UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

HRI 2021 Virtual

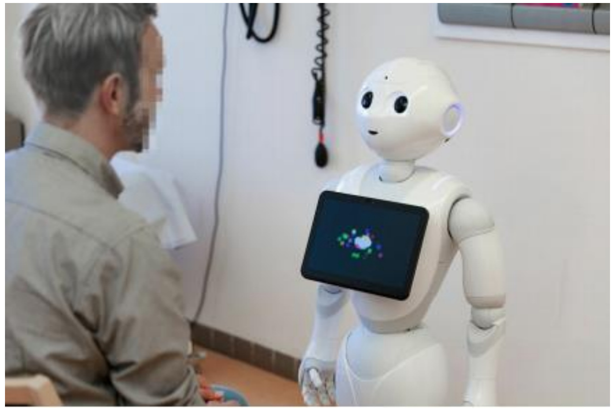
Towards Fair Affective Robotics: Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition



Ozgur Kara   Nikhil Churamani   Hatice Gunes

Affective Robotics & Facial Analysis

- Affective robotics become integral in human life
- Successful long-term HRI can be used in:
  - Providing **physical** and **emotional support** to the users
  - **Healthcare**, **education** and **entertainment**
  - Child-robot interactions
- **Fair** analysis of facial expressions is of vital importance in affective robotics (e.g. design of emotion-aware robots)

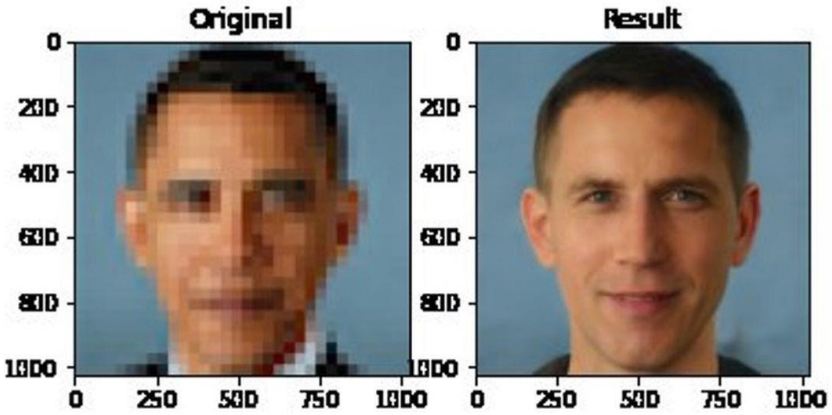


(a)

(b)

(a) Boumans R, van Meulen F, Hindriks K, et al Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial BMJ Quality & Safety 2019;28:793-799.
(b) https://www.wired.com/2010/09/darpa-robot-smarts/

Towards Fairness



(a) A high resolution image is generated from **Barack Obama's** low-resolution image using a generative model

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

(b) **Gender classification** performances of 3 different classifiers

(a) https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias
(b) J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. volume 81 of Proceedings of Machine Learning Research, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR

Popular Bias Mitigation Approaches

Multi-Task Learning

- Adds the biased attribute to the learning objective

Data Augmentation

- Generates additional samples for underrepresented groups.

Adverserial Learning

- Sets up a minimax objective function

Continual Learning ?

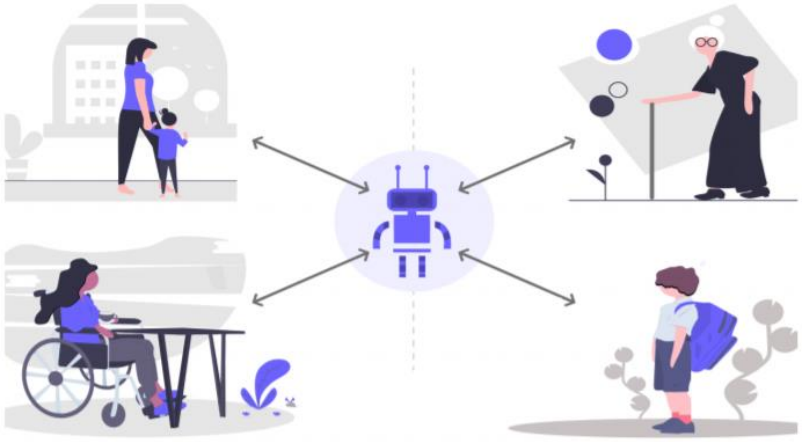
Continual Learning

Continual Learning:

- Can learn with **incrementally acquired** data
- Has the ability **to adapt** with the new data **without forgetting** the seen information

Why Continual Learning:

- Step by step learning manner can allow for **robustness** against biased attributes
- CL can **balance** learning across different domains which leads to development of **fairer models** for affective robots

