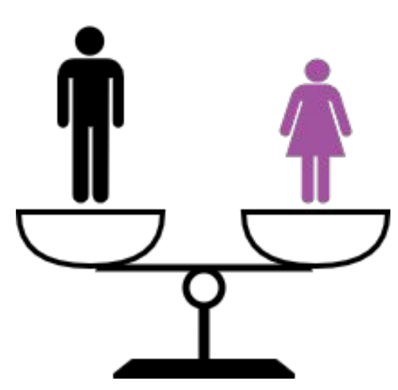
Google Research

# Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, Ed H. Chi
{yuyanw,xuezhiw,alexbeutel,fprost,jilinc,edchi}@google.com

## Introduction

### Fairness

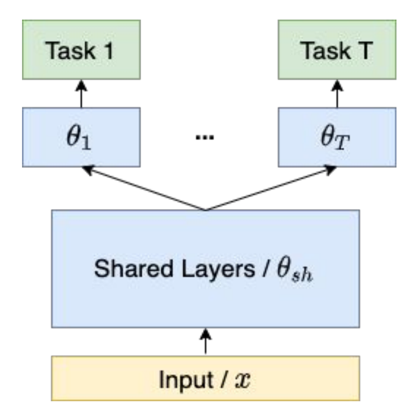


**Objective:** Subgroups are treated equally.

**Why:** Critical for decision making in employment, education etc.

**Mostly studied in single-task learning problems.**

### Multi-Task Learning (MTL)



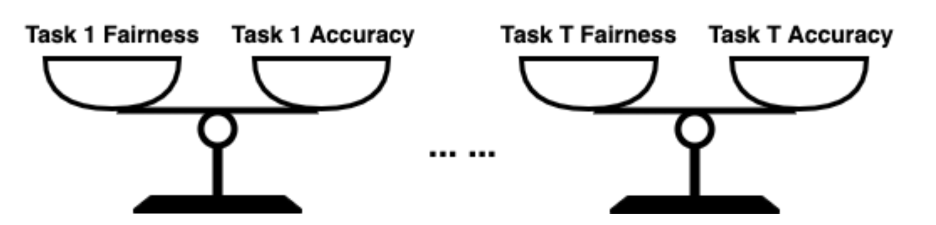

**Objective:** Jointly learn multiple tasks.

**Why:** Transfer learning / regularization / model efficiency/...

**Mostly only focused on optimizing accuracy across multiple tasks.**

### What we know:

- For single task, fairness comes at a cost of accuracy;
- MTL comes with an accuracy trade-off among tasks;




### What we don't know:

- How does fairness play out in the multi-task scenario?
- How to characterize the multi-dimensional fairness-accuracy trade-off?
- Can we improve the Pareto frontier?

## Understanding

### Fairness Implications in MTL

MTL may have **larger** impacts on fairness goals than on accuracy goals...

|  | T1 Error | T1 FPR Gap | T2 Error | T2 FPR Gap |
|---|---|---|---|---|
| STL-T1 | 0.2030 | 0.2716 | - | - |
| STL-T2 | - | - | 0.0784 | 0.0145 |
| MTL | 0.2035 | 0.2846 | 0.0783 | 0.0137 |
| Difference | +0.24% | **+4.78%** | -0.08% | **-5.39%** |

**(a) CelebA**: MTL hurts Task 1 fairness but improves Task 2 fairness.

.. or **hurt** the fairness of some tasks while benefiting from its accuracy gains.

|  | T1 Error | T1 FPR Gap | T2 Error | T2 FPR Gap |
|---|---|---|---|---|
| STL-T1 | 0.1659 | 0.1200 | - | - |
| STL-T2 | - | - | 0.1313 | 0.0661 |
| MTL | 0.1656 | 0.1205 | 0.1299 | 0.0738 |
| Difference | -0.20% | +0.34% | **-1.10%** | **+11.60%** |

**(b) UCI-Adult**: MTL improves Task 2 accuracy but hurts its fairness.

**STL-T1**: single-task learning for Task 1;
**STL-T2**: single-task learning for Task 2;
**MTL**: multi-task learning with equal task weight.

Training multiple tasks together by simply pooling the accuracy objectives may lead to **unwanted fairness consequences.**

### Measuring Fairness in MTL

Can we efficiently summarize and visualize the multi-dimensional Pareto frontier?

- Moreover, fairness/accuracy metrics could differ largely across different tasks (e.g. some tasks are intrinsically harder to learn / have more bias).

