Towards Unbiased and Accurate Deferral to Multiple Experts

Vijay Keswani
Yale University

Matthew Lease
University of Texas at Austin
Amazon AWS AI

Krishnaram Kenthapadi
Amazon AWS AI

Human-in-the-loop frameworks

- Desirable to augment ML model predictions with expert inputs. Useful for
  - Improving accuracy
  - Incorporating human expertise
  - Auditing models
- **Popular examples**
  - Healthcare models
  - Content moderation
  - Risk assessment and screening

Errors and biases in ML

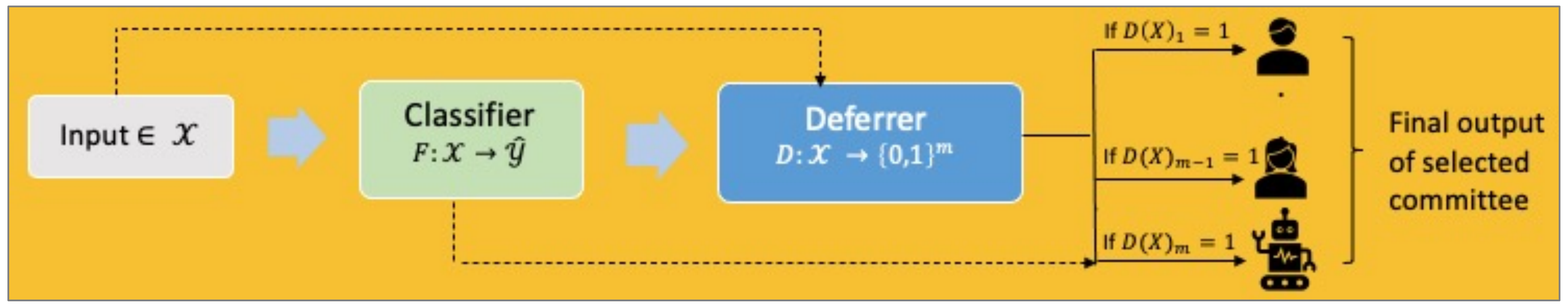
- Human-in-the-loop frameworks can reflect biases or inaccuracies of the human experts. Examples
  - Racial bias in human-in-the-loop framework for recidivism assessment (Green, Chen 2019)
  - Ethical concerns regarding audits of facial processing technologies (Raji et al. 2020)
  - Automation bias in time critical decision support systems (Cummings 2004)

Can we design human-in-the-loop frameworks that account for expertise and biases of human experts?

Prior work

- *Rejection learning* - Pass when not sure; experts are not explicit here (El-Yaniv et al. 2010, Cortes et al. 2016, Kamiran et al. 2012, Li et al. 2011)
- *Learning to defer or joint decision-making with explicitly specified human(s)*
  - Theoretical analysis limited to single expert (Madras et al. 2018, Mozannar, Sontag 2020)
  - Empirical analysis limited to studying correlations from data (Green and Chen 2019, De-Arteaga et al. 2020, Yaghini et al. 2019)

Can we design frameworks that
- **can handle multiple (kinds of) experts,**
- **has feasible optimization formulation,**
- **improves accuracy and fairness of predictions?**



Our joint learning framework

- X – non-protected attributes; Y – class label; Z – protected attribute
- $m-1$ experts available : $E_1, \dots, E_{m-1}$; classifier F is the m-th expert
- For any input X, decision vector $Y_E(X) \coloneqq [E_1(X), E_2(X), \dots, E_{m-1}(X), F(X)]$

Learn classifier $F: \mathcal{X} \to \mathcal{Y}$ using loss L_{clf} (e.g. log-loss)

Learn deferrer $D: \mathcal{X} \to [0,1]^m$ as follows:

