# Transfer Fair Representation Across Domains

Mentors: Dr. Furong Huang, Bang An
REU-CARR 2021

# Goal of Transferring Fairness Across Domains

- There are scenarios where we may have labeled data in one domain but unlabeled data in another domain.
  - **feature, label, sensitive attribute** triplet in source domain and **features** in target domain

- We would like to extend fairness guarantee in not just the source domain, but also a target domain

- We propose an adversarial network structure for ensuring a fair and accurate classifier for target domain with available data

# What is Fairness?

- Choosing a fairness metric is dependent on the domain concerns and the problem setting

- Our research specifically addresses ensuring
    - **Equalized Odds**
        - A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal true positive rate and equal false positive rate

$$P(R=+|Y=y, A=a) = P(R=+|Y=y, A=b) \quad y \in \{+,-\} \quad \forall a, b \in A$$

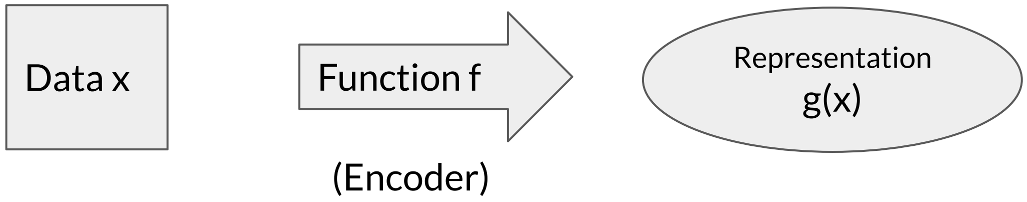
**sensitivity, recall, hit rate, or true positive rate (TPR)**

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

**fall-out or false positive rate (FPR)**

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

# Overview of Fair Representation Learning

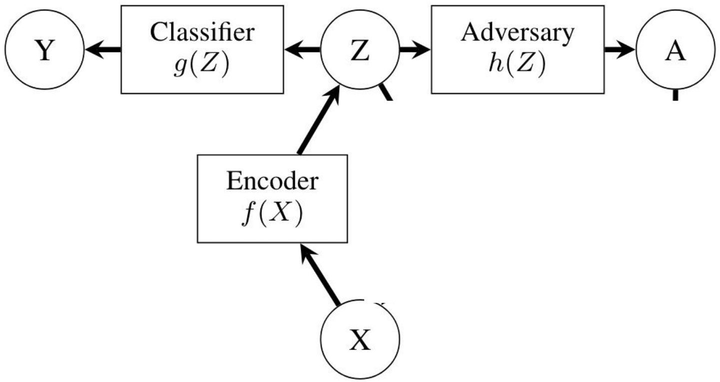



- Finding a function **f** that encodes data containing sensitive attribute to a representation that is debiased
  - Raw data may contain both explicit and implicit sensitive attributes
    - For race, the implicit attribute may be zip code

- Deriving fair representation is beneficial for avoiding exploitation or discrimination using the data representation
  - Data vendor may provide the fair representation instead of raw data

# Learning Adversarially Fair and Transferable Representations

Madras, Creager, Pitassi, Zemel




X: input data (images of faces)
Z: representation
A: predicted sensitive attribute (race)
Y: predicted output (gender)

Proposes adversarial network structure and theoretical theorems for bounds on fairness definitions.

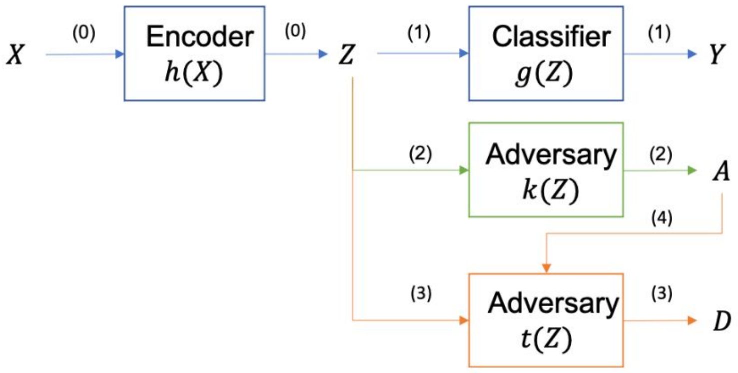
- The adversarial network seeks to predict the sensitive attribute
  - Maximize adversarial objective ,

$$L_{Adv}^{DP}(h) = 1 - \sum_{i\in\{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a)\in\mathcal{D}_i} |h(f(x,a)) - a|$$

- The classifier seeks to predict the label. A decoder is optional to reconstruct original data.
  - Minimize classifier loss and reconstruction loss

Can be supervised or unsupervised

# Extending Adversarially Fair Representation to Target Domain

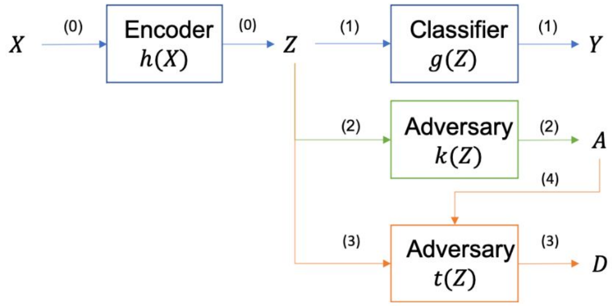



X: input data (images of faces)
Z: representation
A: predicted sensitive attribute (race)
D: predicted domain (source or target)
Y: predicted output (gender)

