Addressing Concept Mislabeling in Concept Bottleneck Models Through Preference Optimization

Emiliano Penaloza [1,2] Tianyue H. Zhang [1,2] Laurent Charlin* [1,3] Mateo Espinosa Zarlenga* [4]

[1]Mila [2]Université de Montréal [3]HEC Montréal [4]University of Cambridge

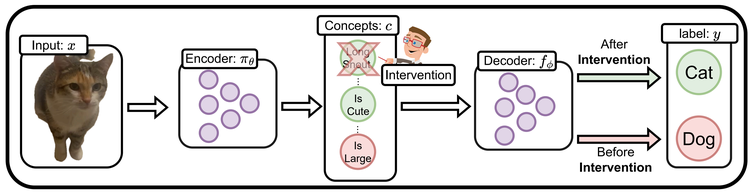
Paper in 30 Seconds

Rather than assuming concept label correctness, we assume **preference** over concept labels—this simple shift enhances concept accuracy, task performance, and intervention effectiveness in **Concept Bottleneck Models** under noisy and clean concept labels.

Concept Bottleneck Models

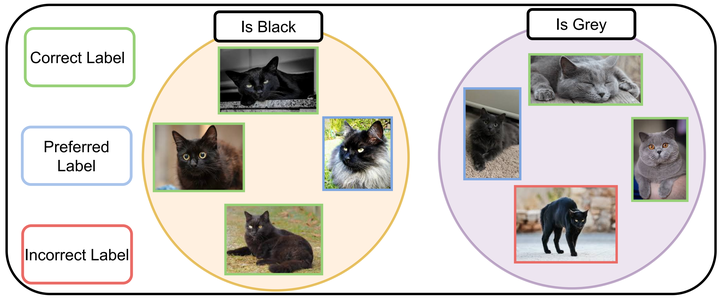
Concept Bottleneck Models (CBMs) are interpretable deep learning architectures that constrain model outputs using a concept loss on pre-defined concepts.

$$\mathcal{L}_{\text{CBM}} = \mathcal{L}_{\text{CE}}(y, g_\phi(c)) + \lambda \mathcal{L}_{\text{BCE}}(c, \pi_\theta(c|x)).$$

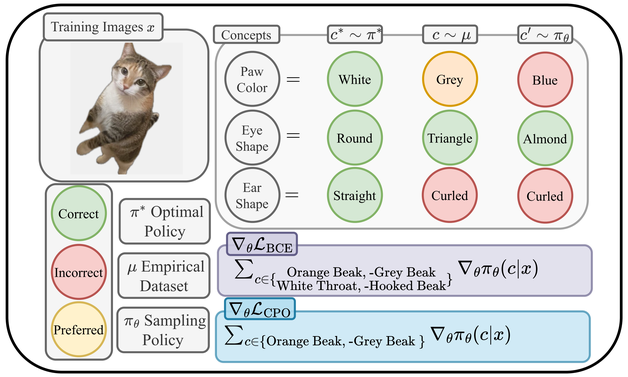


Noisy Concepts

Image datasets are known to be prone to be specifically prone to labeling noise. While not a huge problem for traditional models, this is drastically exacerbated in CBMs where there are potentially hundreds of labels per data sample making this a huge problem for these types of models in reality.



Concept Preference Optimization



CBMs rely on Binary Cross Entropy (BCE) for training which assumes label correctness—being an unrealistic assumption for most CBM datasets—so we design Concept Preference Optimization (CPO), an online algorithm inspired by Direct Preference Optimization that only assumes preference over concept pairs.

$$\mathcal{L}_{\text{CPO}} = -\mathbb{E}_{\substack{(x,c)\sim\mu \\ c'\sim\pi_\theta}} \left[\log \sigma \left(\log \frac{\pi_\theta(c|x)}{\pi_0(c|x)} - \log \frac{\pi_\theta(c'|x)}{\pi_0(c'|x)} \right) \right]$$

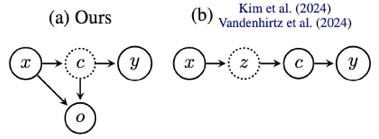
But why should preferences be more robust to noise?

