IMPERIAL

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Causal Representation Learning in Healthcare

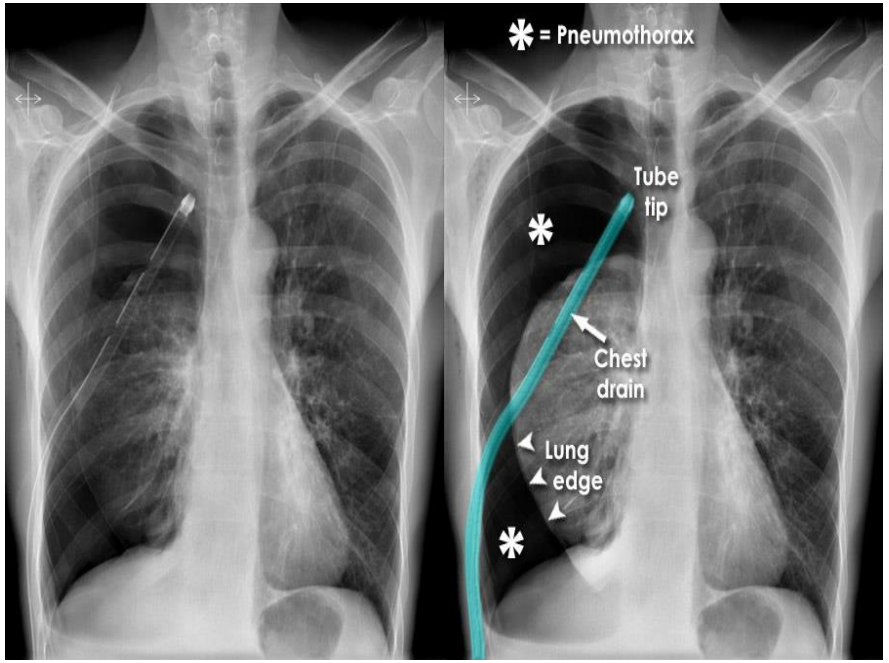
Sonali Parbhoo

Royal Statistical Society's Emerging Applications Seminar

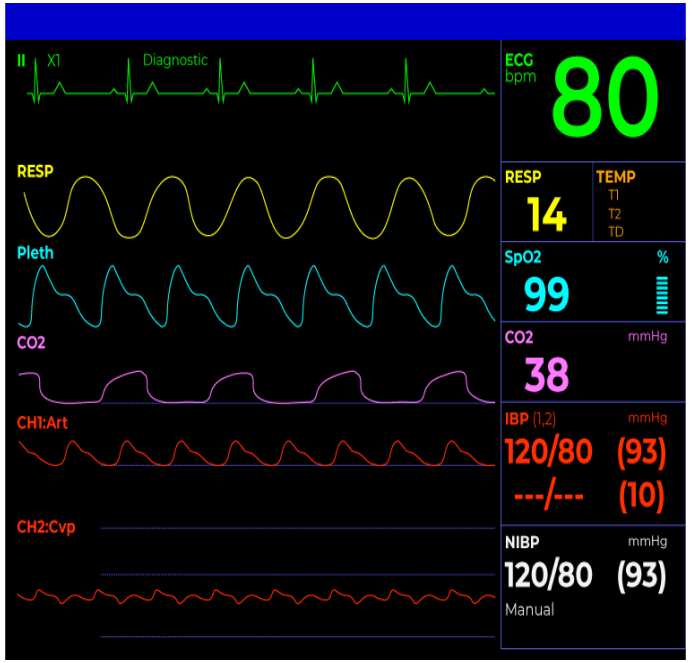
December 12, 2025

In collaboration with Haoyang Hong, Ioanna Papanikolaou, Indiradharshini Thanigaikumar, and discussions with the AI4AI lab and many others.

Healthcare systems face spurious correlations everywhere



Hospital annotations or artifacts can leak into X-ray images.



ICU bedside waveform monitoring can contain device-driven artifacts.

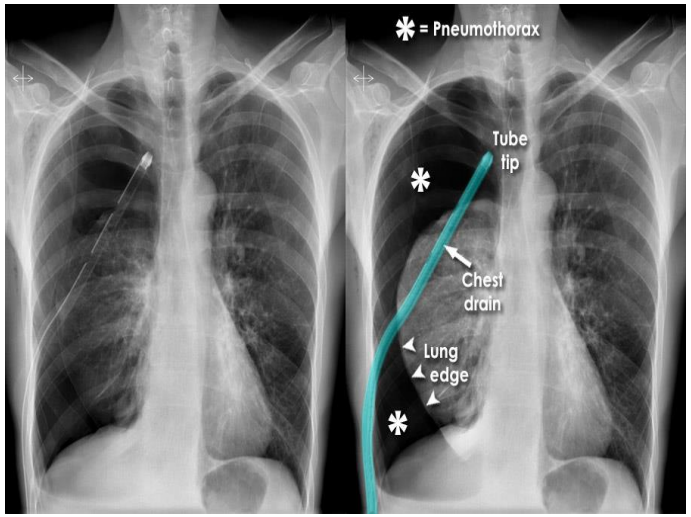


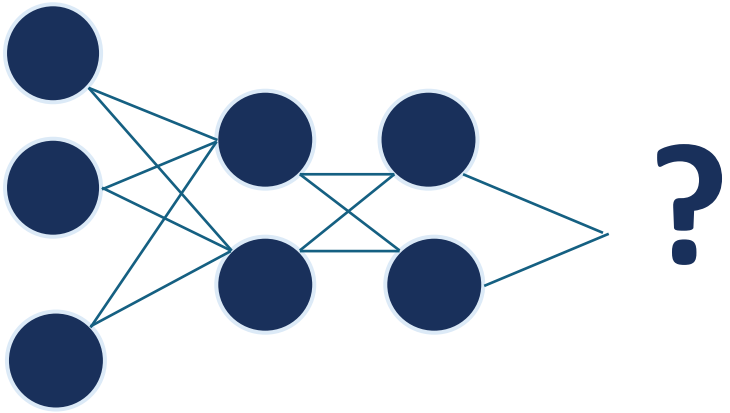
Electronic Health Records frequently have missingness.

Downstream decisions depend on representations that capture **causal signal**, not spurious correlations.

Shortcuts and spurious correlations

Models can learn to exploit **shortcuts (e.g. presence of tubes)** instead of pathology for a downstream prediction





A **shortcut** is a spurious attribute or set of attributes that correlates with outcomes, but is **irrelevant for the problem**.

Why Shortcuts Matter

- Shortcut reliance in a model can result in **poor out-of-distribution behaviour.** If a model learns to rely on shortcuts, it may perform well in development settings but fail catastrophically in deployment.

- **May also result in incorrect diagnosis/triage in healthcare.**

Shortcut reliance **undermines trust, safety, fairness, and generalizability**

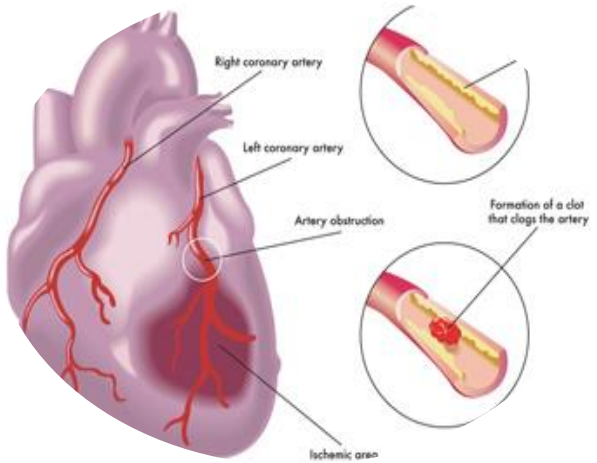
Why Shortcuts Matter: Example

- Troponin is only measured at hospital A when clinicians suspect a heart attack. Missing troponin thus strongly correlates with low mortality, but not causally.

