Supplementary: A Framework for Learning Ante-hoc Explainable Models via Concepts

Limitations & Broader Impact

The increasing need for explainability in the use of Deep Neural Network (DNN) models has in turn necessitated the design of ante-hoc explainable models that jointly learn to predict and explain. The limited efforts in this space such as SENN [1] and CBM [14] have their own set of limitations when used in practice. They either require concept-level supervision to train the model or need a significant number of additional parameters in the network, which prohibits their use in deeper models more commonly used in practice. Our framework aims to alleviate these issues. The proposed work addresses this need, and provides a framework for learning ante-hoc explainable models via concepts with significantly lesser additional parameters when compared to [1] (please see Sec 4 of main paper).

A limitation of our method is this increase in computational cost during training due to adding additional components such as the decoder network and the concept generator to an existing DL pipeline (although this is still better than baseline methods such as SENN). We keep additional components like the decoder network in the model despite this limitation since the decoder encourages the interpreter (or the concept generator) to generate meaningful explanations and faithfully capture semantics of an input image. While our model needs additional training time, its inference time, which matters in practice, isn't significantly higher than the backbone architecture itself. Our framework works by jointly learning to generate ante-hoc explanations via concepts and predict the label for the given image. Ante-hoc methods like ours can help understand a model's decision and gain intuition into its inner workings. In turn, this can help to improve the transparency and trustworthiness of Deep Learning models. In addition to the above, our framework also allows a user to intervene on learned concepts to understand a model's decisions. In turn, this could be a valuable tool, especially when models are deployed for sensitive or critical tasks.

Appendix

In this appendix, we provide details that we could not include in the main paper owing to space constraints, including:

- Comparison of the faithfulness of explanations generated by our technique against different post-hoc explainability methods such as LIME and Grad-CAM
- Predictive performance of our framework with different backbone architectures
- Interpretation of concepts in the unsupervised setting where no ground truth is known
- Details of the hyperparameters used in the experiments
- Comparison of concept-based explanations generated by our method with and without concept supervision, in cases where annotations of attributes are available
- Comparison of concept-based explanations generated by our framework and baselines methods such as SENN and CBM
- Comparison of concept-based explanations generated by our framework with different number of concepts.

A. Comparison with Post-hoc Methods

In the recent past, there has been a lot of interest in developing techniques that try to explain a model's prediction after training. While these methods are helpful, the separation of explanation from prediction is not ideal. Ideally, we would like the techniques which generate interpretations to explain the model's prediction faithfully. But in the case of post-hoc methods, when an explanation goes wrong, it is not trivial to understand if the explanation method is incorrect or if the model itself relied on spurious correlations to make a prediction. To illustrate that our framework generates more faithful explanations than post-hoc methods, we compare the *faithfulness* (predictive capacity of the generated concepts, i.e., from the output of $s_{\theta_{cce}}$) of well-known post-hoc explainability methods such as LIME [23] and Grad-CAM [27] with that of our framework. We generate explanations with these methods for every image and pass the modified image through the model. From Table 8, we see that our method outperforms other post-hoc explainability methods in terms of the faithfulness metric.

Dataset	Post-hoc Baselines		OURS	
	LIME	Grad-CAM	w/o sup	w sup
CIFAR10	34.23	76.91	**90.86**	NA
ImageNet	29.39	47.48	**59.73**	NA
AwA2	46.21	75.00	79.29	**83.30**
CUB-200	27.68	43.51	61.49	**62.59**

Table 8. Comparison of *faithfulness* (predictive capacity of the generated concepts) our method against post-hoc explainability methods on CIFAR10, ImageNet, AwA2 and CUB-200 data sets. (w/o sup = without supervision; w sup = with supervision)

B. Plug and Play: Integrating with Other Backbone Architectures

One of the advantages of our method is that we can plug ante-hoc interpretability to different existing DNN backbone architectures. To illustrate this, we incorporate ante-hoc interpretability for 4 different popular backbones architectures i.e. ResNet34, ResNet50, EfficientNet-B0 and DenseNet-121 in Table 9. We consider all four datasets i.e.

CIFAR10, ImageNet, AwA2 and CUB-200 for our experiments to show these results. We find that a stronger backbone (such as EfficientNet-B0 and DenseNet-121) helps improve the performance of our framework.

