	73.52	74.30	**74.32**
			75.36	64.67	74.26	74.57	
	Bias	5.48	3.83	**2.35**	3.75	3.27	2.39
			2.38	1.92	2.83	3.17	

Post-processing as a transitional method

- ***Transformation from basic AI to trustworthy AI***

Long time

- ***AI service provider***
  - There may be sufficient obstacles (unbalanced data, time-consuming repeated testing, unpredictable effect, etc.) to be unable to retrain the model
- ***User***
  - They want to get the fair predictions about themselves
- ***Third-party***
  - They want to get the fair predictions across the population

Post-processing can meet the needs of the three parties, even they can’t retrain the model.

ACM multimedia
Chengdu, China OCT 20-24 2021

BEIJING JIAOTONG UNIVERSITY 1896

Thanks!

Contact me: yi.zhang@bjtu.edu.cn

http://www.adam-bjtu.org