Overcoming Interpretability and Accuracy Trade-off in Medical Imaging

Ivaxi Sheth [12] Samira Ebrahimi Kahou [123]

[1]Mila [2]ÉTS Montréal [3]CIFAR AI Chair

Motivation

With neural networks applied to safety-critical applications, it has become increasingly important to understand the defining features of decision-making, especially in healthcare. Therefore, the need to uncover the black boxes to rational representational space of these neural networks is apparent.

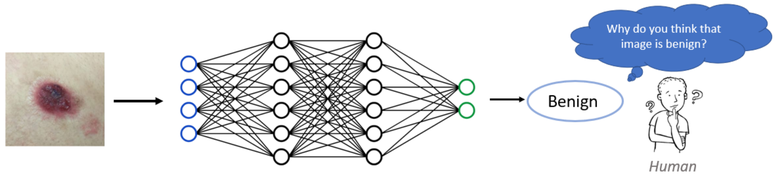

Figure 1. Human would like to reason a model's predictions of a black box model

What are Concept Based Models?

1. Concept Bottleneck Models (CBM) essentially map input images to such interpretable concepts which in turn predicts the label.
2. The intermediary concept prediction allows for the user to interact with the network.
3. This interaction is facilitated by test time interventions that allows an expert to "correct" wrongly predicted concepts, possibly improving downstream predictions.

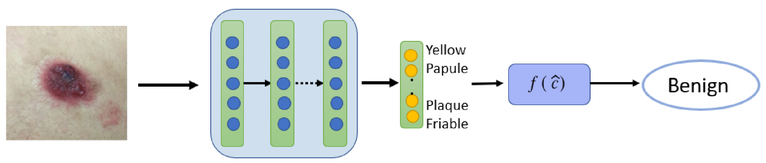

Figure 2. Typical Concept Bottleneck Model that predicts concepts before final prediction

Disadvantages of current CBMs

- CBM gives an explainable model at the expense of the lower accuracy of the model.
- This poses a trade-off between concept accuracy and task accuracy.
- CBMs have not been optimized for medical imaging applications where dense concept annotations are absent.

Our contributions

1. We propose a novel concept-based architecture, *coop-CBM* that overcomes the trade-off between interpretability and accuracy.
2. We make the first attempt to study the robustness of CBMs in realistic medical image settings where fine-grained concept annotations are absent.
3. Our model achieves state of art both concept and task accuracy.
4. We further evaluate the effect of coop-CBM on test-time interventions.

Coop-CBM

Our model, *coop-CBM*, is a hybrid multi-task model that predicts both labels and concept-based explanations.

- **Standard supervised learning** Models $\mathcal{M}$ are trained on a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{K}$ with K data samples. Standard models aim to predict the true distribution $p_{\mathcal{M}}(y|x)$ from an input x.

- **Supervised concept-based model** The dataset $\mathcal{D} = \{x_i, c_i, y_i\}_{i=1}^{K}$ is the input to concept-based model. The model has prediction at two levels, the first model $\mathcal{G}_{X\to C}$ maps the input image x to concepts c denoted by $p_{\mathcal{G}}(c|x)$, while the second model $\mathcal{F}_{C\to Y}$ maps the concepts c to the label y denoted by $p_{\mathcal{F}}(y|c)$.

- **Coop-CBM** To preserve the standard model's performance, our model, coop-CBM uses a supplementary predictor. Therefore inspired by the literature on multi-task learning, we introduce an additional predictor, $\mathcal{H}_{X\to Y}$ that predicts supplemental label. This additional stream is separate from the concept prediction pipeline. We hypothesize that this supplementary label prediction helps the concept prediction stream to recover model performance in the absence of fine-grained concept labels.

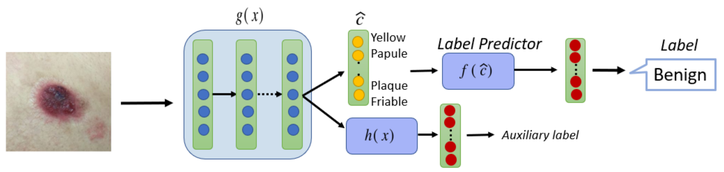

Figure 3. Coop-CBM model with an image sample from DDI dataset. In addition to predicting concepts in the bottleneck, our model also predicts the auxiliary label. The final label is predicted from concepts

Interventions During inference, Coop-CBM is particularly advantageous as they allow model editing based on human feedback. If a supervisor observes incorrect concepts related to a label, they can correct the output of $p_{\mathcal{G}}(c|x)$ which effectively changes, often improves, the downstream label prediction $p_{\mathcal{F}}(y|c)$.

Results

We evaluate on two classification datasets, TIL [?] and DDI [?, ?] to classify cancer tumors and skin diseases respectively. These two datasets are different in their concept representation, metadata for TIL includes non-image features such as age and gender along with clinical descriptor terms.
To evaluate the performance, here, we are concerned with the final prediction accuracy, i.e. performance of $p_{\mathcal{F}}(y|c)$.

Model type	TIL	DDI
Standard *[No concepts]*	51.1	83.4
CBM [1]	49.0	79.9
CEM [?]	51.3	83.9
CBM-AR [?]	49.5	80.6
Coop-CBM (ours)	**53.4**	**84.0**

Table 1. Accuracy of different on TIL and DDI dataset.

From Table 1, we notice our method has the most superior performance in comparison to the baselines on both TIL and DDI datasets. We observe that Concept Bottleneck Models [1] observe a big drop in performance in comparison to the Standard model that does not use concepts.
Therefore if the doctor observes an incorrect concept explanation during test time, they can intervene and alter the concepts often resulting in superior downstream performance.

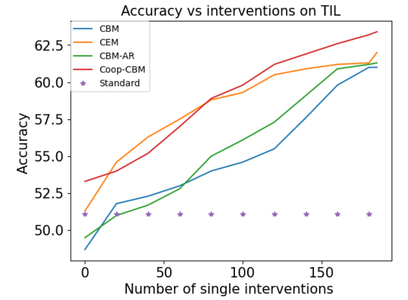

Figure 4.

To quantify the effectiveness of interventions, we compare the accuracy with increasing intervention by choosing concepts randomly and correcting them to ground truth. Figure ?? shows that coop-CBM is highly receptive to concept correction on TIL.

Conclusion

- In this work we propose *coop-CBM*, a multi-task-based explainable concept based model.
- The proposed model achieves state of art task accuracy performance.
- We overcome the interpretability and accuracy trade-off in medical imaging.
- In addition, we perform test-time interventions and observe that coop-CBM is the most receptive to interventions suggesting higher downstream accuracy.

References

[1] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ArXiv*, abs/2007.04612, 2020.

ivaxi-miteshkumar.sheth.1@ens.etsmtl.ca

Paper link