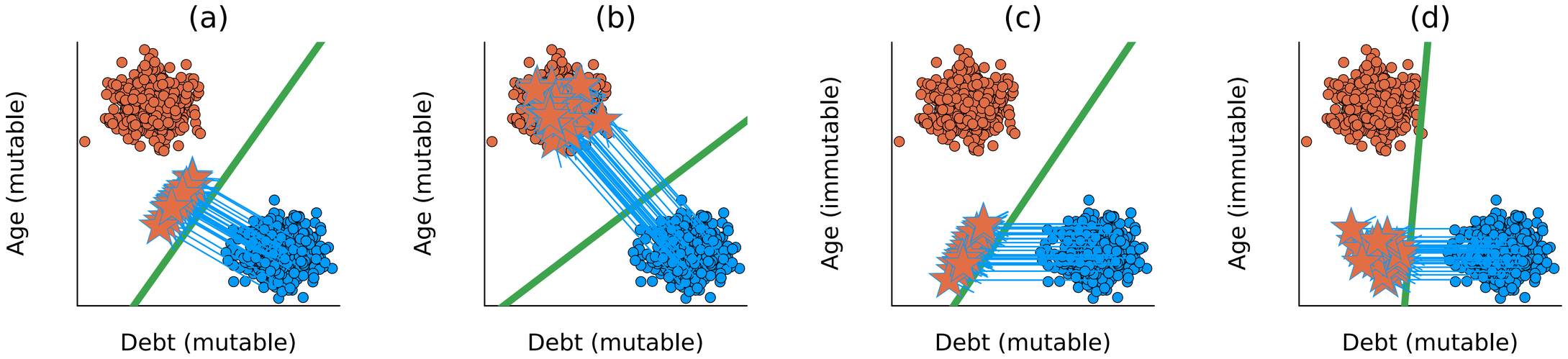
Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Patrick Altmeyer, Aleksander Buszydlik, Arie van Deursen, Cynthia C.S. Liem
(Delft University of Technology)

Counterfactual training leverages counterfactual explanations to make models more 1) explainable, 2) sensitive to actionability constraints and 3) adversarially robust.



Methodology

Counterfactual search objective (equation (1)):

$$\min_{\mathbf{x}' \in \mathcal{X}^D} \{\text{yloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}^+) + \lambda \text{reg}(\mathbf{x}')\}$$

Counterfactual training objective (equation (2)):

$$\min_\theta \text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}} \text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y^+; \theta)$$
$$+\lambda_{\text{adv}} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{\text{AE}}), \mathbf{y}_{\text{AE}}) + \lambda_{\text{reg}} \text{ridge}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta)$$

Pseudo-Code for counterfactual training:

Require: Training dataset $\mathcal{D}$, initialize model $\mathbf{M}_\theta$
while not converged **do**
  Sample $\mathbf{x}'_0 \sim \mathbf{X}$, $y^+ \sim \mathcal{U}(\mathcal{Y})$ and $\mathbf{x}^+ \sim \mathbf{X}^+$
  for $t = 1$ to T **do**
    Backpropagate $\nabla_{\mathbf{x}'}$ through equation (1)
    Store $\mathbf{x}'_{\text{CE}}, \mathbf{x}'_{\text{AE}}, \mathbf{y}_{\text{AE}}$
  end for
  Sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from dataset $\mathcal{D}$
  Distribute $(\mathbf{x}'_{\text{CE}i}, y^+{}_i, \mathbf{x}'_{\text{AE}i}, \mathbf{y}_{\text{AE}i}, \mathbf{x}^+{}_i)_{i=1}^{n_{\text{CE}}}$
  for each batch **do**
    Backpropagate ∇_θ through equation (2)
  end for
end while
return $\mathbf{M}_\theta$

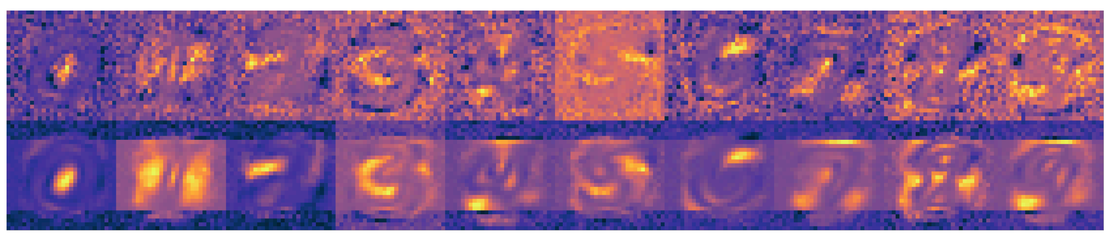
Summary

On-the-fly counterfactual explanations are used to induce explainability and actionability through contrastive divergence and recycled as adversarial examples to further improve robustness.

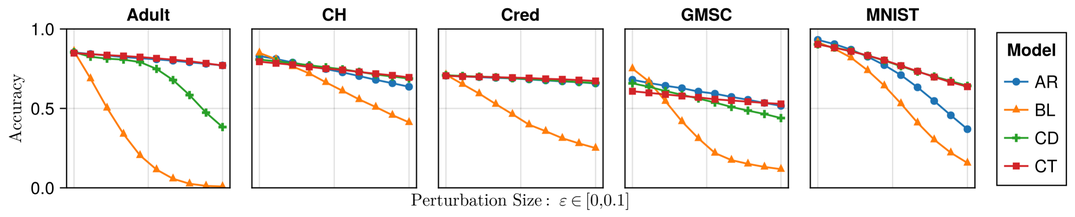
Improved Plausibility

Data	IP (−%)	IP* (−%)	Cost (−%)
LS	$26.26 \pm 0.67^*$	$51.28 \pm 2.01^*$	$16.41 \pm 0.57^*$
Circ	$58.88 \pm 0.37^*$	$93.84 \pm 6.70^*$	$42.99 \pm 0.85^*$
Moon	$19.59 \pm 0.73^*$	8.00 ± 9.44	$5.16 \pm 1.00^*$
OL	-1.93 ± 1.12	-27.70 ± 14.59	$40.86 \pm 2.30^*$
Adult	0.19 ± 1.05	$34.35 \pm 5.61^*$	4.03 ± 4.03
CH	$10.65 \pm 1.47^*$	$63.06 \pm 4.25^*$	$44.23 \pm 1.43^*$
Cred	$10.14 \pm 1.59^*$	$50.35 \pm 12.26^*$	$-18.17 \pm 4.40^*$
GMSC	$10.65 \pm 2.28^*$	$24.75 \pm 4.84^*$	$66.01 \pm 1.41^*$
MNIST	$6.36 \pm 1.70^*$	-70.31 ± 217.60	$-35.11 \pm 6.96^*$
Avg.	15.64	25.29	18.49

Improved Actionability



Improved Robustness



Further Resources

arxiv.org/abs/2601.16205
IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)
p.altmeyer@tudelft.nl