Architecture	CIFAR10	ImageNet	AwA2	CUB-200
ResNet34	91.82	65.55	85.88	65.69
ResNet50	92.04	66.12	86.11	65.98
EfficientNet-B0	91.79	66.58	85.95	66.09
DenseNet-121	92.85	65.91	86.73	66.03

Table 9. Comparison of *accuracy* (in %) on CIFAR10, ImageNet, AwA2 and CUB-200 datasets using different backbone architectures as concept (or base) encoder. All numbers with AwA2 and CUB-200 are generated with concept supervision.

C. Interpret Concepts in Unsupervised Setting

The complexity of concept interpretation in the unsupervised setting depends on dataset complexity i.e. number of classes and number of concepts used during training. For e.g., number of concepts difficult to interpret for ImageNet are more than for CIFAR10. This is also seen in Figs 5 & 4 in main paper, as number of homogeneous properties captured by ImageNet concepts are more than CIFAR10 concepts, and hence more difficult to interpret.

We conducted a human study by selecting 5 CIFAR10 concepts (Fig 5 left) and asked 10 users about the concepts that are closest to 5 images activated by each concept (concept activations). Both classes and the concepts (pointy, gray, round, black, white) were provided as options to the user. Users agreed on 90%, 100%, 90%,90% and 100% images for the 5 concepts respectively.

We conducted the same study for ImageNet and asked the same 10 users (as above) to choose closest concepts activated by 10 ImageNet concepts (Fig 4). As the users aren't aware of the fine-grained ImageNet class labels, we considered crude labels here as options as concepts i.e. dog, swan, cucumber, watch, pen, wolf, cat, rugby, drawer and cheetah. We also added abstract concepts (white, blue, green, circular, sharp, wolf face, cat face, game, rectangular, stripes) as added options for the users. Here, they agreed on 90%, 70%, 80%, 60%, 100%, 70%, 50%, 80%, 70%, 70% and 60% images for the 10 concepts respectively. As ImageNet has way more classes with many images and more complex concepts compared to CIFAR10, the user agreements are lower than above (for CIFAR10), as expected. While we observe varying numbers for different datasets based on complexity, evidently, there is overall high agreement on concept interpretations generated by our model.

D. Hyperparameter details

We provide details for coefficients of loss terms (presented in Sec 3) here for all the datasets i.e. CIFAR10, ImageNet, AwA2 and CUB-200. We have used similar set of coefficients for CIFAR10 and ImageNet. Also, similar set of coefficients were used for AwA2 and CUB-200 datasets. All these results are shown in Table 10. Please note that, for CIFAR10 and ImageNet, γ is 0 as there is no concept supervision, and μ can be 0 based on use of self-supervision on concepts. On the other hand, μ is 0 for AwA2 and CUB-200, as self-supervision is not required due to presence of ground truth concepts.

Coefficients	α	β	γ	μ
CIFAR10 / ImageNet	0.0001	0.1	0	0/0.1
AwA2 / CUB-200	0.0001	0.1	1	0

Table 10. Hyperparameter details i.e. values of α, β, γ and μ coefficients of our experiment on CIFAR10, ImageNet, AwA2 and CUB-200 datasets.

E. Qualitative Evaluation

This section presents additional qualitative studies on the quality of concepts and an ablation study on change in number of concepts.

E.1. Quality of Concepts

Concepts: *With and Without Supervision*: In this section, we observe that the concepts discovered by our framework correspond to some of the ground truth attributes even when trained without concept supervision. We considered AwA2 and CUB-200 for this study, as these datasets have ground-truth concepts for every image. We generated a set of images that maximally activate each concept for both types of models, i.e., with and without concept supervision, and present such results for some of the concepts in Figures 7 and 8 for AwA2 and CUB-200 respectively. In these figures, five concepts on the left are generated from the model trained without concept supervision. The rest of the concepts were generated by the model trained with concept supervision. For example, the images for ψ^5 (from the model without concept supervision) are visually similar to the images for the LONGNECK concept (from the model with concept supervision) from Figure 7. Also, the images for ψ^3 (from the model without concept supervision) are visually similar to the images for HAS_WING_COLOR::RED concept (from the model with concept supervision) from Figure 8.