In short, we show that under label noise, the gradient of $\mathcal{L}_{\text{CPO}}$ closer approximates its optimal gradient compared to $\mathcal{L}_{\text{BCE}}$.

$$\left\| \mathbb{E}_{(c^*,x)\sim d}[\nabla_\theta \mathcal{L}_{\text{CPO}}] - \mathbb{E}_{(c,x)\sim\mu}[\nabla_\theta \mathcal{L}_{\text{CPO}}] \right\|_2 \leq \left\| \mathbb{E}_{(c^*,x)\sim d}[\nabla_\theta \mathcal{L}_{\text{BCE}}] - \mathbb{E}_{(c,x)\sim\mu}[\nabla_\theta \mathcal{L}_{\text{BCE}}] \right\|_2$$

Posterior Approximation

We show that since $\mathcal{L}_{\text{CPO}}$ optimizes the maximum entropy RL objective, it is equivalent to doing posterior inference over the concepts c, thus yielding enhanced uncertainty estimates at no cost.



Non-Noisy Performance

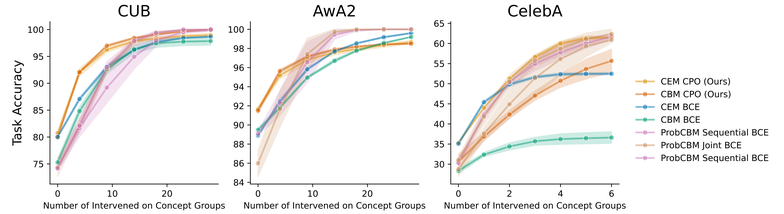
Base Perfomence

We find that even when there is no noise models $\mathcal{L}_{\text{CPO}}$ is able to outperform models trained with $\mathcal{L}_{\text{BCE}}$. Interestingly base CBMS trained with $\mathcal{L}_{\text{CPO}}$ are able to match/outperform more parametrized models such as Concept Embedding Models (CEMs).

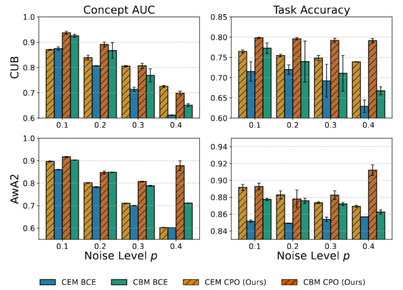
	CUB		AwA2		CelebA	
	Task Accuracy	Concept AUC	Task Accuracy	Concept AUC	Task Accuracy	Concept AUC
ProbCBM Sequential	0.742 ± 0.004	0.900± 0.007	0.891 ± 0.003	0.960 ± 0.003	0.302 ± 0.008	**0.878 ± 0.006**
CBM BCE	0.753 ± 0.009	0.937 ± 0.001	0.900 ± 0.008	0.959 ± 0.003	0.283 ± 0.007	0.873 ± 0.002
CBM CPO (Ours)	0.800 ± 0.003	**0.952 ± 0.001**	0.915 ± 0.004	**0.971 ± 0.001**	0.310 ± 0.009	0.857 ± 0.003
CEM BCE	0.800 ± 0.003	0.946 ± 0.001	0.889 ± 0.001	0.953 ± 0.000	0.351 ± 0.006	0.875 ± 0.004
CEM CPO (Ours)	**0.807 ± 0.004**	0.931 ± 0.003	**0.917 ± 0.003**	0.965 ± 0.001	**0.352 ± 0.004**	0.853 ± 0.003

Interventions

We find that the uncertainty estimates yielded by $\mathcal{L}_{\text{CPO}}$ greatly aid in getting better performance when intervening based on model uncertainty (variance of the concept prediction).



Performance Under Noisy Concept Labels



We study noisy concepts by randomly flipping the concepts labels with probability $p \in \{0.1, 0.2, 0.3, 0.4\}$. See the for more structured noise where we noise according to related concepts and uncertainty of the labeler.

Link to Paper

emiliano.penaloza@mila.quebec