  "**Troponin missing → low risk; Troponin present → high risk"**

- When deployed at test time, if troponin is measured more frequently at hospital B, the shortcut breaks and performance of the model significantly drops.



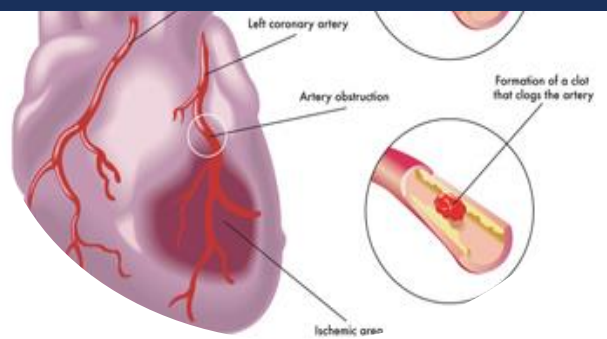


Why Shortcuts Matter: Example



How do we learn models that mitigate shortcuts to preserve causal signal?

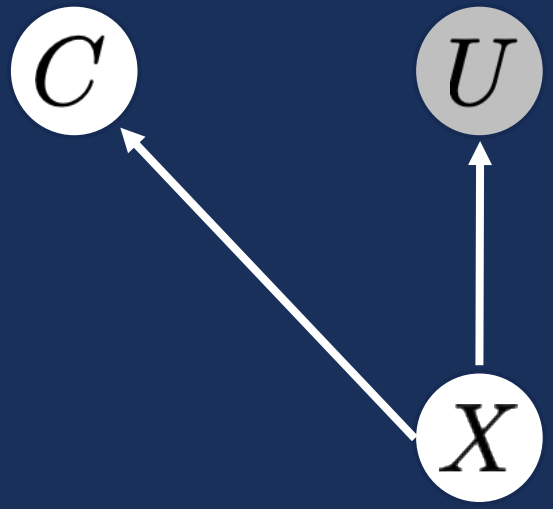
- When deployed at test time, if troponin is measured more frequently at hospital B, the shortcut breaks and performance of the model significantly drops.



Problem Setup

Known representation

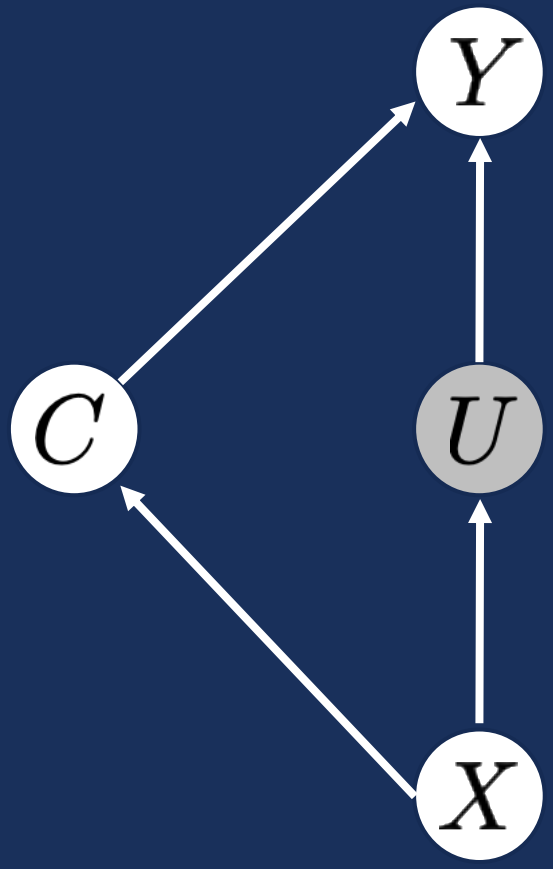
E.g. Radiographical annotations/findings