Measuring **relative change** over single-task learning (STL), and average across tasks:

Average Relative Fairness Gap

$$ARFG := \frac{1}{T}\sum_{t=1}^{T} FPRGap^{(t)} / \underline{FPRGap_S^{(t)}},$$

single-task FPR gap w/o fairness remediation

Average Relative Error

$$ARE := \frac{1}{T}\sum_{t=1}^{T} Err^{(t)} / \underline{Err_S^{(t)}}.$$

single-task error w/o fairness remediation

## Methods

### Improving Fairness in MTL

Using **FPR gap** as the measure for group fairness...

$$\hat{\mathcal{L}}_{STL}(f) = \hat{\mathcal{L}}(f) + \lambda \hat{\mathcal{F}}(f|N),$$

set of negative examples

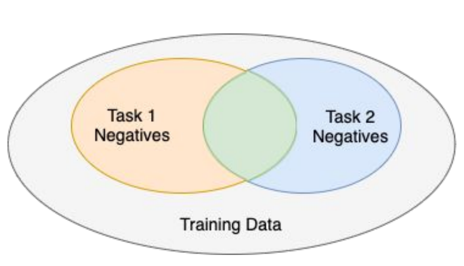
accuracy loss

group fairness loss

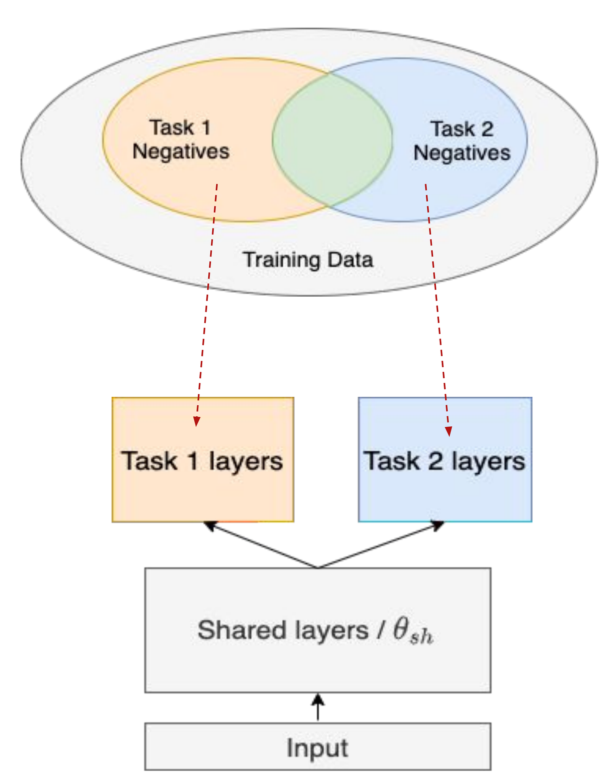
- correlation loss [1-2]
- Max Mean Discrepancy [3]
- FPR difference [4-6]

closing the gap between predictions on **negative examples (Y=0)**

If we generalize this naively to MTL...



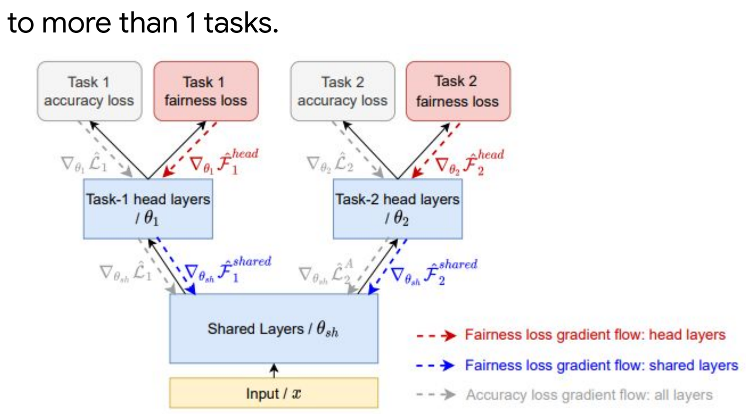

- Baseline: Fairness loss computed on & & ;
- However, is **only** relevant to Task 1 fairness;
- Likewise, is **only** relevant to Task 2 fairness;
- But Baseline method does **not** distinguish between them => A **suboptimal** use of model capacity!




### MTA-F: Multi-task-aware fairness treatment

Let's address the fairness in a more targeted way:

