Sustaining Fairness via Incremental Learning

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi
{somnath, snigdha}@cs.unc.edu
UNC Chapel Hill

UNC NLP

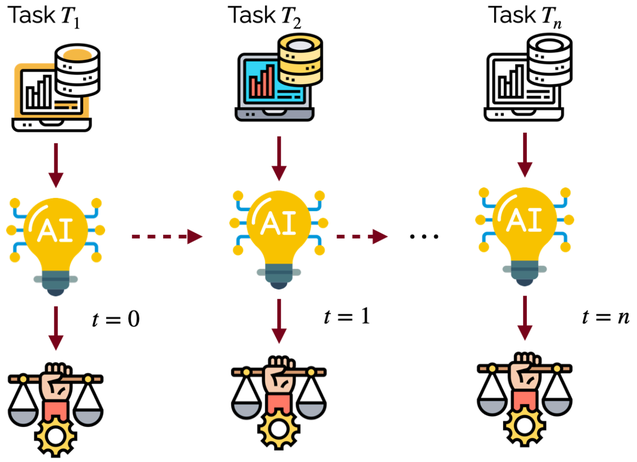
brcsomnath/FaIRL

Introduction

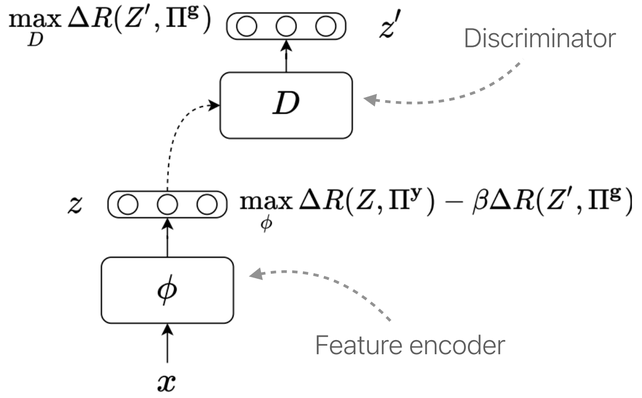
- Ensuring fairness of ML approaches is crucial in sensitive applications like hiring
- Current fairness approaches are trained and tested on a single data domain
- These approaches fail to remain fair under domain shift
- This calls for approaches that can function in the wild

Fair Incremental Learning

- We tackle the above problem by incrementally learning new tasks while ensuring fairness
- Most incremental learning systems focus on target task
- We introduce FaIRL, that learns fair representations while acquiring knowledge of new tasks



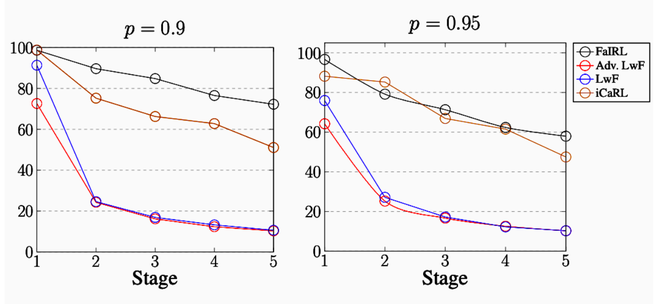
Fairness-aware Incremental Representation Learning (FaIRL)



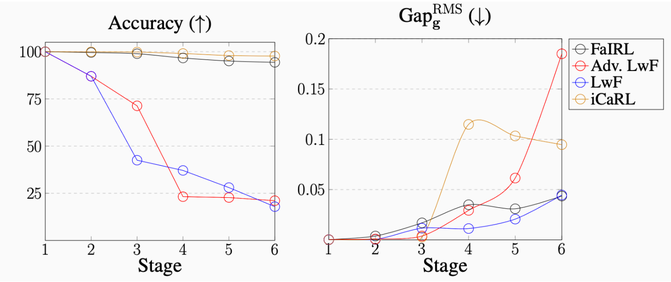
- The debiasing framework shown above is learns fair representations for a single domain
- This framework learns compact representations, given a task, making it amenable to incremental learning
- Feature encoder is modified to achieve the following:
  - Discriminative representation for X_{new}: $\max \Delta R(Z_{new}, \Pi^{y}_{new})$
  - Protect leakage for X_{new}: $\min \Delta R(Z'_{new}, \Pi^{g}_{new})$
  - Retain old subspaces: $\min \Delta R(Z_{old}, \bar{Z}_{old})$
  - Protect leakage for X_{old}: $\min \Delta R(Z'_{old}, \Pi^{g}_{old})$
- We sample X_{old} using either random sampling, prototype sampling or submodular optimization

Evaluation

- Accuracy over different training stages on Biased MNIST



- Accuracy and TPR-GAP over different training stages on Biography Classification dataset



Conclusion

- We propose FaIRL, that learns fair representations in an incremental fashion
- FaIRL controls the rate-distortion function of representations
- FaIRL outperforms existing methods by significant margin
- FaIRL is a first step towards achieving fairness in the wild