A Framework for Learning Ante-hoc Explainable Models via Concepts

Anirban Sarkar
IIT Hyderabad
Telangana, India
cs16resch11006@iith.ac.in

Deepak Vijaykeerthy
IBM Research
Bangalore, India
deepakvij@in.ibm.com

Anindya Sarkar
IIT Hyderabad
Telangana, India
anindyasarkar.ece@gmail.com

Vineeth N Balasubramanian
IIT Hyderabad
Telangana, India
vineethnb@iith.ac.in

Abstract

Self-explaining deep models are designed to learn the latent concept-based explanations implicitly during training, which eliminates the requirement of any post-hoc explanation generation technique. In this work, we propose one such model that appends an explanation generation module on top of any basic network and jointly trains the whole module that shows high predictive performance and generates meaningful explanations in terms of concepts. Our training strategy is suitable for unsupervised concept learning with much lesser parameter space requirements compared to baseline methods. Our proposed model also has provision for leveraging self-supervision on concepts to extract better explanations. However, with full concept supervision, we achieve the best predictive performance compared to recently proposed concept-based explainable models. We report both qualitative and quantitative results with our method, which shows better performance than recently proposed concept-based explainability methods. We reported exhaustive results with two datasets without ground truth concepts, i.e., CIFAR10, ImageNet, and two datasets with ground truth concepts, i.e., AwA2, CUB-200, to show the effectiveness of our method for both cases. To the best of our knowledge, we are the first ante-hoc explanation generation method to show results with a large-scale dataset such as ImageNet.

1. Introduction

Recent years have seen an exponentially increasing interest in explainability of decisions of Deep Neural Network (DNN) models across domains including biometrics, healthcare, autonomous navigation and many more. Existing efforts in computer vision including occlusion-based, gradient-based and Shapley value-based efforts largely perform post hoc analysis [23, 27, 36], of an already trained model to identify what a DNN model looked at in an input image while making a prediction. While this is useful, the separation of explanation from prediction is not ideal. 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. This has paved the need for ante hoc methods that jointly learn to explain and predict, and thus learn inherently interpretable models.

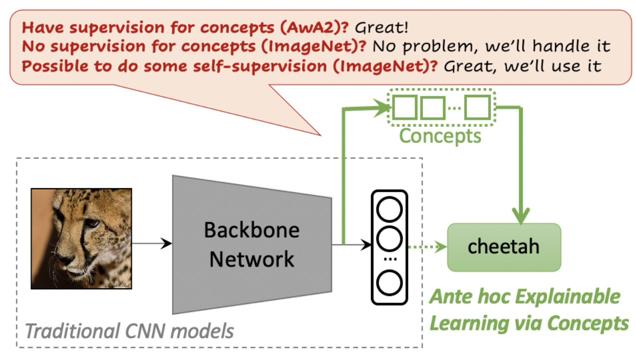

Figure 1. Illustration of the proposed framework. Our framework offers a way to train models that can not only predict but also explain their predictions. It can be easily integrated with existing backbone networks. Compared to existing techniques, it provides the flexibility to incorporate different forms of supervision (including weaker forms of supervision like *self-supervision*) whenever available or feasible.

Efforts on envisioning interpretable learning models by Rudin [25] and Lipton [18] have stressed on the importance of implicitly interpretable methods over post hoc explanations in elaborate terms. In a more recent exposition, Rudin et al [26] identified ten challenges of interpretable machine learning, which also highlighted the need for placing constraints into models to learn with better interpretability during training itself. The last few years have seen a few efforts in ante hoc methods that explain through concepts, which are learned during the training of the DNN itself such as Self-Explaining Neural Networks [1], Concept Bottleneck Models [14], Concept-based Model Extraction [12], and Concept Whitening [2]. Learning concepts during training provides a natural pathway for ante hoc explanations that are global (concepts that are most activated on a dataset or a class) or local (concepts that are most activated for prediction on given input image). Existing methods however either require concept-level supervision to train the model [14], or require a significant number of additional parameters in the network [1], which prohibits their use in deeper models more commonly used in practice.

In this work, we propose a new method towards learning ante-hoc explanations via concepts, that: (i) can be added easily to existing backbone classification architectures with minimal additional parameters; (ii) can provide explanations for model decisions in terms of concepts for an individual input image or for groups of images; and (iii) can work with different levels of supervision, including *no concept-level supervision* at all. This is achieved by an architectural modification added to a backbone network along with additional loss terms that allow such ante hoc learning. Importantly, we show that our framework allows learning of concepts with no supervision, self-supervision as well as full supervision at a concept-level. An overview of our proposed model is shown in fig.1.

