Probing Classifiers are Unreliable for Concept Removal and Detection

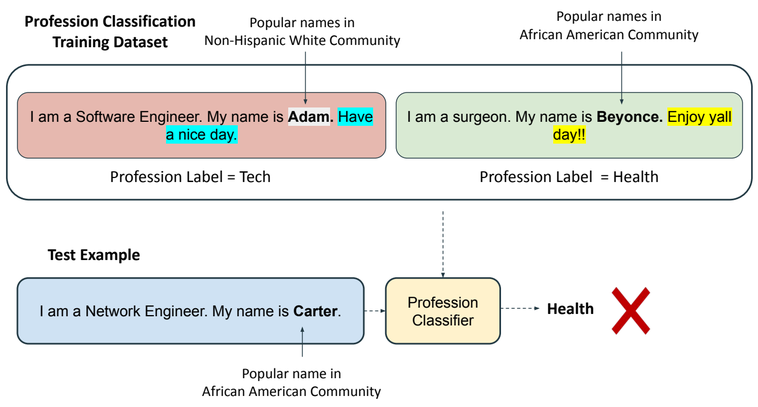
Abhinav Kumar[1], Chenhao Tan[2], Amit Sharma[1]

[1]Microsoft Research India, [2]University of Chicago

[1]{t-abkumar,amshar}@microsoft.com, [2]chenhao@uchicago.edu

arXiv: 2207.04153

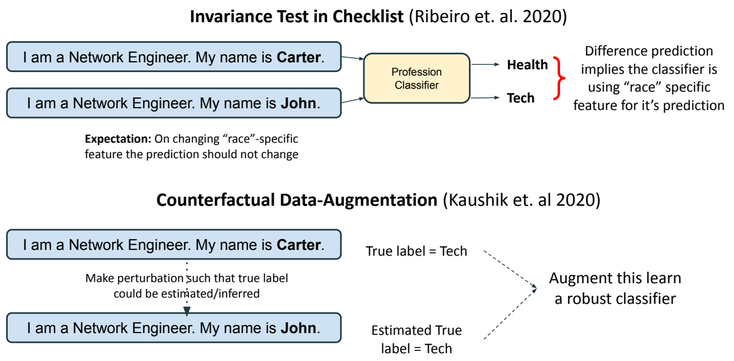
Background



- ML models tend to rely of spurious signal for prediction
- Eg. Race specific names correlated to Profession feature.
- At test time if correlation breaks → wrong prediction.

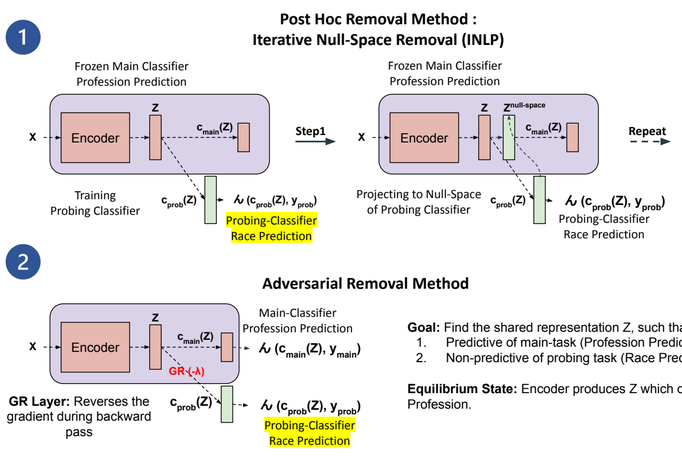
Input Space Based Removal Methods

- Methods like Checklist → discover such bias
- Makes perturbation in input to test model behavior.
- **Counterfactual Augmentation**: Add perturbed data with expected label to breaks spurious correlation.



Latent Space Based Removal Methods

- Making perturbation in input is not always possible.
- Make perturbation or changes in latent space.
- **Null Space Removal (INLP)**: Removes spurious features by projecting latent space to null-space of spurious feature classifier.
- **Adversarial Removal (ADV)**: Jointly trains main-task and spurious feature classifier adversarially.



Our Contribution

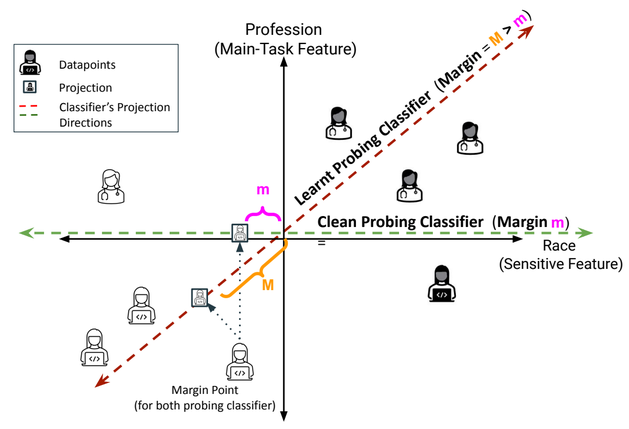
1. Any method using Probing classifier will fail.
2. Theoretical and Empirical Results showing failure of INLP and Adversarial Removal on 3 real-world dataset and 1 synthetic-Text.