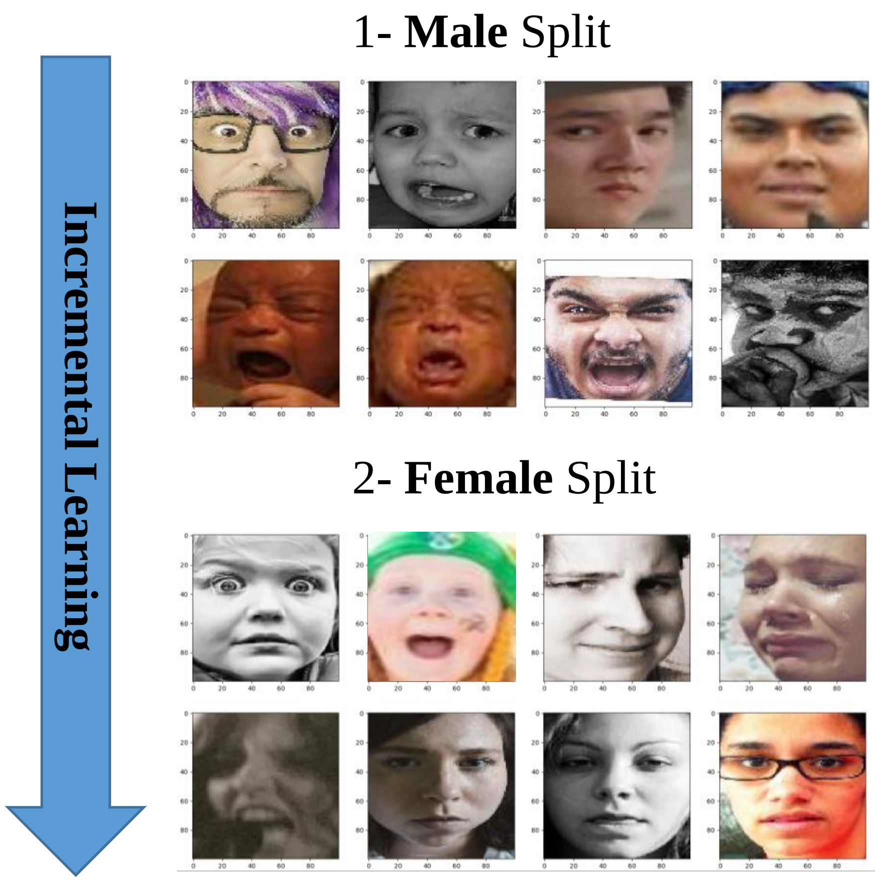


(a)

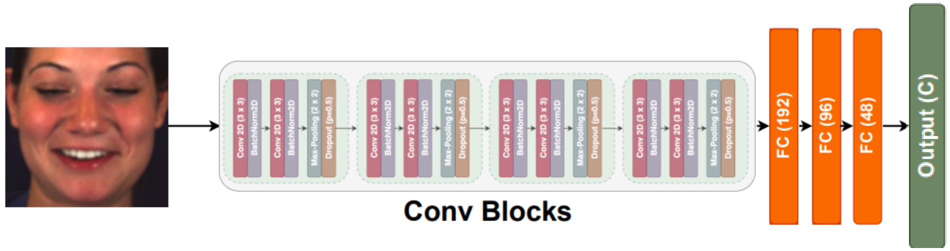
(a) Churamani et al.. Continual Learning for Affective Robotics: Why, What and How?. In IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 425--431, IEEE, Virtual Conference, 2020.

Domain Incremental Learning Settings

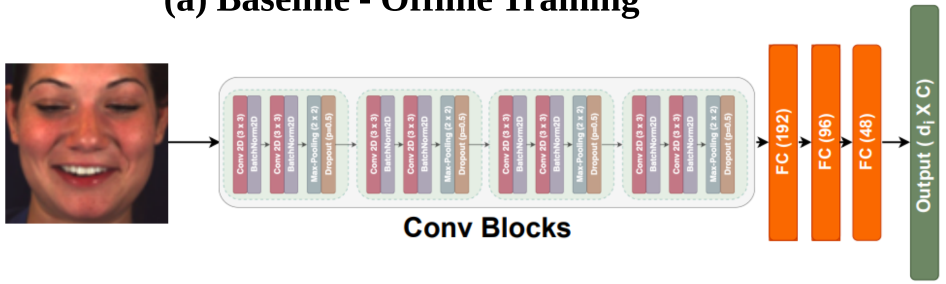
- The task to be learnt by the model does not change but the **input data distribution changes**
- **Continual learning algorithms** are trained under this setting **for each sensitive attribute**
- An example of the settings for gender attribute
  - **Task:** Classifying facial expressions
  - **Attribute:** Gender
  - **Domains:** Male and Female
  - **Splits:** Each split involves samples from one domain
  - **Training:** Model encounters with one split at a time and learns incrementally
  - **Evaluation:** Model is evaluated on each split after training



