[Re] On the Reproducibility of Post-Hoc Concept Bottleneck Models

Nesta Midavaine, Diego Cánez Ildefonso,
Gregory Hok Tjoan Go, Ioana Simion

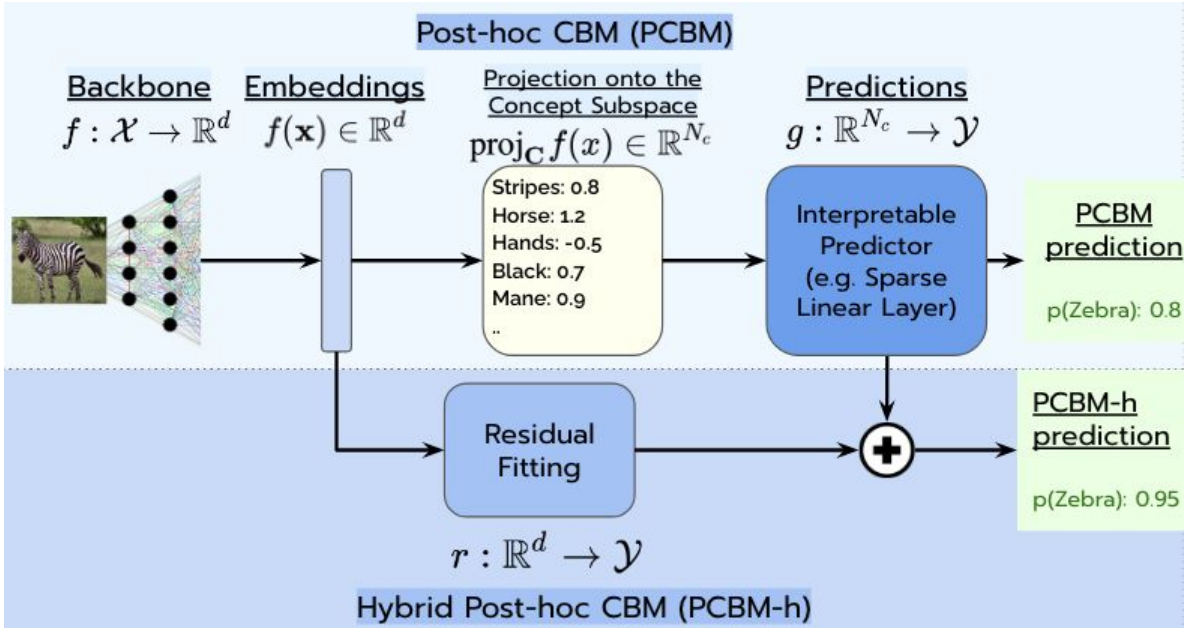
NEURAL INFORMATION PROCESSING SYSTEMS

PCBMs

Given model (backbone + classifier):

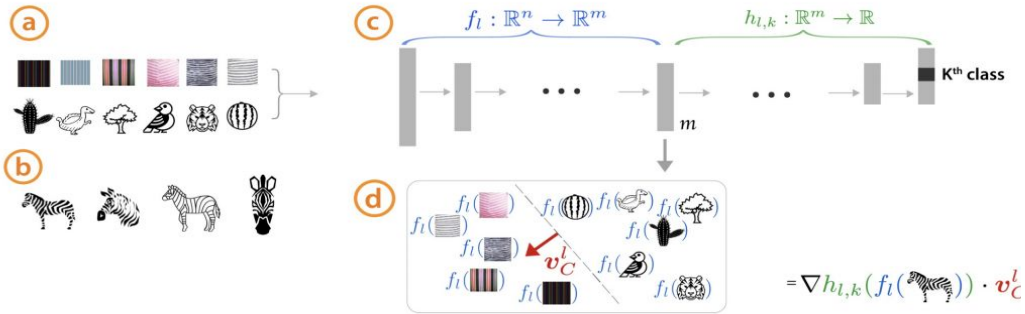
a. Drop classifier
b. Add Concept Projection Layer
c. Re-train new classifier

If concept vectors are not sufficiently expressive, add residual layer!

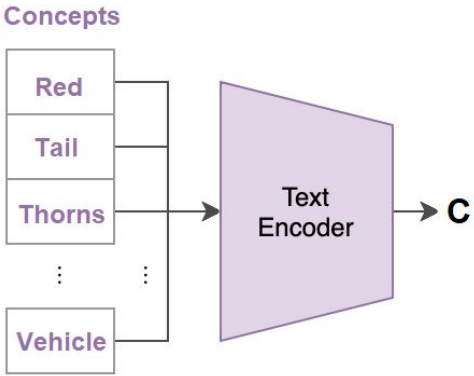


Computing the Concept Matrix: CAVs & CLIP

CAVs: Normal vectors of the linear boundaries between positive and negative examples for each concept.



CLIP: Get concept vectors directly from CLIP text encoder.