Dataset Description

- **Multi-NLI**: Main Task = Contradiction Prediction, Spurious Feature = Negation Words.
- **Twitter AAE**: Main Task = Sentiment Prediction, Spurious Feature = Race.
- **Synthetic-Text**: Main Task = Presence of Numbered Word, Spurious Feature = Length of Sentence.

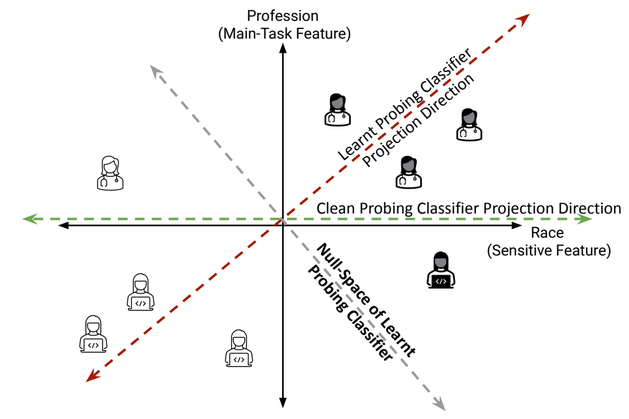
Example: Probing Failure

Goal: Learn a *clean* probing (Race) classifier which just uses probing (Race) feature (green dashed line).



Observation: Slanted *unclean* classifier (red) is better than *clean* probing classifier (green) when trained with max-margin objective.

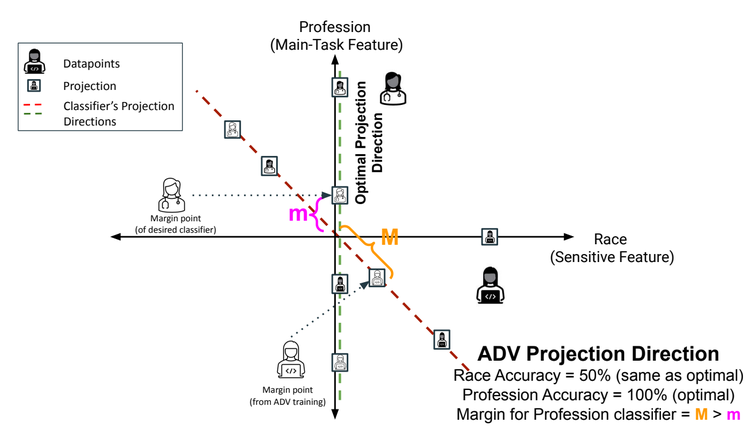
Example: INLP Probing Classifier



- From **Lemma 2.1** → Probing Classifier is slanted
- Null-Space of Probing Classifier is also slanted (wrong).

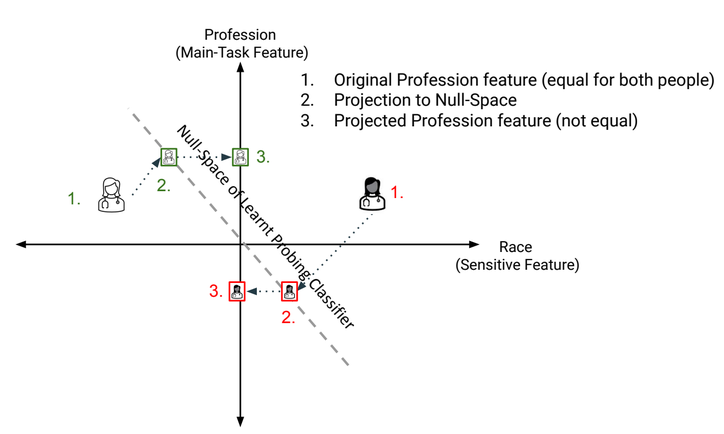
Example: ADV Removal Failure

Goal: Learn shared 1D latent representation s.t. only main-task feature (Profession) is present.



Observation: Slanted Projection Direction (red) better in Main-task objective and equivalent in Adversarial Objective than desired direction (green).

Example: INLP Feature Corruption



- Two individual with same profession but different race.
- Projecting to *wrong* Null Space → inverts Profession.

INLP Empirical Result

Expectation: Clean Main-Task classifier is given as input to INLP, so it should have no effect on main-task classifier.