- Use the fairness adversary's predicted sensitive attribute as input for domain adversary in predicting the domain.
- By minimizing the loss of the two adversaries and the loss of the classifier, we expect to find a feature Z that is indistinguishable in its sensitive attribute and its domain.

$$\mathcal{L}(h,g,k,t) = \underbrace{\alpha\mathcal{L}_C(g(h(X_S)),Y_S)}_{\text{Classifier loss}} + \underbrace{\beta\mathcal{L}_{Adv}(k(h(X_S)),A_S)}_{\text{Fair adversary loss}} + \underbrace{\gamma\mathcal{L}_{Adv}(t(h(X),k(h(X))),D)}_{\text{Domain adversary loss}}.$$

# Implementation




X: input data (images of faces)
Z: representation
A: predicted sensitive attribute (race)
D: predicted domain (source or target)
Y: predicted output (gender)

- Dataset: UTK Face
  - Creates source and target domain by choosing source to be ages 10-40, and target to be ages over 40
  - Generates binary case by subsetting the white and black races only
  - Predicting for gender
- Encoder and classifier: VGG network
  - Not pretrained since pretraining dataset might include target domain distribution
  - Last two fully connected layers as classifier, rest as encoder
- Adversaries: two fully connected layers
  - Strong adversary achieving 95%+

# Results for Determining Gender

* mean

| | Source (10-40 age) | | | | Target (>40 age) | | | |
|---|---|---|---|---|---|---|---|---|
| | White TPR | Black TPR | White FPR | Black FPR | White TPR | Black TPR | White FPR | Black FPR |
| pretrained | 0.913* | 1* | 0.097* | 0.06* | 0.69* | 0.75* | 0.036* | 0.012* |
| Baseline (0-1) | 0.725 | 0.500 | 0.246 | 0.250 | 0.503 | 0.333 | 0.185 | 0.091 |
| Baseline-fair (0-1-2) | 0.706 | 0.600 | 0.201 | 0.250 | 0.471 | 0.417 | 0.166 | 0.091 |
| Baseline-fair-transfer (source-only) (0-1-2-3) | 0.791 | 0.600 | 0.276 | 0.333 | 0.551 | 0.333 | 0.274 | 0.091 |
| Baseline-fair-transfer (source+target) (0-1-2-3) | 0.758 | 0.700 | 0.299 | 0.333 | 0.556 | 0.417 | 0.236 | 0.182 |

White acc:0.739, Black acc:0.636

White acc:0.645, Black acc:0.609

# Results for Determining Gender

* mean

| | Source (10-40 age) | | | | Target (>40 age) | | | |
|---|---|---|---|---|---|---|---|---|
| | White TPR | Black TPR | White FPR | Black FPR | White TPR | Black TPR | White FPR | Black FPR |
| pretrained | 0.913* | 1* | 0.097* | 0.06* | 0.69* | 0.75* | 0.036* | 0.012* |
| Baseline (0-1) | 0.725 | 0.500 | 0.246 | 0.250 | 0.503 | 0.333 | 0.185 | 0.091 |
| Baseline-fair (0-1-2) | 0.706 | 0.600 | 0.201 | 0.250 | 0.471 | 0.417 | 0.166 | 0.091 |
| Baseline-fair-transfer (source-only) (0-1-2-3) | 0.791 | 0.600 | 0.276 | 0.333 | 0.551 | 0.333 | 0.274 | 0.091 |
| Baseline-fair-transfer (source+target) (0-1-2-3) | 0.758 | 0.700 | 0.299 | 0.333 | 0.556 | 0.417 | 0.236 | 0.182 |

White acc:0.732, Black acc:0.682

White acc:0.651, Black acc:0.609

# Results for Determining Gender

* mean

| | Source (10-40 age) | | Target (>40 age) | |
|---|---|---|---|---|
| | TPR Diff | FPR Diff | TPR Diff | FPR Diff |
| Pretrained | 0.087 | 0.037* | 0.06 | 0.024 |
| Baseline (0-1) | 0.225 | 0.004 | 0.170 | 0.094 |
| Baseline-fair (0-1-2) | 0.106 | 0.049 | 0.054 | 0.075 |
| Baseline-fair-transfer (source-only) (0-1-2-3) | 0.191 | 0.057 | 0.218 | 0.183 |
| Baseline-fair-transfer (source+target) (0-1-2-3) | 0.058 | 0.034 | 0.139 | 0.054 |

# Results for Determining Gender

* mean

| | Source (10-40 age) | | Target (>40 age) | |
|---|---|---|---|---|
| | White Acc | Black Acc | White Acc | Black Acc |
| Baseline (0-1) | 0.739 | 0.636 | 0.645 | 0.609 |
| Baseline-fair (0-1-2) | 0.749 | 0.682 | 0.637 | 0.652 |
| Baseline-fair-transfer (source-only) (0-1-2-3) | 0.760 | 0.636 | 0.631 | 0.609 |
| Baseline-fair-transfer (source+target) (0-1-2-3) | 0.732 | 0.682 | 0.651 | 0.609 |

## Conclusion and Next Steps

- We produced fairness improvement in source and target domain using fair and domain adversaries for our task of gender classification while maintaining accuracy

- Tune the regular VGG to increase baseline accuracy on gender classification task and additional datasets
- Compare Fair-Transfer with fair representation result fed into the domain adversarial network
- Hyperparameter search for improving Fair-Transfer results
  - Balance the hyperparameter in the loss function to account for the two adversarial loss
- Extend the theoretical bounds for transfer fairness

# Thank you!

*Thank you to Dr. Furong Huang, Bang An, my group, as well as the REU program for supporting my research.*