Concepts: *Ours vs CBM*: Apart from the quantitative results presented in the main paper, which compare our method with the baseline method that considers concept supervision, i.e., Concept Bottlencek Methods (CBM), we herein show sample qualitative results. We generated images that maximally activate every concept learned by our model as well as CBM and show in Figures 9 and 10 for AwA2 and CUB-200 respectively. While all the concepts are meaningful visually, we can see a better selection of representative images for the concepts generated by our model.

For example, images activated for the FIELDS concept for our model represent the concept better than the images activated for CBM from fig.9. Similarly, images activated for HAS_BILL_SHAPE::SPATULATE concept for our model represent the concept better than the images activated for CBM from Figure 10.

Concepts: ***Ours vs SENN***: We present more qualitative results to compare the concepts generated by our method (without concept supervision) and SENN for CIFAR10 and ImageNet datasets. Figures 11 and 12 represent results with CIFAR10 for our method (without concept supervision) and SENN respectively. We observe that the concepts captured by SENN tend to repeat more and are less diverse than those generated by our method, thus leaving out some important aspects of the dataset. Similar issues are observed from the concept activations generated by SENN for ImageNet in Figure 13. For example, many concepts capture round-shaped objects or objects with a round head and miss some important hidden concepts in the dataset (for concept-based explanations generated by our method on ImageNet, please see Figure 4).

E.2. Ablation Study: Number of Concepts

While the number of concepts is available apriori for the datasets with ground truth concepts, it's not known beforehand for other datasets like CIFAR10. Hence it is a choice (or hyperparameter) left to the user. To understand the impact of the number of concepts on the performance of our framework, we experimented with different number of concepts for CIFAR10 (i.e., 5 and 15) and achieve 91.51% and 91.58% accuracies, respectively. These numbers are very close to the model's accuracy with ten concepts (i.e., 91.68%). We further analyzed the concepts generated by these two models (with 5 and 15 concepts) for more insights. We present the maximally activated images for every concept in Figures 14 and 15 for our models with 5 and 15 concepts respectively. It is evident that the model with 5 concepts (from Figure 14) is not able to capture all artifacts of the CIFAR10 dataset, whereas the model with 15 concepts (Figure 15) captures the most number of dataset artifacts (compared to models with 5 and 10 concepts), but it has repetitions of concepts. For example, the concept representing "deer" is not captured by any model except the model with 15 concepts. Exploring adaptive number of concepts and enforcing concept exclusivity while training could be interesting directions of future extensions of our work.

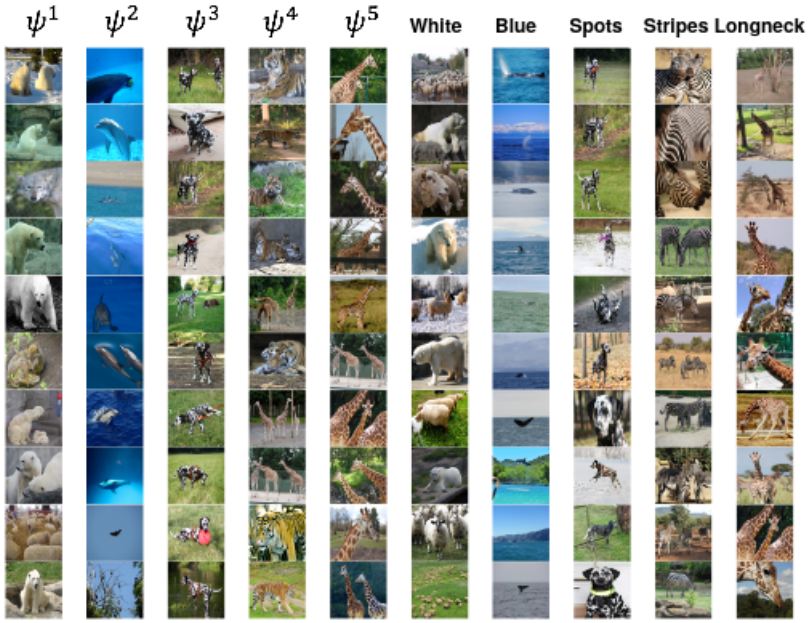

Figure 7. A subset of 5 concepts learned by our framework on AwA2 each with (right) and without (left) concept supervision. Please note that some of the concepts learned without concept supervision capture similar concepts that are learned with concept supervision.