Main Claims

Claim 1:

PCBMs do not compromise the original model's performance

✔️

Claim 2:

PCBMs do not require labelled concept datasets

✔️

Claim 3:

PCBMs allow for global model editing

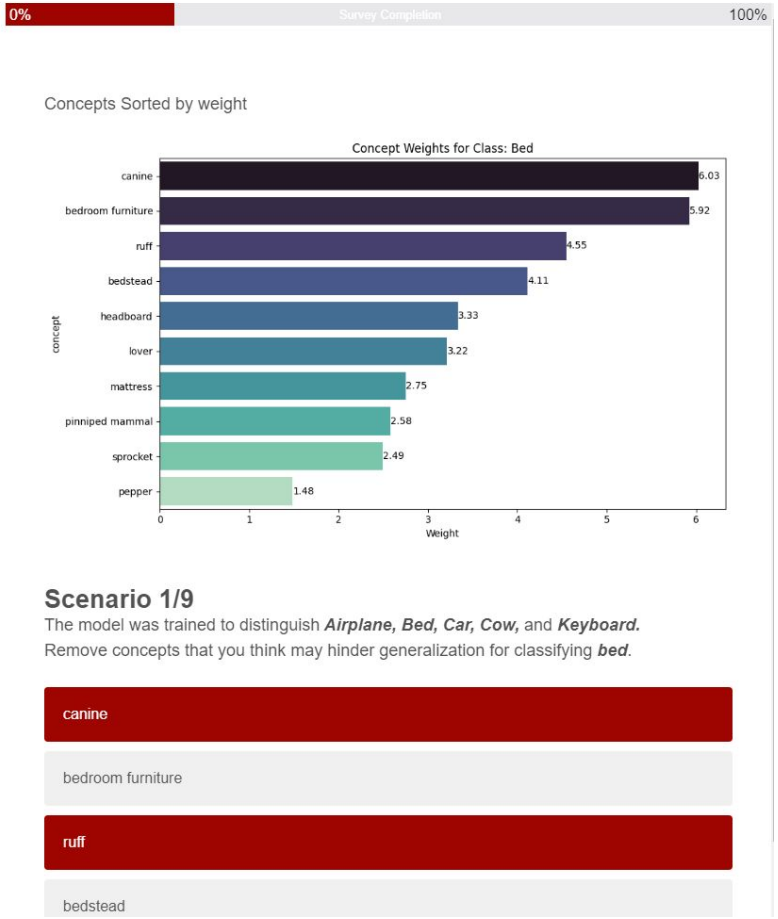
❌

Claim 3

1. Tackle spurious concept
2. Using Metashift Dataset
3. Also, a survey

Train Set

Test Set



Metashifts

	Unedited	Prune	Prune + Normalize
PCBM Accuracy	0.8637	0.8639	0.8638
PCBM Edit Gain	-	0.0002	0.00001

User Study

	Unedited	User Prune	Random prune	Greedy Prune
PCBM Accuracy	0.822	0.733	0.741	0.747
PCBM Edit Gain	-	-0.089	-0.081	-0.075

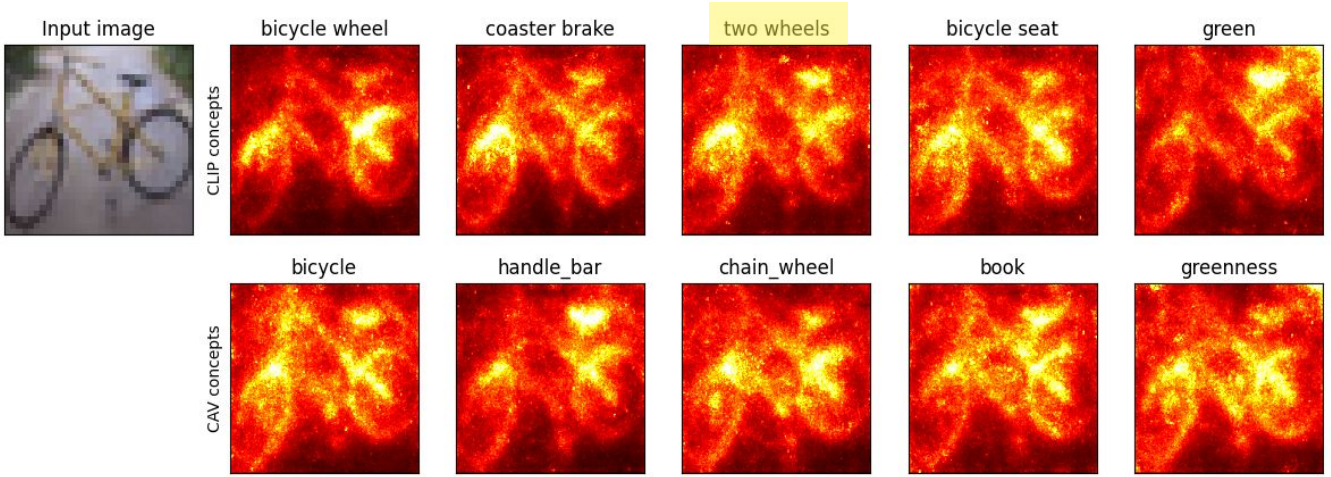
Pruning time

Users	Greedy
36.6s	52.49s

Claim 3

Limitations in Interpretability

- Concept saliency map highlights <u>entire</u> object instead of specific part



Additional Generalization

1. Evaluate PCBMs on new modality (audio) → AudioCLIP
2. Textual concepts from:
   a. Audioset Ontology
   b. UrbanSound8K
   c. ESC-50

3. AudioCLIP text encoder

	UrbanSound8K	ESC-50
Original	0.613	0.670
PCBM + CLIP	0.558	0.280
PCBM-h + CLIP	0.603	0.280
PCBM	0.411	0.400
PCBM-h	0.462	0.410

Discussion

1. Results generally reproducible
2. Extension supports main claim
3. Not without issues:
    - Missing details
    - Missing code

Image sources:
https://www.kpsol.com/wp-content/uploads/2017/09/speech-marks-2.jpg

Thank you!