Our key contributions in this work can be summarized as follows:

- We propose a simple and effective method that jointly learns to predict and explain (through concepts) in an ante hoc manner (i.e. learning to explain during training itself, as opposed to post hoc explainability methods popularly used today).
- Our method can learn to explain through concepts with different levels of supervision: (i) with no concept-level supervision; (ii) through weak supervision (self-supervised learning of concepts); as well as (iii) with concept-level supervision.
- We perform a comprehensive suite of experiments to study accuracy and explainability of our method on multiple benchmark datasets quantatively and qualitatively, and show ablation studies on different choices made in the method. In this context, we introduce a metric based on concept intervention for ante hoc explainable models such as ours.
- Our method outperforms existing methods on accuracy and explainability metrics, and achieves these results with negligible computational overhead over baseline models with no explanation component.

2. Related Work

The main objective of concept learning is to obtain a lower-dimensional representation that faithfully explains the downstream tasks, such as - object classification.

Unsupervised Concept Learning: Most of the existing methods generate meaningful explanations in an unsupervised manner, i.e., when ground truth concepts are not available for the dataset. Such methods either work as a post-hoc approach on a trained model [13] or learn an inherently interpretable model [1, 33]. TCAV [13] leverages the directional derivatives with intermediate model features to quantify the importance of a user-defined concept towards final model predictions. Though this method doesn't require full concept annotations, the explanations are generated based on the prior knowledge of concepts over the data points. Zhou et al. [38] proposed a method to decompose the model prediction in terms of projections onto concept vectors using the model generated saliency by CAM [37]. Another recent method [33] leverages Shapley values to quantify the sufficiency of a set of concepts in explaining the model predictions through completeness measure of the concepts. Being a post-hoc explainability method, it works on trained deep networks. Unlike our method, it doesn't allow a user to intervene on the concepts to explore the interactions between concepts and the class predictions.

The first fully-unsupervised ante-hoc concept learning method, SENN [1], employs a concept encoder $h(x)$ with corresponding relevances $\theta(x)$ for an image x and outputs the final logit as $\theta(x)^T h(x)$. SENN is trained, following a joint training approach, with cross-entropy loss for the logits and a stability loss to enforce closeness of the similar concept relevances i.e. $\theta(x)$. Similar to SENN, our method also uses a concept encoder to extract concepts. But, we replace the heavy relevance network with a couple of simple, fully connected networks that generate explanations and perform classification.

Supervised Concept Learning: Methods such as concept bottleneck models (CBM) [14] divides the complete model into two parts. The first part is a function $g : X \rightarrow C$, that generates an intermediate concept representation c from an image x, which is followed by the label predictor part $f : C \rightarrow Y$ to output a class label from c. The model predicts a class label for an image x by computing $f(g(x))$. This model is trained with both concept and class label supervision, either training individual parts sequentially or both parts jointly. Kazhdan et al. proposed CME [12], a post-hoc data-efficient version of CBM, that captures intermediate representations from a pre-trained model to improve the sensitivity to the dependence between the concepts and the final prediction. Concept whitening (CW) [2] proposed a method to plug an intermediate layer in place of the batch-normalization layer of any pre-trained CNN model that helps in concept extraction by constraining the latent layer output to represent a target concept. As opposed to CBM, we decouple the process of generating explanations and predictions. This helps us to learn concept-based explanations without losing much in predictive performance and enables the user to use the model with different levels of supervision.

Self-Supervised Concept Learning: Different self-supervised methods have been proposed to help learn better representations and boost classification accuracy. Tasks such as predicting the relative position of image patches [5], predicting rotation angle [9], recovering color channels [34], solving jigsaw puzzle games [20], and discriminating images created from distortion [6] have been extensively used in recent years. Another class of methods reconstruct images from corrupted versions or just part of it such as denoising autoencoders [28], image inpainting [21], and split-brain autoencoder [35]. Contrastive learning is another paradigm where representations are learned in such a way that similar data points are brought closer, and dissimilar data points are pushed further away [29] in the representation space. Predicting natural ordering or topology of data has also leveraged as pretext tasks in video-based [8,19,30], graph-based [11,32], and text-based [4,22] self-supervised learning. While self-supervision has been used to learn better model representations, their utility for learning concept-based explanations hasn't been explored in the past. In our work, we explore how self-supervision can be used for learning better concept-based explanations.