Benchmark – Non CL Based Approaches

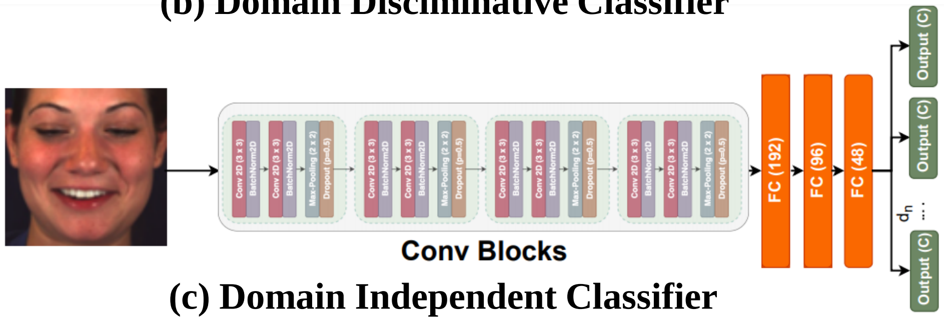


(a) Baseline - Offline Training



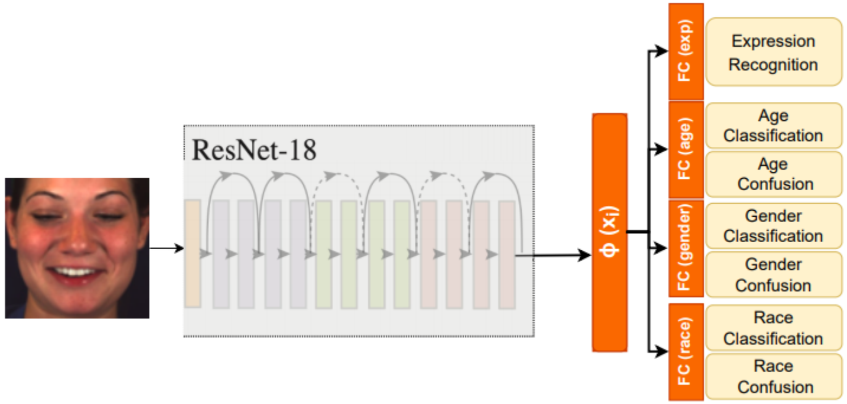
N x M classifier
(N: domain, M: class)

(b) Domain Discriminative Classifier



Multi-head

(c) Domain Independent Classifier



(d) The Disentangled Approach

Weight

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N} w_i \sum_{s=1}^{S} y_s^{(i)} \log \hat{y}_s^{(i)}$$

(e) Strategic Sampling

Benchmark – CL Based Approaches

EWC

- Adds **quadratic penalty** on the difference between the parameters for the old and new tasks

EWC Online

- A single quadratic penalty is applied in an **online fashion**.

SI

- Adds **importance value** to parameters of the network, high important parameters change less

MAS

- MAS enables importance weight estimation in an unsupervised manner

Naive Rehearsal

- While training a new task, each mini-batch is constructed by an **equal amount of new data** and **the rehearsal data**.

- Note that we don't use **complex rehearsal algorithms** for fair comparison

Fairness Measure

- We use '*equal opportunity*' definition of fairness [1]
- It quantifies the largest gap among scores on different domains
- We use **accuracy** as a scoring metric for models

$$\mathcal{F} = \min\left(\frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_0, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}, \ldots, \frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_n, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}\right)$$

x: *input*
ŷ: *predicted label*
f: *scoring function*

y: *ground truth label*
s: *sensitive attribute*
d: *dominant attribute*

Green denotes the minimum accuracy score
Blue denotes the maximum accuracy score
Fairness = **Green** / **Blue** => largest gap

	Black	Asian	White	Latino	Fairness
Baseline	0.659	0.720	0.771	0.764	0.855
Naive Rehearsal	0.767	0.779	0.788	0.762	0.967

Example: Accuracy and fairness table for 2 methods evaluated on race attribute

UNIVERSITY OF CAMBRIDGE

[1] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity insupervised learning," inAdvances in neural information processingsystems, 2016, pp. 3315–3323.

Experiment Setup

- We conducted **2 experiments** with **11 different approaches**:
  - **Experiment 1**: Facial Expression Recognition
  - **Experiment 2**: Action Unit Detection
- With reporting models' fairness performances on:
  - **Gender** attribute,
  - **Race** attribute,
- Under two versions
  - **With** data augmentation (horizontal flipping)
  - **Without** data augmentation.
- We use the same CNN architecture for all models except for the Disentangled Approach [1]

RAF-DB Dataset

- Provides 7 expression labels:
  - Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral
- Provides **gender**:
  - *Male – Female*
- and **race** information:
  - *Caucasian – African American – Asian*

BP4D Dataset

- We use 12 most frequent Action Units (AU)
- Provides **gender**:
  - *Male – Female*
- and **race** information:
  - *Black – White – Latino - Asian*

UNIVERSITY OF CAMBRIDGE

[1] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias andfairness in facial expression recognition," in Computer Vision –ECCV 2020 Workshops, A. Bartoli and A. Fusiello, Eds.Cham:Springer International Publishing, 2020, pp. 506–523.

