Transfer of Machine Learning Fairness across Domains

PRESENTER: **Candice** Schumann

INTRO

- It is common for models to be used on multiple tasks.
- This is concerning for machine learning **fairness**.
- Traditionally **domain adaptation** is used when the distribution of training and validation data does not match the target distribution.
- We ask the **question**: if the model is trained to be "fair" on one dataset, will it be "fair" over a different distribution of the data?

CONTRIBUTIONS

1. We provide theoretical bounds on transferring equality of opportunity and equality of odds metrics across domains and discuss the insights gained from these bounds.
2. We offer a general, theoretically-backed modeling objective that enables transferring fairness across domains.
3. We demonstrate when transferring machine learning fairness works successfully, and when it does not, through both synthetic and realistic experiments.

THEOREM

Theorem 1. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_0^0}$, $\mathcal{U}_{S_1^0}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^0}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^0}$, $\mathcal{D}_{S_1^0}$, $\mathcal{D}_{T_0^0}$, and $\mathcal{D}_{T_1^0}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0}) + 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0,$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

IMPLICATIONS

Each relevant subspace in the source and target space should be "close" (for instance the negatively labeled minority in the source should be "close" to the negatively labeled minority in the target). Additionally, the model should have enough capacity to perform in both domains.

The First **Theoretical** Examination of **Transfer** of Machine Learning **Fairness** Across **Domains**

Take a picture to **download** the **full paper**

GENERAL MODEL

- Using the theorem we provide a general model described in the figure below.
- A loss function can be defined as follows

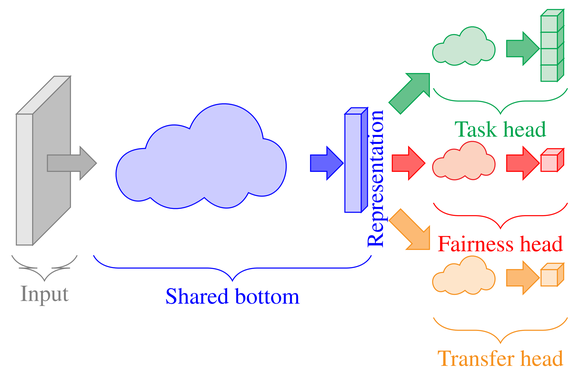
$$\min \Big[\sum_{Z \in \mathcal{D}_S \cup \mathcal{D}_T} L_Y(f(Z), g(Z)) + \sum_{(A, Z^0) \sim \mathcal{D}_{S^0}} \lambda_{Fair} L_{MMD}\left(a(h(Z^0)), A\right) + \sum_{(d, Z^0) \sim (\mathcal{D}_{S^0} \cup \mathcal{D}_{T^0})} \lambda_{DA} L_{MMD}\left(d(h(Z^0)), d\right) \Big]$$

- The first term is the loss for the given task (green task head).
- The second term minimizes the difference between sensitive groups in the source domain using MMD (red fairness head). This term minimizes the first term in Theorem 1:

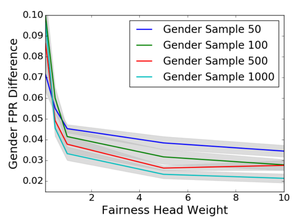
$$\Delta_{EOp_S}(g)$$

- The final term minimizes the difference between the source and target domain using MMD (orange transfer head). The term minimizes the second two terms in Theorem 1 when balanced data is used:

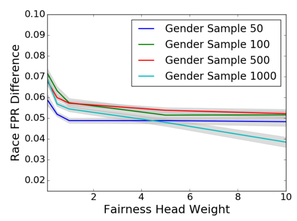
$$\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})$$



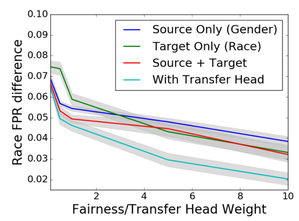
RESULTS



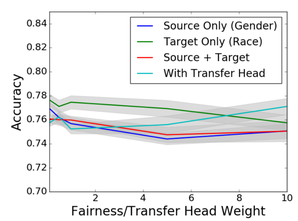
(a) Effect of fairness head: Improving $\Delta_{EOp_{\text{gender}}}$ with varying number of gender-balanced samples.



(b) Some natural transfer occurring without explicit transfer: $\Delta_{EOp_{\text{race}}}$ is improved with gender data.



(c) Effect of transfer head: better transfer from gender (1000 samples) to race (50 samples).



(d) Accuracy graph for transferring from gender (1000 samples) to race (50 samples).

Candice Schumann,
Xuezhi Wang,
Alex Buetel,
Jilin Chen,
Hai Qian,
Ed H. Chi

Google

UNIVERSITY OF MARYLAND