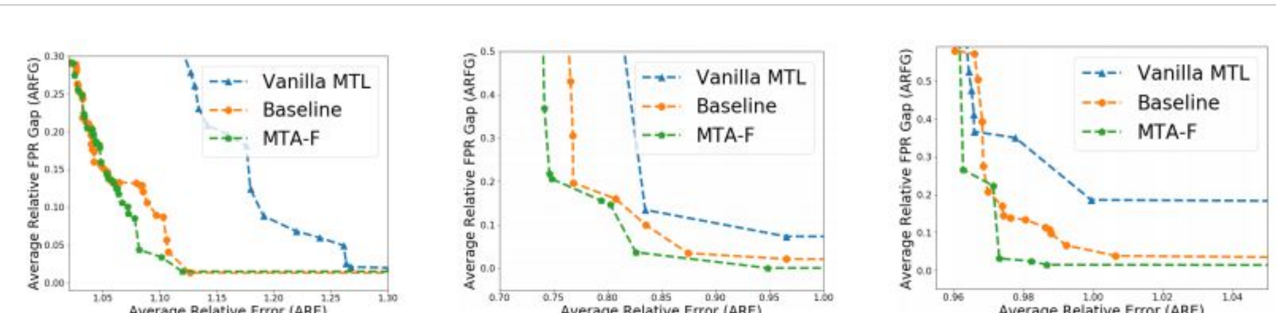
- Head layers address fairness issues that are ***specific*** to the task itself;
- Shared layers address fairness issues that are ***common*** to more than 1 tasks.




**(b)** Backpropagation with MTA-F: We backpropagate task-specific fairness losses $\hat{\mathcal{F}}_t^{head}$ to head layers, and the remaining fairness loss $\hat{\mathcal{F}}_t^{shared}$ to shared layers ($t = 1, 2$).

## Experiments

- Datasets:
  - **UCI-Adult**: *Income > $50k* (T1), *Capital Gain > 0* (T2)
  - **German Credit Data**: *Good loans* (T1), *Credit > 2000* (T2)
  - **LSAC Law School**: *Pass bar* (T1), *high GPA* (T2)
- Methods:
  - **Vanilla MTL**: plain MTL **without** fairness mitigation
  - **Baseline**: Per-task fairness treatment
  - **MTA-F**: our proposed method
- Fairness loss: correlation loss / MMD loss / FPR gap loss
- Fairness metric: Equal Opportunity between females and males




(a) UCI-Adult. (b) German Credit Data. (c) LSAC Law School.

**Figure 2:** *ARFG-ARE* Pareto frontier. Lower-left indicates better Pareto optimality, i.e. better overall fairness-accuracy trade-off.

Overall fairness gap

Overall error rate

| Dataset | UCI-Adult |  | German Credit |  | LSAC Law School |  |
|---|---|---|---|---|---|---|
| Metric | *ARFG* | *ARE* | *ARFG* | *ARE* | *ARFG* | *ARE* |
| Vanilla MTL | 0.3444 | 1.1040 | 0.1336 | 0.8367 | 0.3497 | 0.9778 |
| Baseline | 0.0871 | 1.1032 | 0.0999 | 0.8356 | 0.1126 | 0.9864 |
| **MTA-F** | **0.0437** | **1.0820** | **0.0364** | **0.8264** | **0.0310** | **0.9731** |

**Table 3:** Average relative fairness gap (*ARFG*) and average relative error (*ARE*) on **UCI-Adult**, **German Credit Data** and **LSAC Law School** datasets, as defined in Section 4. Lower metric values indicate better overall fairness / accuracy across all tasks.

Task 1 Task 2

|  |  | $T_1$ *Err* | $T_1$ *FPRGap* | $T_2$ *Err* | $T_2$ *FPRGap* |
|---|---|---|---|---|---|
| UCI-Adult | Vanilla MTL | 0.1911 | 0.0715 | 0.1359 | 0.0091 |
|  | Baseline | 0.1938 | 0.0186 | 0.1336 | 0.0020 |
|  | **MTA-F** | **0.1891** | **0.0083** | **0.1319** | **0.0016** |
| German Credit | Vanilla MTL | 0.205 | 0.0150 | 0.220 | 0.0084 |
|  | Baseline | 0.255 | 0.0879 | **0.180** | 0.0069 |
|  | **MTA-F** | **0.200** | **0.0033** | 0.220 | **0.0034** |
| LSAC Law School | Vanilla MTL | 0.1555 | 0.0503 | **0.1565** | **0.0004** |
|  | Baseline | 0.1568 | 0.0119 | 0.1580 | 0.0006 |
|  | **MTA-F** | **0.1540** | **0.0015** | **0.1565** | **0.0004** |

**Table 4:** Per-task metrics for **UCI-Adult, German Credit Data** and **LSAC Law School** datasets.

## References

[1] Beutel et al. Fairness in recommendation ranking through pairwise comparisons. KDD 2019.
[2] Beutel et al. Putting fairness principles into practice: Challenges, metrics, and improvements. AIES 2019.
[3] Prost et al. Toward a better trade-off between performance and fairness with kernel-based distribution matching. NeurIPS 2019 "ML with Guarantees" workshop.
[4] Feldman et al. Certifying and removing disparate impact. KDD 2015.
[5] Menon et al. The cost of fairness in binary classification. FAccT 2018.
[6] Zafar et al. Fairness Constraints: A Flexible Approach for Fair Classification. JMLR 2019.