3. Method

Let $\mathcal{X}$ denote the input space, and $\mathcal{Y}$ the output space, we assume that the training instances (or examples) $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ are sampled i.i.d from the source distribution P defined over $\mathcal{X} \times \mathcal{Y}$. We also assume that $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} = \{y \in \{0,1\}^M, \sum_{j=1}^{M} y^k = 1\}$, where M is the number of classes, and y is an one-hot encoded vector.

We propose a generic framework to incorporate *ante-hoc explanation (or self-explanation) modules* into existing deep learning pipelines. In this paper we demonstrate it for a classification task. In practice, for classification tasks we learn a Deep Neural Network $f_\theta = \{\eta_{\theta_e}(.), g_{\theta_c}(.)\}$ which consists of a base encoder (or a feature extractor) $\eta_{\theta_e(.)}$, that extracts the representation the representation vectors which are fed into a classifier function $g_{\theta_c}(.)$ (a classifier function takes the latent representation $\mathbf{z} = \eta_{\theta_e}(x_i)$, and then predicts the label). Typically the base encoder & the classifying function are trained together by optimizing for $\theta = \{\theta_e, \theta_c\}$ such that the output of the network $\tilde{y}_i = f_\theta(x_i)$ minimizes a loss $\mathcal{L}_C(\tilde{y}_i, y_i)$ over the set of training instances $\mathcal{D}$.

To incorporate implicit learning of interpretable concepts, in addition to the existing components of classical classification pipelines described previously, we introduce a concept encoder $\Psi_{\theta_{ce}}(.)$ which takes the representation $\eta_{\theta_e}(.)$ and learns a set of *intepretable concepts* $\{\psi^1, \ldots, \psi^C\}$ (where C is the number of concepts), to explain the predictions provided by f_θ. In general, concepts are low dimensional representation that can be characterized as $C \in \mathbb{R}^{K \times d}$, i.e. every concept $c \in \mathbb{R}^d$ belongs to one of the total $k \in K$ concepts. In our work, we learn one-dimensional concepts, i.e., our setup uses k concepts, with every concept is represented by a scalar value.

To encourage the model to learn concepts $\{\psi^1, \ldots, \psi^C\}$ that capture the semantics of the input image x_i we pass the concepts to a decoder $h_{\theta_d}(.)$ which reconstructs the image $\hat{x}_i$. We then add a loss $\mathcal{L}_R(x_i, \hat{x}_i)$ which measures the reconstruction error to the overall loss $\mathcal{L}$. $\mathcal{L}_R$ penalises the model f_θ, if the concepts aren't suffice to generate an accurate reconstruction $\hat{x}_i$ of the input image x_i. In our paper, we use an L_2 loss.

Since the concepts $\{\psi^1, \ldots, \psi^C\}$ explain the prediction of a DNN f_θ. Ideally, they should be informative enough by themself to predict the input instance x_i correctly. To enforce that the learnt concepts not only explain the prediction but are also informative, we penalize the model f_θ, if the predictions $s_{\theta_{cce}}(\Psi_{\theta_{ce}}(.))$ (where $s_{\theta_{cce}}$ is a classification function which predicts the class labels taking the concepts as input) based on the concepts $\{\psi^1, \ldots, \psi^C\}$ and prediction by the DNN f_θ differs. We enforce that the concepts learned should be individually informative by adding a fidelity loss $\mathcal{L}_F$ to the overall loss $\mathcal{L}$.

Taking the proposed modifications into consideration, the overall loss $\mathcal{L}_O$ of the model can be written as follows:

$$\begin{aligned}\mathcal{L}_O = {} & \mathcal{L}_C(y_i, \tilde{y}_i) + \alpha \mathcal{L}_R(x_i, \hat{x}_i) \\ & + \beta \mathcal{L}_F(f_\theta(x_i), s_{\theta_{cce}}(\Psi_{\theta_{ce}}(x_i)))\end{aligned} \tag{1}$$

In practice, most data sets seldom include annotations of concepts (or attributes) that could be used to learn a self-explaining model. However, few exceptions contain concepts (or attributes) that the models can leverage while learning to explain their predictions. The majority of existing frameworks either work when only annotation of concepts is available, or data sets don't contain any additional annotations, but not both. Often it is neither trivial nor efficient to incorporate alternate forms for supervision in these existing frameworks. In comparison, our framework offers the flexibility to incorporate different forms of supervision whenever available easily. To illustrate this, we demonstrate how to incorporate i) complete supervision (supervised learning of interpretable concepts), ii) zero-supervision (unsupervised learning of interpretable concepts), and iii) a weaker form of supervision that is cheaply available like self-supervision.