Experiment 1 – Facial Expression Recognition

Experiment 1: Fairness Scores across Gender and Race for the RAF-DB Dataset. **Bold** values denote best while [bracketed] denote second-best values for each column.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.834	0.943	0.816	0.937
Offline	0.944	0.925	0.954	0.974
Non-CL-based Bias Mitigation Methods				
DDC [44]	0.968	0.985	0.961	0.976
DIC [44]	0.938	0.989	0.962	0.965
SS [15]	0.955	0.961	0.954	0.975
DA [45]	0.975	0.858	[*0.997*]	0.919
Continual Learning Methods				
EWC [23]	0.972	0.987	0.983	0.990
EWC-Online [39]	0.970	0.987	0.974	0.990
SI [47]	**0.990**	**0.996**	**0.999**	**0.996**
MAS [2]	[*0.980*]	[*0.990*]	0.990	[*0.994*]
NR [22]	0.928	0.974	0.923	0.974

Experiment 2 – Action Unit Detection

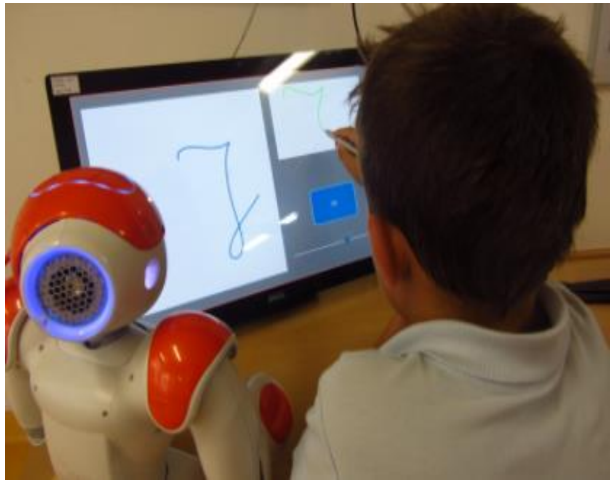
Experiment 2: Fairness Scores across Gender and Race for the BP4D Dataset. **Bold** values denote best while [bracketed] denote second-best values for each column.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.962	0.855	0.941	0.858
Offline	0.984	0.878	[*0.994*]	0.901
Non-CL-based Bias Mitigation Approaches				
DDC [44]	[*0.990*]	0.920	0.991	0.924
DIC [44]	0.979	0.925	0.985	0.922
SS [15]	0.977	0.920	0.983	0.919
DA [45]	**0.994**	[*0.954*]	**0.995**	[*0.962*]
Continual Learning Approaches				
EWC [23]	0.981	0.949	0.992	0.943
EWC-Online [39]	0.976	0.937	[*0.994*]	0.957
SI [47]	0.986	0.946	0.965	0.954
MAS [2]	0.966	0.920	0.967	0.909
NR [22]	0.983	**0.966**	0.954	**0.974**

Conclusion

- Proposed the novel usage of continual learning for developing fairer models
- Highlighted how CL can help **mitigate bias**
- Showed that CL methods are able to **balance learning** across different domains
- **Outperformed** non-CL based approaches
- Can be used in:
    - real-word scenarios with **embedding them onto a humanoid robot**
    - long-term social interactions with under-represented population groups
    - investigating how CL-based FER systems respond to users from different demographics.

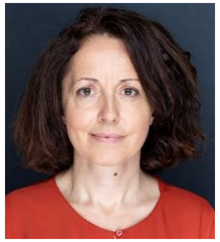
(a)



(b)

(a) https://www.scmp.com/lifestyle/health-wellness/article/3024028/how-robot-nurses-could-help-care-worlds-elderly-and
(b) https://navigator8972.github.io/navigator8972.github.io/portfolio/portfolio-3-hwhri/

Acknowledgement



Ozgur Kara **Nikhil Churamani** **Hatice Gunes**

- N. Churamani is funded by the EPSRC under grant EP/R513180/1(ref.2107412).
- H. Gunes' work is supported by the EPSRC under grant ref. EP/R030782/1.
- The authors also thank Prof Lijun Yin from Binghamton University (USA) for providing access to the BP4D Dataset and the relevant race attributes; and Shan Li, Profs Weihong Deng and Jun Ping Du from Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.