Unknown representation

Features from data

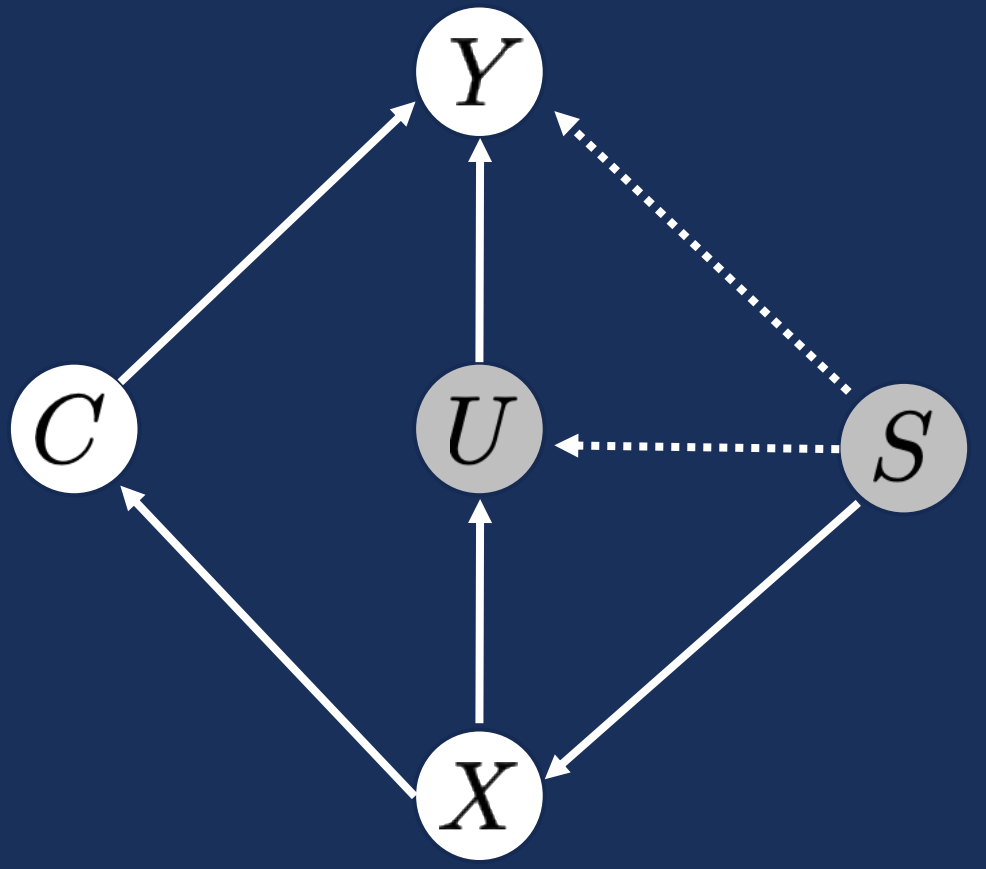
Problem Setup



Label

E.g. Diagnosis

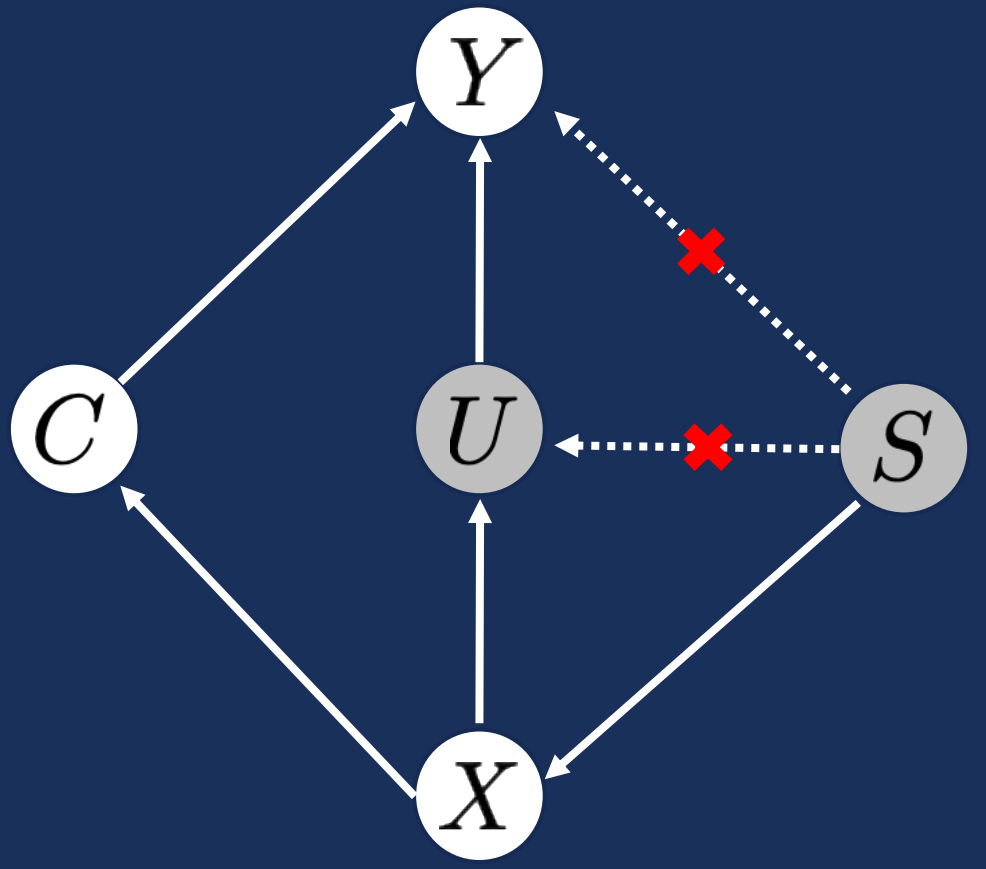
Problem Setup



Shortcut features

E.g. Scanner Type

Problem Setup



Shortcut features

E.g. Scanner Type

As we've all probably tried in our papers…..

- Use more data modalities to regularize out the effects of shortcut features

As we've all probably tried in our papers…..

- Use more data modalities to regularize out the effects of shortcut features

  …. but more modalities can **increase the number of shortcuts affecting the representation** (Hong et al 2025, AISTATS)

Do Regularization Methods for Shortcut Mitigation Work as Intended?, Hong, Papanikoulaou and Parbhoo, AISTATS 2025.

As we've all probably tried in our papers…..

- Explicitly **regularize shortcut features** in model to reduce their contribution
  - Introduce some **inductive biases** to reduce reliance on shortcut features

Do Regularization Methods for Shortcut Mitigation Work as Intended?, Hong, Papanikoulaou and Parbhoo, AISTATS 2025.

Standard Approach: Regularization of Shortcut Features

- Requires **explicit knowledge** of shortcut features.
- Reduce the contribution of shortcut features by,
  - **Sparsifying the representation** using L_1 regularization
  - Using L_2 to force the representation to be **close to the MAP solution** (if we assume Gaussian feature contributions)
  - **Using Expert Yielded Estimates (EYE) regularization:** A combination of L_1 and L_2 regularization that incorporates expert knowledge about covariates and outcome of interest.

$$J([\theta_{u,s}, \theta_c]) = ||\theta_{u,s}||_1 + \sqrt{||\theta_{u,s}||_1^2 + ||\theta_c||_2^2}$$

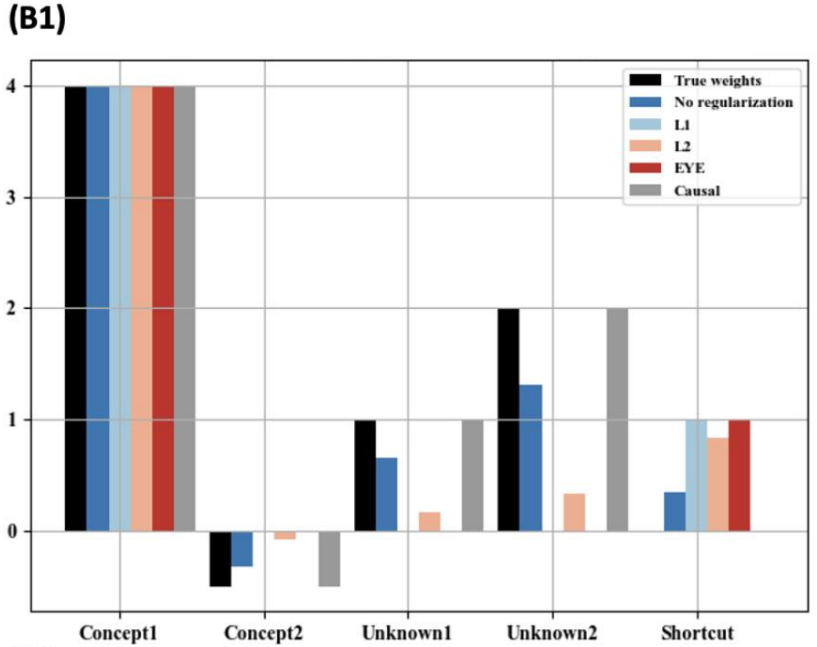
Standard Approach: Regularization of Shortcut Features

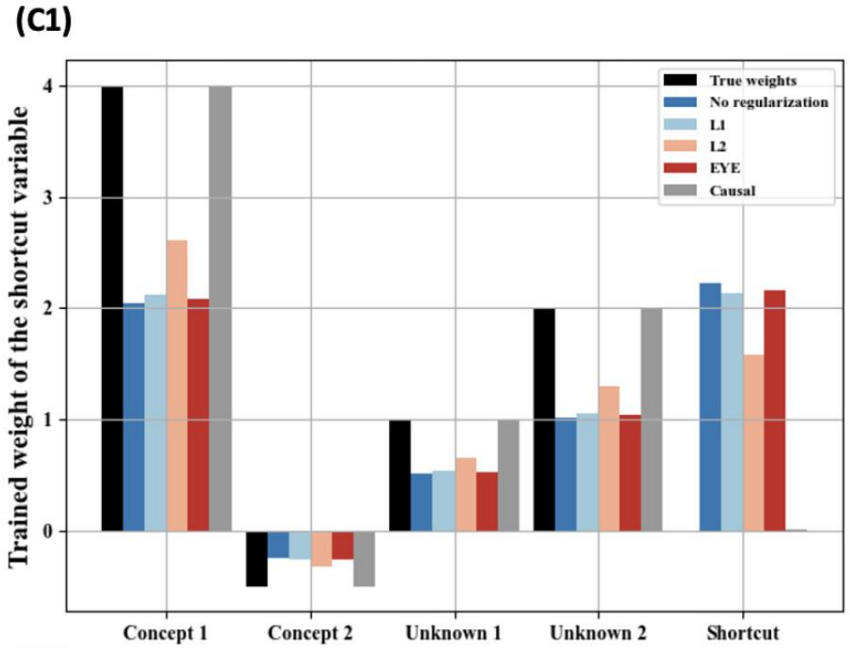
- Requires **explicit knowledge** of shortcut features.
- Reduce the contribution of shortcut features by,
  - **Sparsifying the representation** using L_1 regularization
  - Using L_2 to force the representation to be **close to the MAP solution** (if we assume Gaussian feature contributions)
  - **Using Expert Yielded Estimates (EYE) regularization:** A combination of L_1 and L_2 regularization that incorporates expert knowledge about covariates and outcome of interest.