By default, our framework works with data sets where the annotation of concepts isn't available. In cases when they are available, we can easily incorporate them into the learning process by adding a loss $\mathcal{L}_E(\Psi_{\theta_{ce}}(x_i), a_{x_i})$

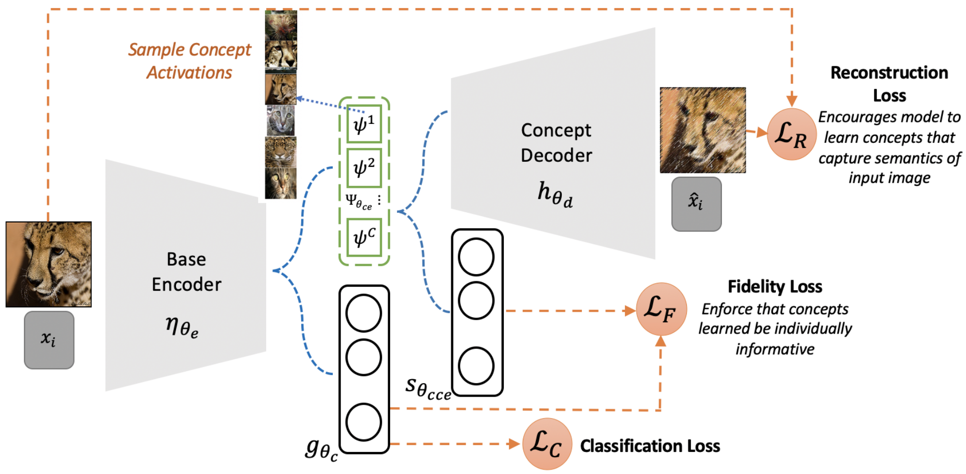

Figure 2. An overview of our proposed framework (concept activations denote images that maximally active each concept)

(where a_{x_i} is the concept (or attribute) annotation of x_i). $\mathcal{L}_E(\Psi_{\theta_{ce}}(x_i), a_{x_i})$ would penalize the model if the concepts learned aren't similar to the annotation in the data set for the corresponding instance. We then train the model by optimising θ such that the output of the network $\tilde{y}_i = f_\theta(x_i)$ minimizes a loss $\mathcal{L}_O + \mu\mathcal{L}_E$ over the training set.

Even when direct supervision isn't available for concepts, it is possible to learn a robust set of high fidelity interpretable concepts $\{\psi^1, \dots, \psi^C\}$ by leveraging the underlying structure of the data by incorporating supervisory signals obtained directly from the data itself. This technique is popularly known as self-supervision. In our framework, we incorporate *self-supervision* as an auxiliary task with a loss $\mathcal{L}_{SS}$, and the auxiliary task shares the parameters with our model until the concept encoder $\Psi_{\theta_{ce}}(.)$. In this paper, we choose rotation prediction as an auxiliary task. The task involves rotating the image by one of 0, 90, 180, or 270 degrees and predicting the rotation angle r_i as a four-way classification problem through an auxiliary head. We can also easily incorporate other self-supervision tasks into our framework.

As opposed to existing techniques where the branch for the auxiliary task uses the output of feature extractor (or base encoder) for the self-supervision tasks, in our case, we use the output of the concept encoder $\Psi_{\theta_{ce}}(.)$. In turn, this helps us to ensure that the set of interpretable concepts from the concept encoder $\Psi_{\theta_{ce}}(.)$ always respects the underlying structure of the data and has high fidelity. To estimate $\mathcal{L}_{SS}$ we pass the output of the concept encoder $\Psi_{\theta_{ce}}(.)$ through a classifier function $\zeta_{\theta_{ss}(.)}$ that predicts the angle of rotation, we then compute cross-entropy between $\zeta_{\theta_{ss}(.)}$ and r_i. Like in other cases, we jointly train the model and the auxiliary head by optimizing θ such that the output of the network $\tilde{y}_i = f_\theta(x_i)$ minimizes a loss $\mathcal{L}_O + \gamma\mathcal{L}_{SS}$ over the training set. In cases where ground truth annotations of concepts aren't available, and an auxiliary self-supervision task isn't used μ, and γ are respectively set to 0.