$$Y_D = \text{sigmoid}\left(D(X)^\top \cdot Y_E(X)\right),$$

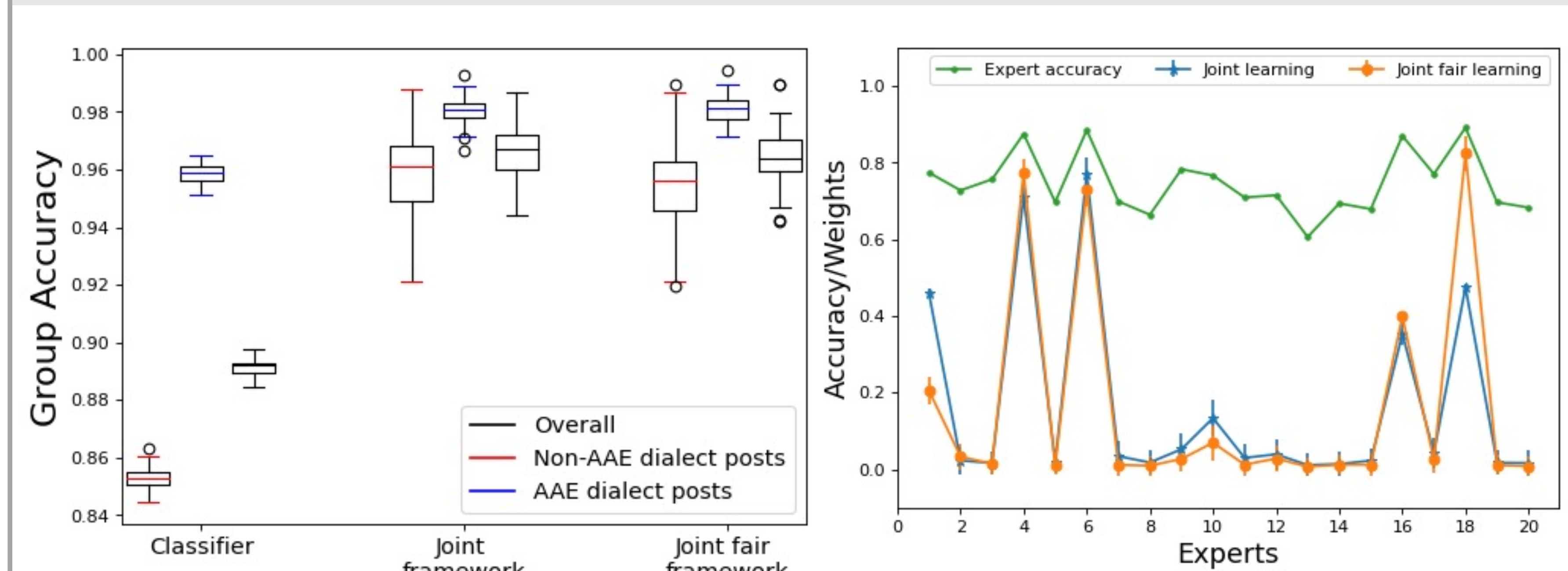
$$L_D = -\mathbb{E}_{X,Y}\left[Y \cdot \log Y_D + (1-Y) \cdot \log(1-Y_D)\right]$$

$$\min_{F,D} L_{clf} + \alpha \cdot L_D$$

- **Fair learning –** Can ensure predictions are *fair* w.r.t Z using additional regularizers or Minimax-Pareto fairness formulation (Martinez et al. 2020; Diana et al. 2021)
- Use **dropout** to prevent overfitting and **cost regularizers** for individual expert costs.

Empirical analysis for content moderation

- Hate-speech detection using Twitter dataset from Davidson et al. (2017)
- Protected attribute – dialect labels of the post (African-American English -AAE- or not)
- 20 synthetic experts with 14 biased against AAE and 6 biased against non-AAE dialect



Our framework learns the classifier and deferrer simultaneously and leads to improved overall and group-specific accuracies

Theoretical Properties

- Projected Gradient Descent can be used to obtain optimal classifier and deferrer
- Intuitive gradient updates- rewards *good* experts
- If L_{clf} is Lipschitz-smooth, then projected-gradient descent converges close to optimal classifier and deferrer in time polynomial in number of experts
- Deferrer weights can be used to choose committees of smaller sizes as well

Analysis using real-world dataset

- Dataset - 1471 Twitter posts
- MTurk survey presented to 170 participants to label whether post is offensive or not
- Overall accuracy of aggregated response – 87%
- Heterogeneous expert domains - 92 participants had higher accuracy for non-AAE posts, 75 participants had higher accuracy for AAE posts

Performance of framework for this dataset

Method	Overall Accuracy	Non-AAE Accuracy	AAE Accuracy
Classifier only	.78 (.02)	.76 (.05)	.80 (.04)
Joint framework	.85 (.03)	.87 (.04)	.83 (.03)
Joint balanced framework	.84 (.03)	.87 (.03)	.81 (.04)
Joint minimax framework	.85 (.02)	.87 (.02)	.83 (.02)

Our framework improves the accuracy and fairness of the final prediction for this real-world dataset as well, despite heterogeneity in expert performances

Discussion, Limitations and Future Work

- *Larger real-world datasets* can be constructed for more robust analysis of hybrid frameworks
- Extension to settings where *experts are replaceable* or when more experts can be added
- Improved selection of *smaller committees*

Paper: https://arxiv.org/abs/2102.13004