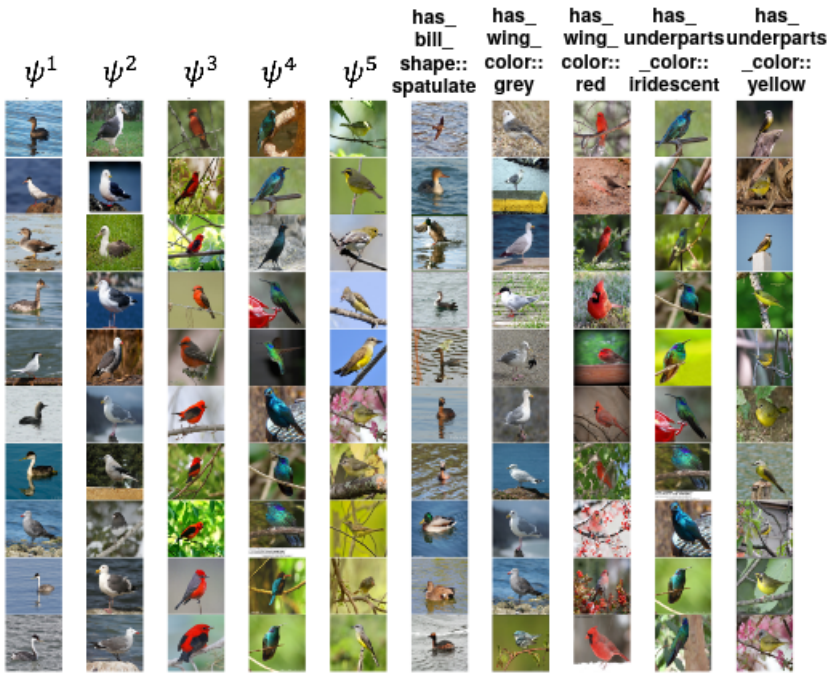

Figure 8. A subset of 5 concepts learned by our framework on CUB-200 each with (right) and without (left) concept supervision. We observe that some of the concepts learned without concept supervision capture similar concepts that are learned with concept supervision.

Figure 9. A subset of 5 concepts learned by our framework on AwA2 each with CBM (left) and our method (right). Here we consider our method with concept supervision for fair comparison with CBM.

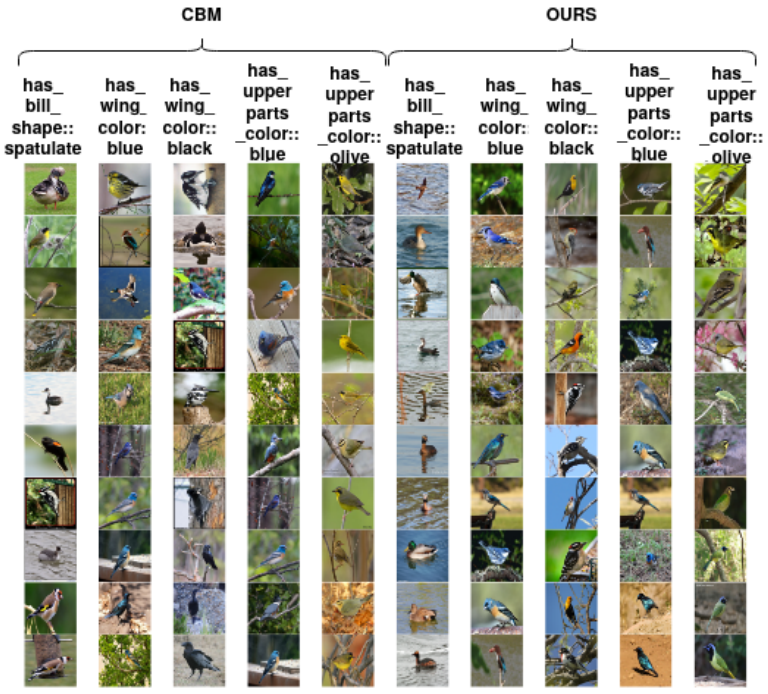

Figure 10. A subset of 5 concepts learned by our framework on CUB-200 each with CBM (left) and our method (right). Here we consider our method with concept supervision for fair comparison with CBM.

Figure 11. Concept activations (i.e. images that maximally activate each concept) learned by our framework on CIFAR10. It can be seen that each concept captures a certain set of homogeneous properties corresponding to a class. For instance, ψ^1 is mostly activated for images from cat class and the same for ψ^6 happens for images from frog class.