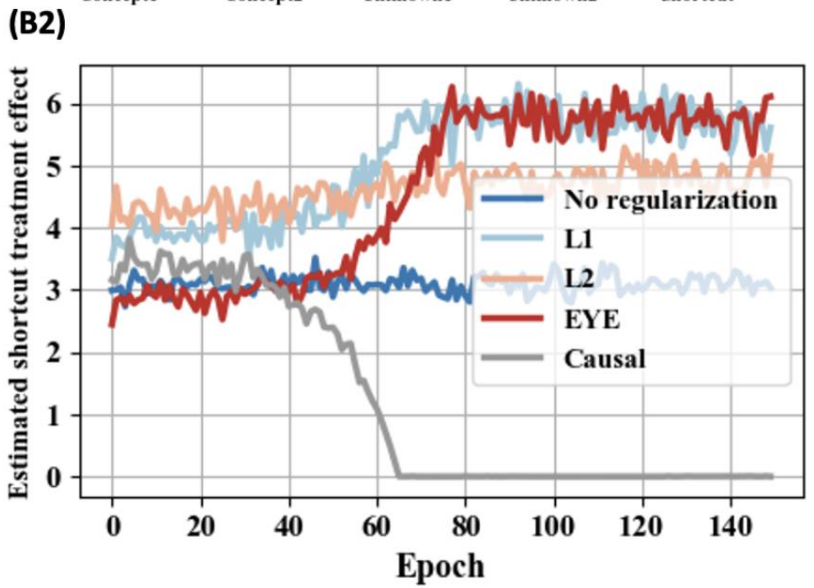
$$J([\theta_{u,s}, \theta_c]) = ||\theta_{u,s}||_1 + \sqrt{||\theta_{u,s}||_1^2 + ||\theta_c||_2^2}$$

What happens empirically?

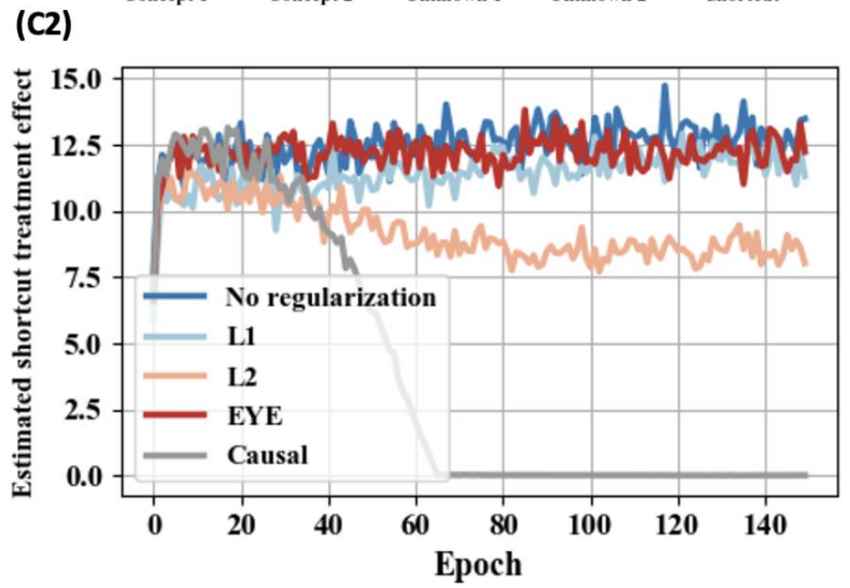
Hong et al, 2025 AISTATS – Do Regularization Methods for Shortcut Mitigation work as Intended?





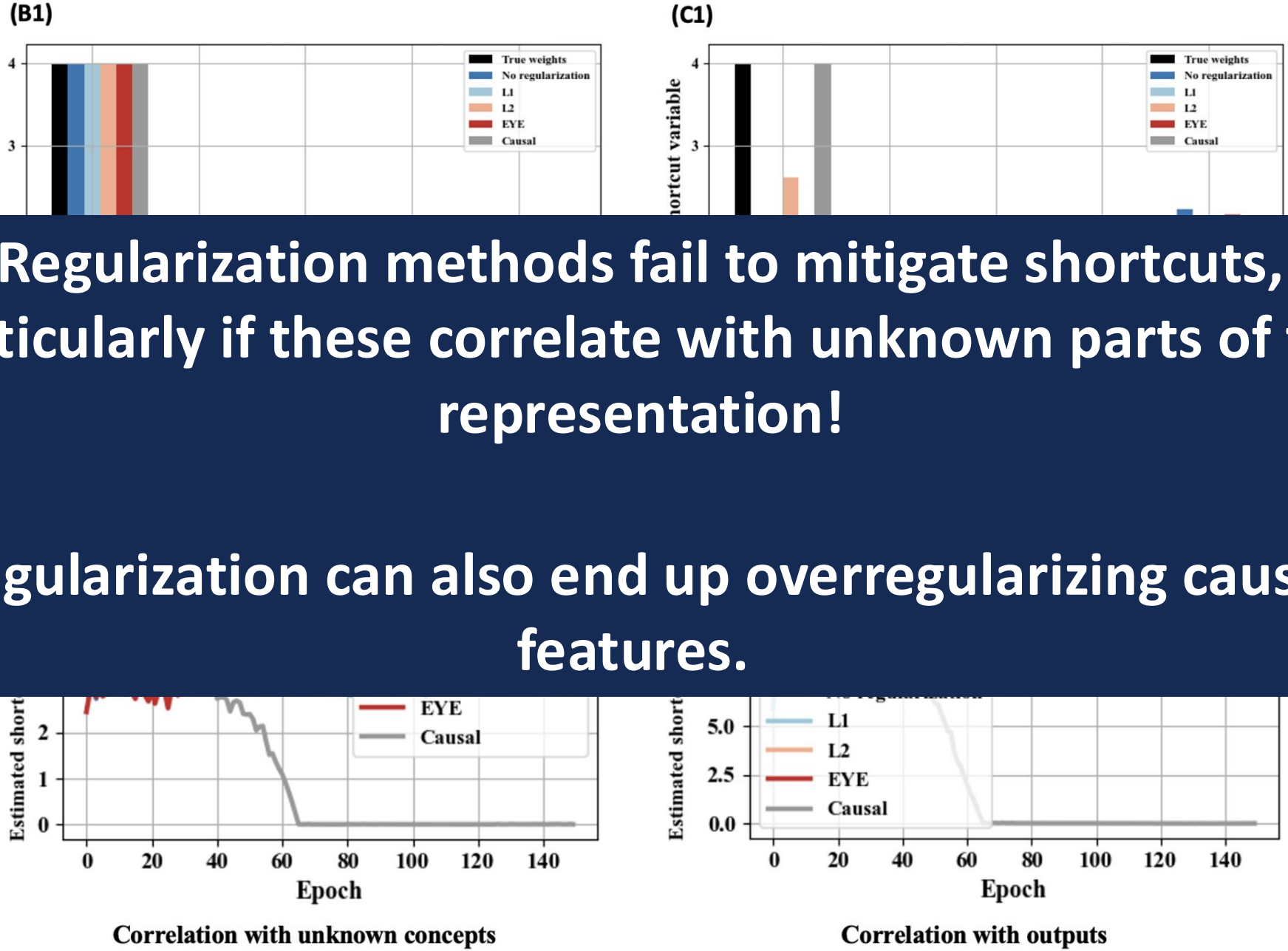


Correlation with unknown concepts



Correlation with outputs

Hong et al, 2025 AISTATS – Do Regularization Methods for Shortcut Mitigation work as Intended?



Regularization methods fail to mitigate shortcuts, particularly if these correlate with unknown parts of the representation!

Regularization can also end up overregularizing causal features.

Why does this happen?

- When a model's inductive bias aligns closely with spurious patterns in the training data, it can **exacerbate the problem of spurious correlations.**

Why does this happen?

