Debugging Concept Bottleneck Models through Removal and Retraining

Eric Enouen, Sainyam Galhotra

Cornell University

CBMs

Bottleneck prediction through an interpretable concept layer

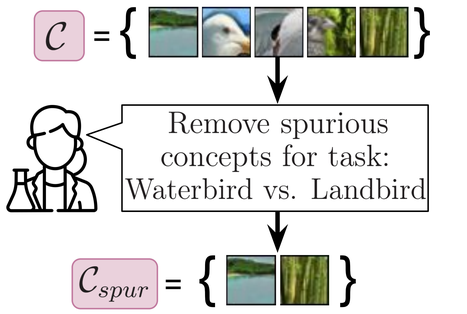
- Enables domain experts to analyze which concepts are being used for predictions
- Enable test-time interventions on mispredicted concepts

VLM-CBMs

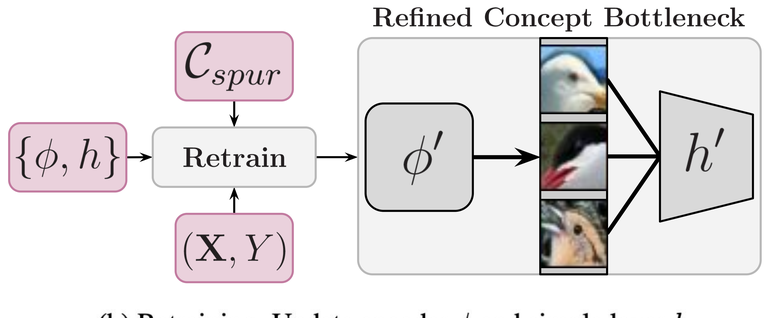
Remove the need for expensive concept annotations using foundation models

- Can introduce spurious concepts that fail under subpopulation shift!
- **How can we realign spurious concepts with expert reasoning?**

Removal + Retraining Framework

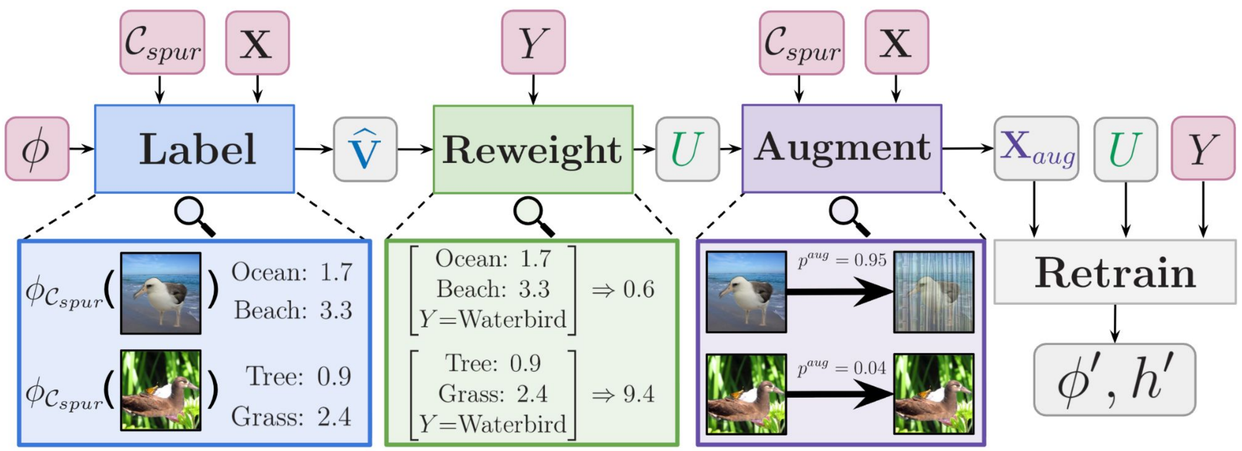


(a) Removal. User removes undesired concepts $\mathcal{C}_{spur}$.



(b) Retraining. Update encoder ϕ and simple layer h based on $\mathcal{C}_{spur}$.

CBDebug (Concept Bottleneck Debugger)



CBDebug Improves Robustness to Subpopulation Shift

Real	**PIP-Net**		**Post-hoc CBM**	
	Waterbirds	MetaShift	Waterbirds	MetaShift
Original	$71.9_{\pm 2.7}$	$52.4_{\pm 2.0}$	$25.8_{\pm 3.0}$	$84.5_{\pm 2.2}$
CBDebug	$\mathbf{79.4}_{\pm 4.3}$	$\mathbf{57.3}_{\pm 3.1}$	$\mathbf{51.9}_{\pm 16.2}$	$\mathbf{89.3}_{\pm 1.3}$

Automated	**Post-hoc CBM**			
	Waterbirds	MetaShift	CelebA	ISIC (AUROC)
Original	$25.8_{\pm 3.0}$	$84.5_{\pm 2.2}$	$8.7_{\pm 0.9}$	$39.3_{\pm 3.7}$
CBDebug	$\mathbf{58.3}_{\pm 6.0}$	$\mathbf{87.5}_{\pm 2.8}$	$\mathbf{51.3}_{\pm 3.9}$	$\mathbf{58.0}_{\pm 11.6}$

Learns Better Core Concepts

Class	Original	Retrain	CBDebug
Waterbird	hooked seabird beak	beach	duck-like body
	sea	gull-like body	hooked seabird beak
	harbor	water	orange wings
	lake	hooked seabird beak	orange eyes
	gull-like body	duck-like body	orange nape
Landbird	olive crown	olive upper tail	olive upper tail
	tree-clinging-like body	bamboo	iridescent bill
	forest	green primary color	blue upper tail
	olive upper tail	tree-clinging-like body	olive crown
	tree	olive breast	hawk-like body

Conclusions

CBDebug improves worst-group accuracy

- **across datasets** (Waterbirds, MetaShift, CelebA, ISIC)
- **and feedback** (real users, automated w/LLM)

Can also outperform classic unsupervised group robustness strategies (JTT, LfF)

Future Work

Robustness is fundamentally sociotechnical

- Want model to reason how a human would reason.
- Interpretability as a tool for improving robustness

How can we use interpretability to make progress on sociotechnical problems?

This research was supported by a gift to the LinkedIn–Cornell Bowers Strategic Partnership, and a grant from Infosys.