Figure 12. Concept activations (i.e. images that maximally activate each concept) learned by SENN on CIFAR10. It can be seen that some of the concepts are not able to capture a certain set of homogeneous properties corresponding to a class. For instance, ψ^5 is mostly activated for images from aeroplane class along with one image from horse class. Also, the maximally activated images for ψ^5 are from frog and dog classes.

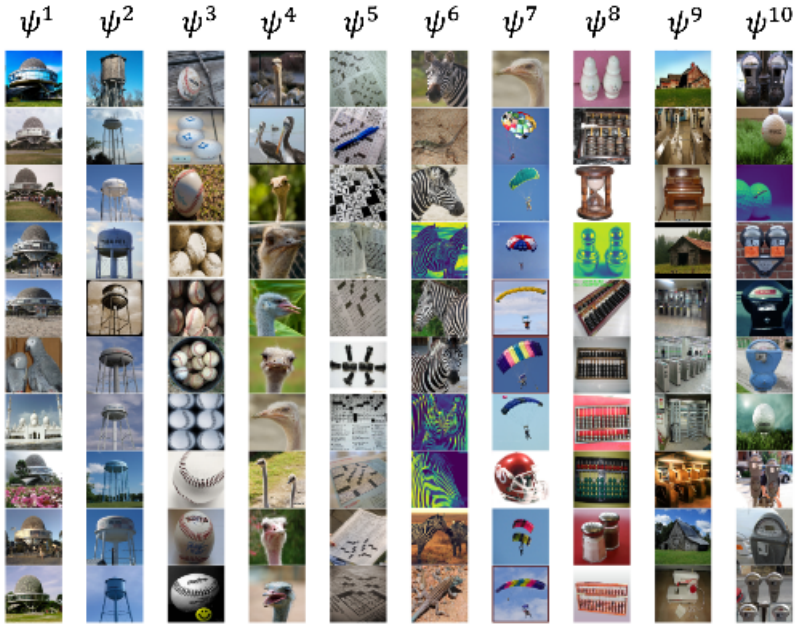

Figure 13. Concept activations (i.e. images that maximally activate each concept) learned by SENN on ImageNet. It can be seen that some concepts are not able to capture a certain set of homogeneous properties corresponding to one or many classes. For instance, ψ^8 is mostly activated for images which are visually not of similar structure. Also, most of the concepts capture a few of the properties, thus miss out on other important artifacts. For instance, ψ^1, ψ^2, ψ^3 and ψ^{10} capture round shaped objects or objects with round shaped head.

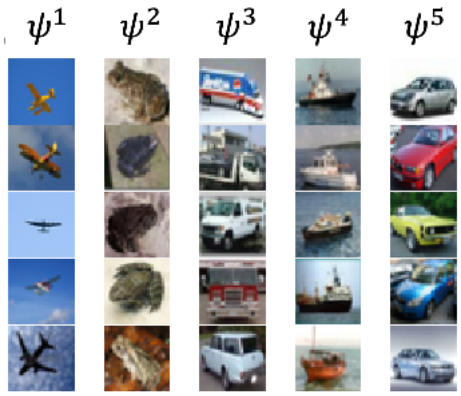

Figure 14. Concept activations (i.e. images that maximally activate each concept) learned by our framework (with 5 concepts) on CIFAR10. It can be seen that each concept captures a certain set of homogeneous properties corresponding to a class. For instance, ψ^1 is mostly activated for images from aeroplane class and ψ^2 is mostly activated for images from frog class. It is also clear that all the important concepts are not captured by this model due to less number of concepts (i.e. 5 concepts).

Figure 15. Concept activations (i.e. images that maximally activate each concept) learned by our framework (with 15 concepts) on CIFAR10. It can be seen that each concept captures a certain set of homogeneous properties corresponding to a class. For instance, ψ^7 is mostly activated for images from bird class and ψ^9 is mostly activated for images from cat class. Also some of the concepts are repeated by this model due to more number of concepts (i.e. 15 concepts). For example, ψ^2, ψ^4 are both activated for images from horse class, ψ^3, ψ^6 are both activated for images from truck class and ψ^8, ψ^{11} are both activated for images from automobile class.