Consider the simplest setting

$$\hat{y} = X\beta \quad \text{where} \quad \beta = [\beta_{\mathbf{c}}, \beta_{\mathbf{s}}]$$

A regularized ERM objective considers

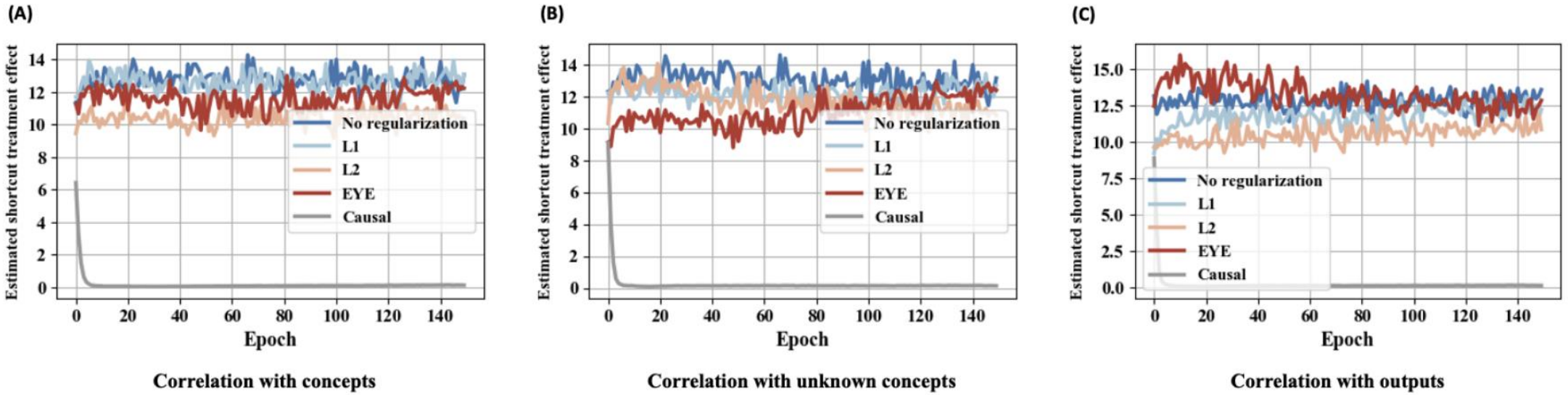
$$-\min_{\beta} \frac{1}{n} ||Y - X\beta||^2 + \lambda R(\beta)$$

Theoretically, it can be shown that if shortcut features **S** are predictive and correlated with causal features **C** or unknown features **U** (Hong et al 2025, AISTATS)

- L2 regularization **shrinks β_c and β_s together**
- L1 /sparsity regularization **may select S over C**
- **||β_s|| ≠ 0 even under strong regularization**

The issue gets worse under more shortcuts

Hong et al, 2025 AISTATS – Do Regularization Methods for Shortcut Mitigation work as Intended?



Is there any alternative to regularization that can help mitigate shortcuts in causal representation learning?

First make an **additional assumption**

Assume we have some small subset of the complete set of data $\mathcal{D}_m \subset \mathcal{D}_1$ for which we know the true labels Y

We use this labelled subset as an **anchor** for unlabelled data.

Key Insight: Edit Gradients, Not Data

Instead of modifying data or using regularization, **edit the gradients** themselves to force the model away from shortcut regions.

- Use the gradients from labelled subset g_L to **redirect shortcut-aligned** components in gradients from unlabelled data g_U

We term this procedure **Fix-A-Shortcut (FAS).**

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

Fix-A-Shortcut

- FAS modifies the standard weight update in gradient descent

$$w \leftarrow w - \eta(g_L + \lambda_u g_U)$$

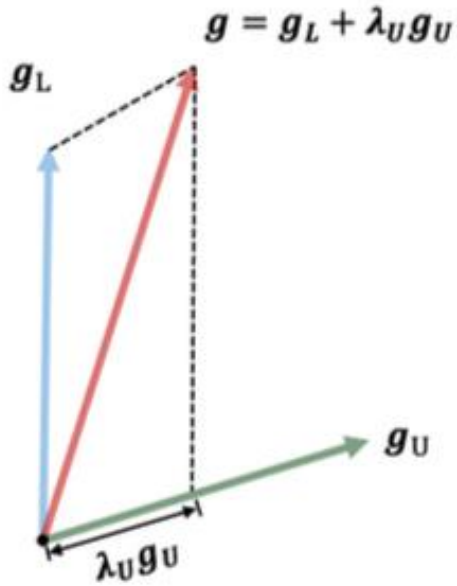
by applying a **gradient-space intervention**:

$$w \leftarrow w - \eta(g_L + \lambda_u M(g_L, g_U))$$

- Gradient operator *M* compares the alignment of g_U with g_L :
  - If aligned, retain or partially retain g_U
  - If misaligned, suppress shortcut-bearing components in g_U
  - **Three variants**: hard-gating g_U based on cosine similarity, soft-gating g_U by alignment, or projecting the shortcut components out of g_U

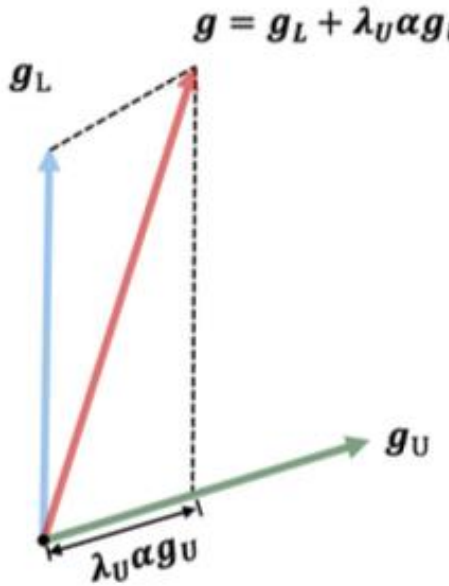
Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

Fix-A-Shortcut: Variants

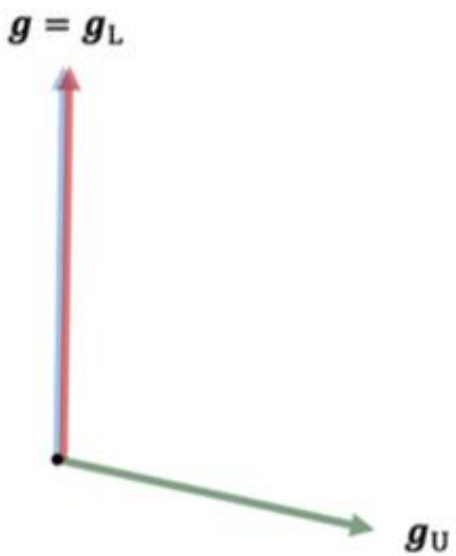


Fix-A-Shortcut

$\langle g_L, g_U \rangle \geq 0$



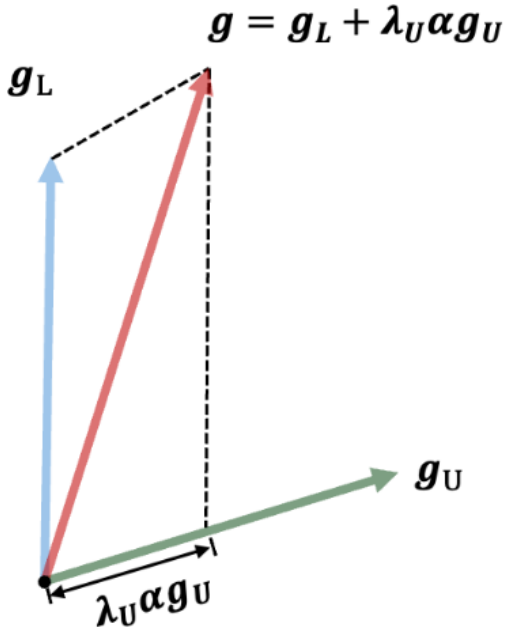
Fix-A-Shortcut Soft



$\langle g_L, g_U \rangle < 0$

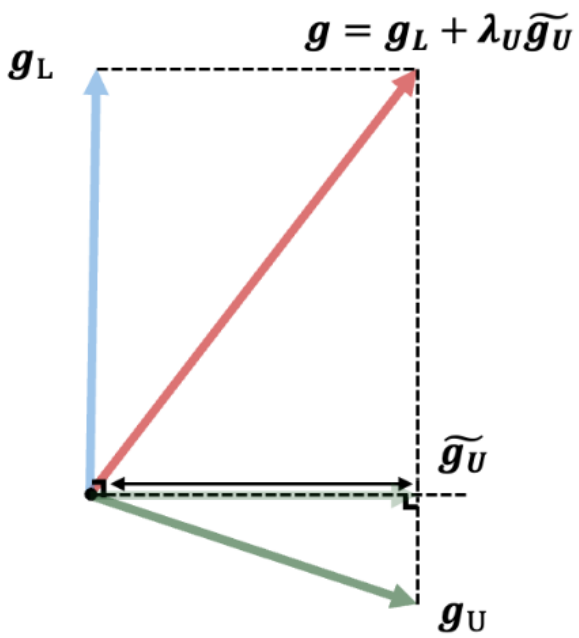
Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