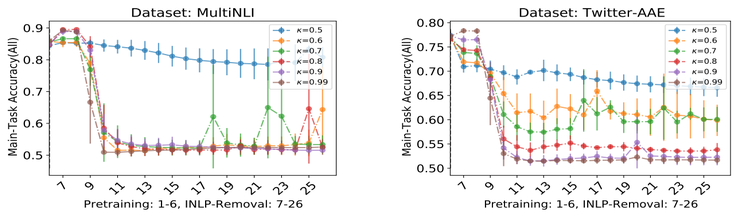


Figure 1: Variation of Main Task accuracy with INLP steps. Main-Task Accuracy goes to random guess as INLP proceeds.

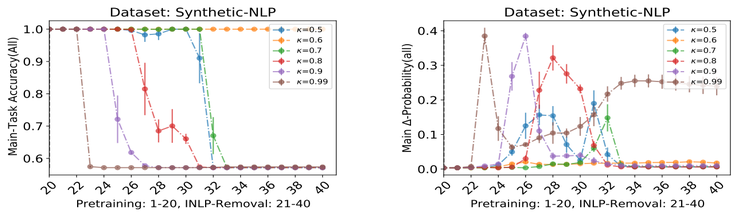


Figure 2: Measuring ΔProb i.e Change in prediction probability of classifier by changing the spurious feature

Early Stopping Doesn't Help: ΔProb increases in the initial phase of INLP. Stopping early could lead to relatively more *unclean* classifier.

ADV Empirical Result

Metric: Post Adversarial Training accuracy on subset of data where spurious correlation breaks (minority group).

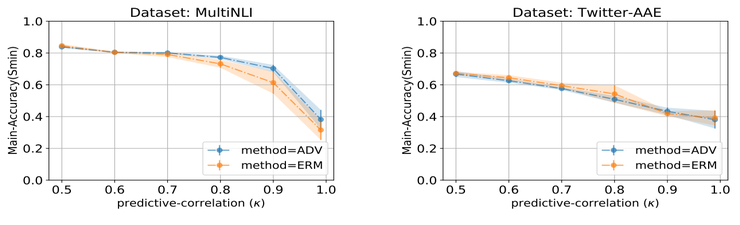


Figure 3: Variation of Main-Task classifier's accuracy on minority group as we vary the degree of correlation between main and spurious feature. Little to no improvement in minority group accuracy across dataset and degree of correlation.

Assumptions

1. Latent Space is disentangled and frozen.
2. Probing feature is linearly separable w.r.t. probing label for binary case.
3. Main-Task feature is linearly separable w.r.t. probing label for the margin point of *clean* probing classifier.

Theory: Probing Failure

Lemma 2.1 *(Informal)* Given Assm 1,2,3 is satisfied:

- $c_{prob}(\boldsymbol{z}) = \boldsymbol{w}_{prob} \cdot \boldsymbol{z}_{prob} + \boldsymbol{w}_{main} \cdot \boldsymbol{z}_{main}$ where $\boldsymbol{w}_{main} \neq \boldsymbol{0}$.
- Generalized version for any classifier in paper.

Theory: ADV Failure

Theorem 2.3 (*Informal*) Given adversarial removal methods is just training the last layer. Then there exist an *unclean* shared latent representation ($\boldsymbol{Z}$) s.t.:

- $\text{Margin}^{main}(\text{unclean } \boldsymbol{Z}) = \text{Margin}^{main}(\text{clean } \boldsymbol{Z})$.
- $\text{Acc}^{adv}(\text{unclean } \boldsymbol{Z}) = \text{Acc}^{adv}(\text{clean } \boldsymbol{Z})$.
- $\mathcal{L}^{adv}(\text{unclean } \boldsymbol{Z}) > \mathcal{L}^{adv}(\text{clean } \boldsymbol{Z})$, when main-task and probing labels are correlated for the margin point of *clean* main-task classifier.

Theory: INLP Failure

Theorem 2.2 (*Informal*) Given the probing classifier used by INLP is trained using max-margin objective (Lemma 2.1), following happens:

Mixing: After first step of INLP we have:

- $\boldsymbol{z}_{main}^{after} = \psi(\boldsymbol{z}_{main}^{before}, \boldsymbol{z}_{prob}^{before})$
- $z_{prob}^{after} = \phi(\boldsymbol{z}_{main}^{before}, \boldsymbol{z}_{prob}^{before})$
- Mixing is non-invertible and removal of sensitive concept could lead to removal of main-task specific feature.

Destruction:

- $\|\boldsymbol{Z}\|$ decreases after every step.
- Continued removal leads to complete destruction of $\boldsymbol{Z}$.