Fix-A-Shortcut: Variants

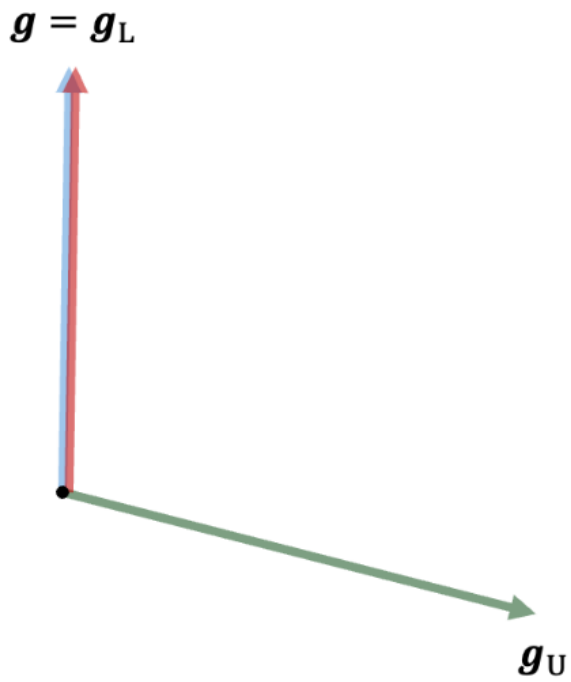


$\langle g_U, g_L \rangle \geq 0$

$|\phi| \leq 1$



$\langle g_U, g_L \rangle < 0$



$|\phi| > 1$

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

Main Theoretical Result: Why FAS works?

Unlabelled multimodal data amplifies shortcut reliance by at least a factor of λ

FAS **bounds shortcut amplification** in representation learning. Specifically, for any iteration t and any FAS operator we show that

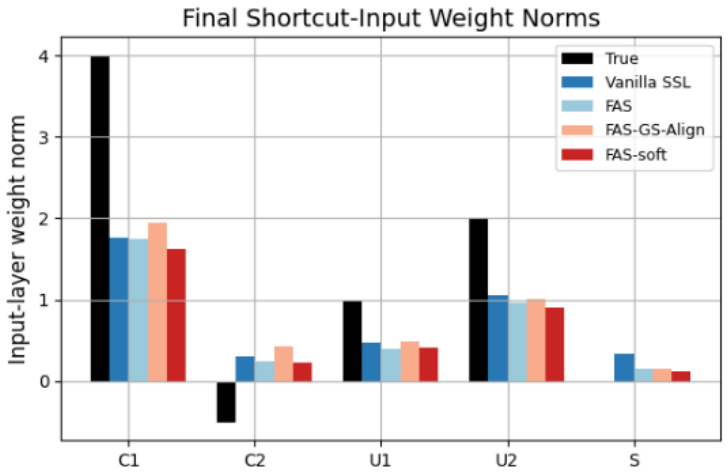
$$|\hat{\beta}^{\mathrm{FAS}}| \leq |\hat{\beta}^{\mathrm{SSL}}|$$

and

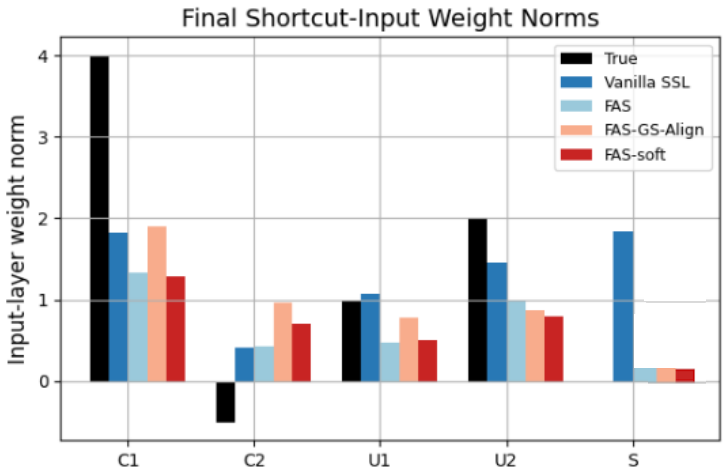
$$|\hat{\beta}^{\mathrm{FAS}}| < |\hat{\beta}^{\mathrm{SSL}}| \quad \text{whenever} \quad Pr(\langle g_L, g_U \rangle < 0) > 0$$

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

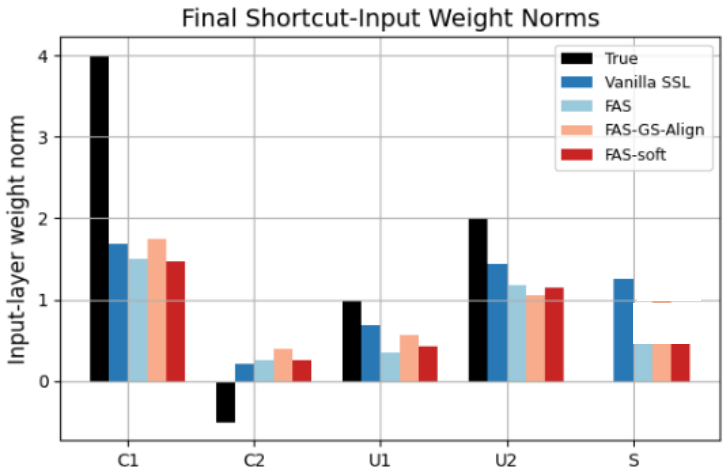
What happens empirically?



(a) Shortcut Weight (Y).



(b) Shortcut Weight with C.

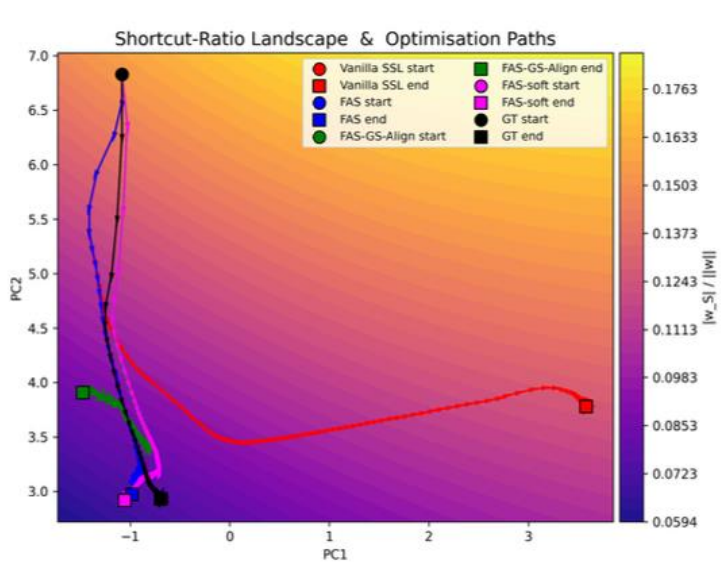


(c) Shortcut Weight (C & U).

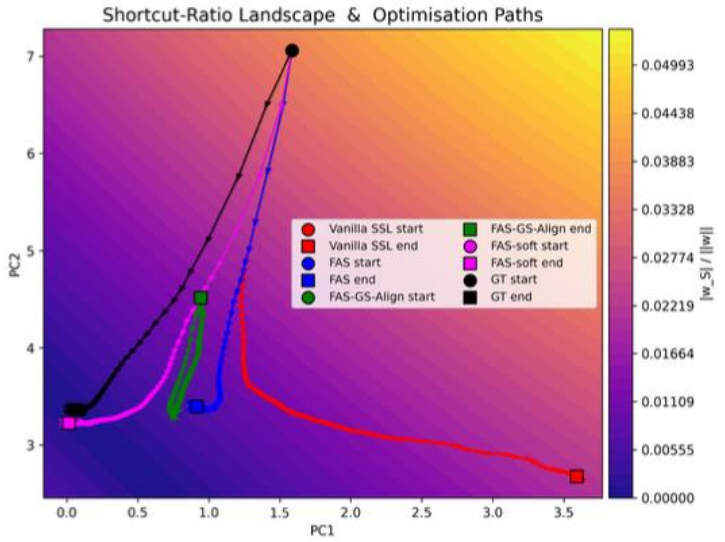
- **Gradient-based methods like FAS reduce shortcuts significantly compared to standard regularization approaches**
- **FAS methods work best when shortcuts correlate with labels or with known parts of the learnt representation.**

What happens empirically?

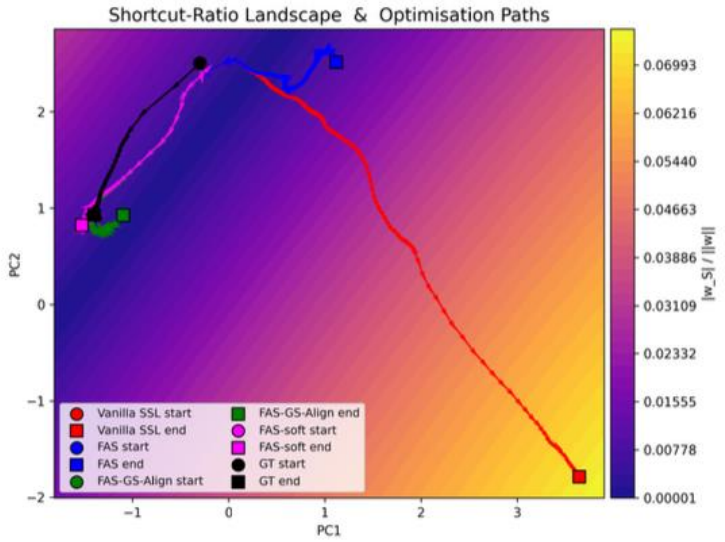
Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.



(a) Shortcut correlated with labels (Y).



(b) Shortcut entangled with causal features (C).

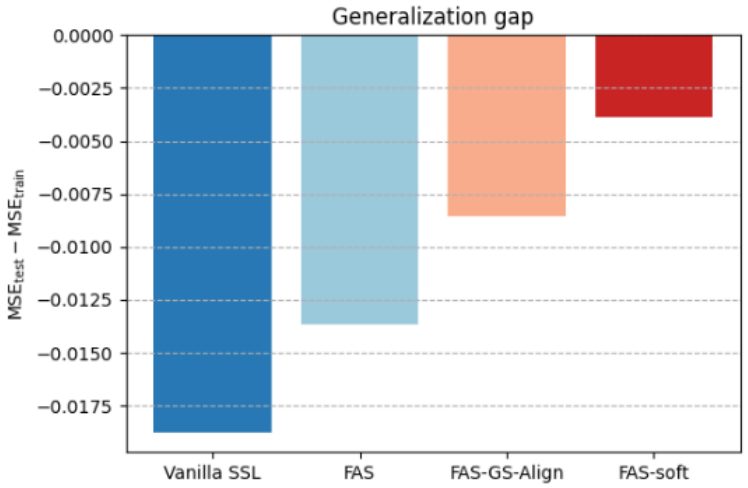


(c) Shortcut entangled with both causal and unknown features ($C\&U$).

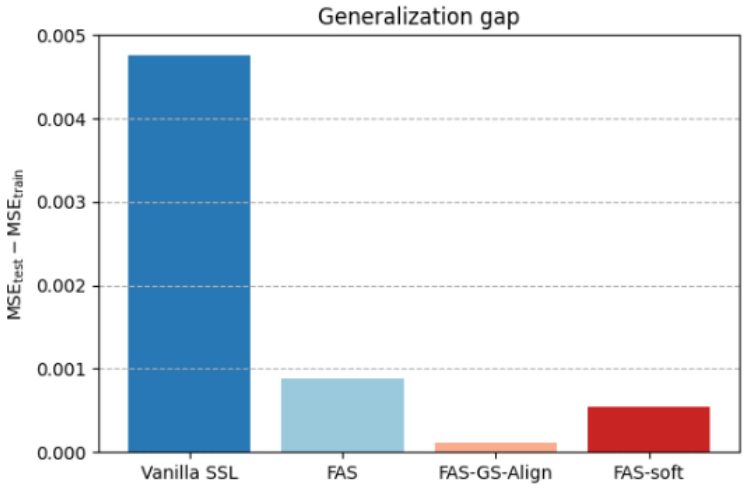
- **Gradient-based methods like FAS force optimization away from shortcut regions by changing optimization paths.**

What happens empirically?

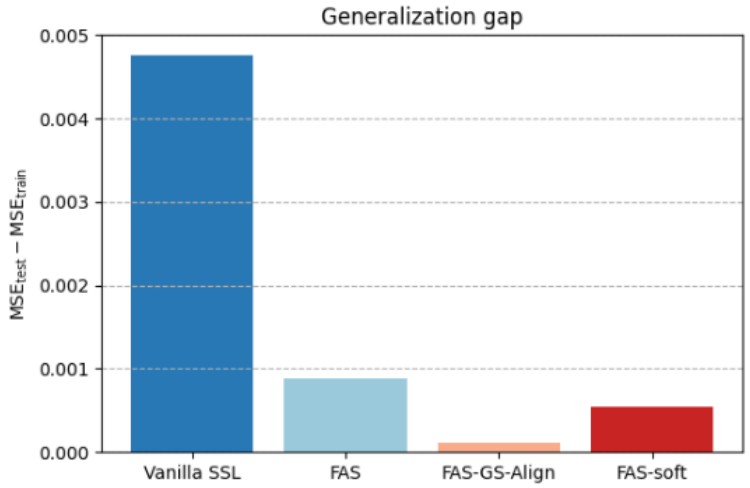
Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.



(a) Generalization gap with Y-correlated shortcuts.



(b) Generalization gap with shortcuts entangled with C.



(c) Generalization gap with shortcuts entangled with $C\&U$.

- **Gradient-based methods like FAS enable better downstream generalization, by reducing overfitting to shortcut features.**

What happens empirically?

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.

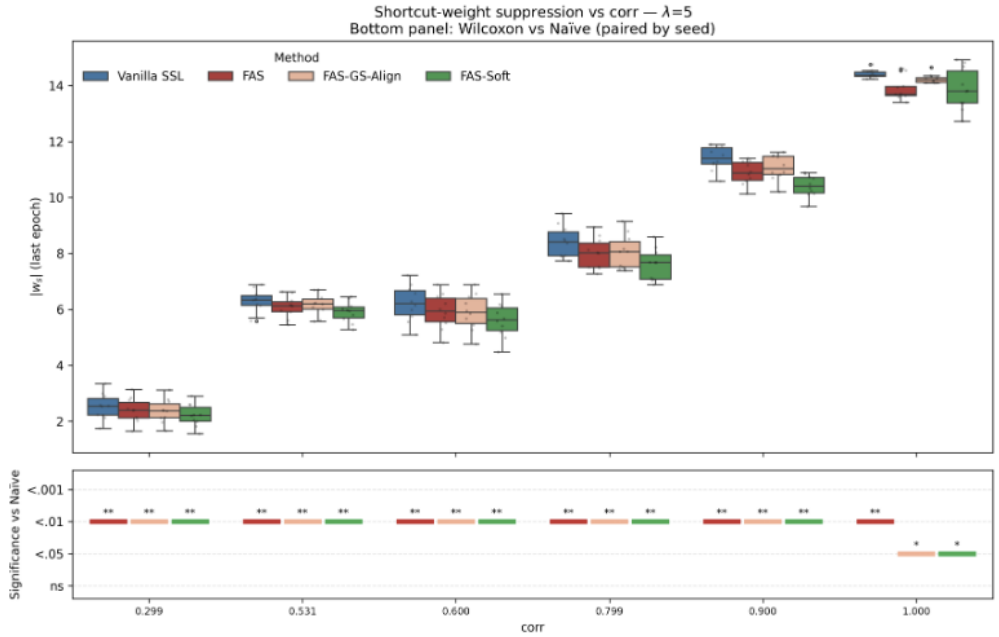


Figure 10: **MIMIC-CXR:** Boxplots showing statistical significance across shortcut–label correlation strengths.

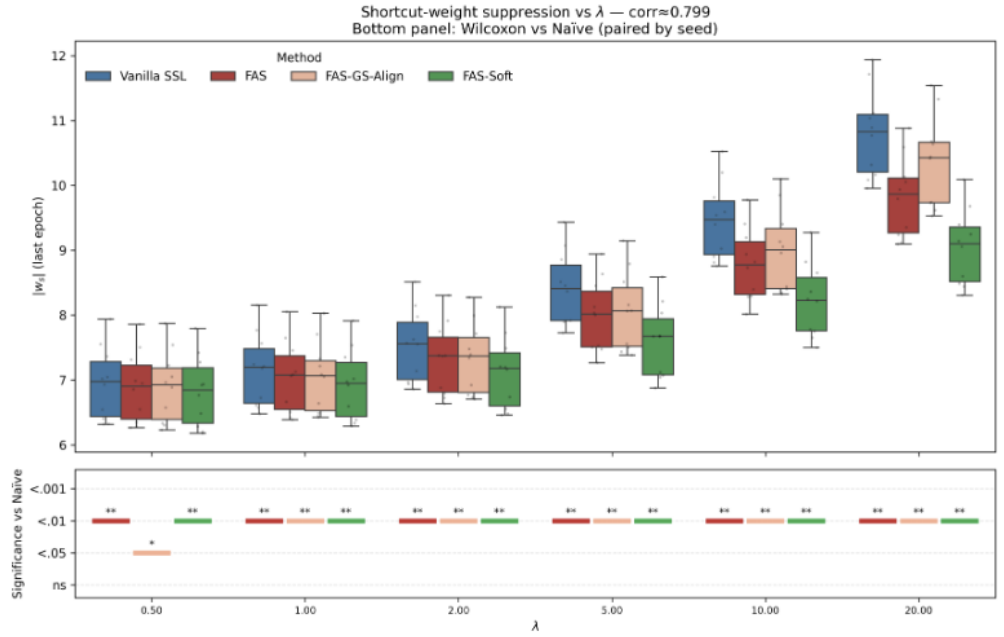
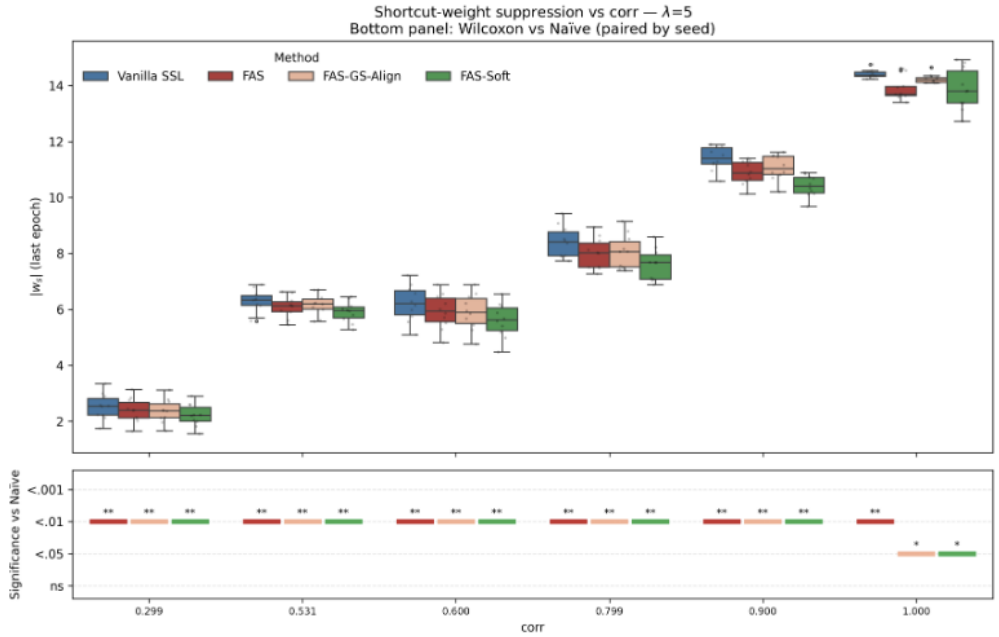
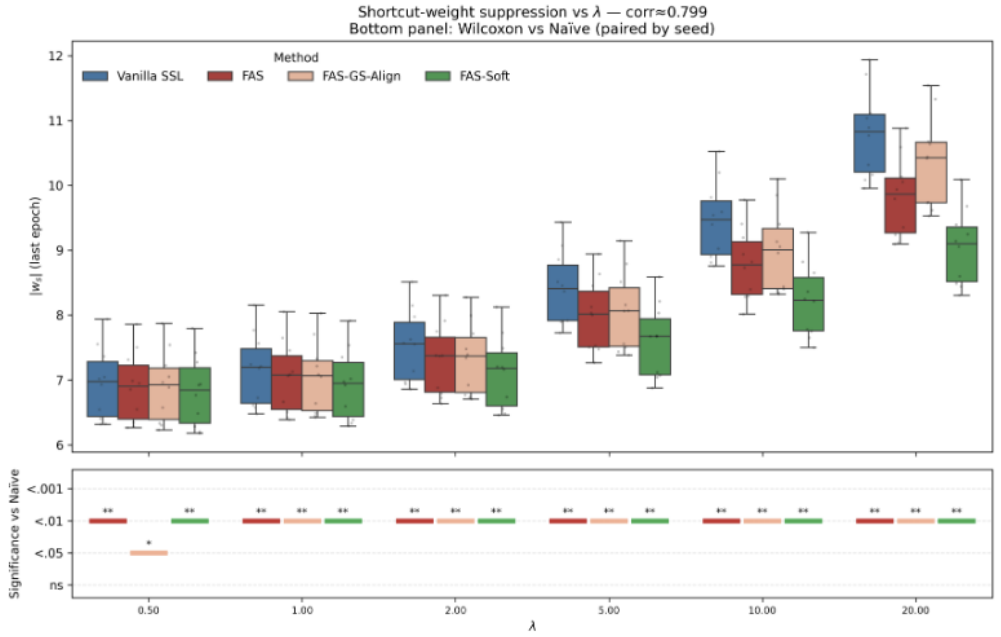


Figure 11: **MIMIC-CXR:** Boxplots showing statistical significance across λ values controlling unlabeled gradient weight.

What happens empirically?

Fix-A-Shortcut: Gradient-Based Shortcut Mitigation for Robust Representation Learning, Thanigaikumar, Hong and Parbhoo, preprint 2025.





- **Gradient-based methods like FAS are more effective than standard methods across real data scenarios with multiple shortcuts.**

Closing (or opening?)

Shortcuts are **prevalent across representations learnt from data**. These shortcuts, when left unresolved, undermine generalizability of models at test time.

We frequently turn to regularization as a solution, but don't spend enough time thinking about how it impacts learnt representations. This issue highlights the need for careful consideration of both the inductive biases of the model and the representativeness of the training data to mitigate the risk of learning misleading correlations.

Gradient-based methods can help under additional assumptions.

What other strategies exist for dealing with shortcuts?