$$\mathcal{L}_O^{'} = \mu\mathcal{L}_E(\Psi_{\theta_{ce}}(x_i), a_{x_i}) + \gamma\mathcal{L}_{SS}(r_i, \zeta_{\theta_{ss}(.)})$$

Even though our framework incorporates additional components to existing deep learning backbones (or pipelines), we can discard most of them after the training. We only retain the sub-network (or module) to generate explanations in addition to the ones on standard deep learning pipelines (i.e., feature extractor and the classifier function) during the prediction time. Hence, compared to existing self-explaining models, the additional cost incurred by our framework is relatively insignificant.

4. Experiments

We show that our framework achieves competitive predictive accuracy compared to standard classification pipelines, as well as meaningful explanations. We report results with our method on CIFAR10, ImageNet, AwA2 and CUB-200 with different levels of concept supervision according to availability of ground truth concepts i.e. unsupervised manner on CIFAR10 [15], ImageNet [3] and with concept supervision on AwA2 [17], CUB-200 [31]. We also report results with AwA2, CUB when our model is trained without concept supervision to show the effectiveness of our method in both cases, i.e., with and without concept supervision. We consider SENN [1] and CBM [14] as our baselines considering the basic methods for unsupervised and supervised concept learning. The implementation of our method is publicly available at this link.

Dataset Details: The CIFAR-10 dataset [15] consists of 32x32 colour images in 10 classes, each with 5000 train images and 1000 test images per class. The ImageNet dataset [3] is comprised of more than 1 million images and 1,000 object classes of natural images. AwA2 dataset [17] consists of 37322 images of total 50 animals classes with 85 numeric attribute. The other attribute dataset we considered is CUB-200 [31], an image dataset with photos of 200 bird categories with a total of 6033 images and 312 attribute annotations for each image.

Architecture Details: We use ResNet18 [10] as our backbone network for all datasets, as there is no standard architecture followed in the literature related to concept learning. The backbone network resembles to $f_\theta = \{\eta_{\theta_e}(.), g_{\theta_c}(.)\}$ as given in Sec.3. The output of the feature encoder $\eta_{\theta_e}(.)$ is also passed to the concept encoder $\Psi_{\theta_{ce}}(.)$ which is a single fully connected layer, that outputs a set of interpretable concepts $\{\psi^1, \ldots, \psi^C\}$ where C is the number of concepts. We considered 10 and 100 concepts for CIFAR10 and ImageNet respectively. The number of concepts (or attributes) for AwA2 and CUB-200 is 85 and 312, respectively. We kept the number of concepts the same for fair comparison while training our model with these datasets for both unsupervised concept learning and learning with concept supervision. The classification function $s_{\theta_{cce}}$ that predicts the class labels, taking the concepts as input, is also a single fully connected layer. The number of parameters of the concept encoder and the classification network, taking the concepts as inputs, vary for different datasets based on the number of concepts and classes. We implement the decoder $h_{\theta_d}(.)$ as a set of deconvolution layers.

Storage and Time Complexity: The architecture proposed by SENN requires a vast number of parameters with both the concept and relevance encoder contributing to it. Our model alleviates this issue by removing the relevance network altogether and adding the concept classification network that serves a similar purpose. However, the decoder network is required to make the concepts capture sufficient information to reconstruct the image. Hence, our overall network requires ∼60% of the space and training time compared to SENN. Compared to CBM, our method requires ∼1.5 times the space and training time. For example, the training times required for one epoch on CIFAR10 are 4.2s, 6.9s & 11.3s for CBM, Ours & SENN methods with batch size 128 in one Tesla V100 GPU. This is due to a decoder that enables our framework to support cases when concept supervision isn't available, which CBM doesn't. Our approach takes almost similar inference time as CBM as we don't use the decoder network during inference and concept extraction. Please note that these storage and time measurements during training are with ResNet18 backbone architectures, and the gap with CBM will further reduce with more complex backbone networks.

Predictive Performance: Table 1 reports the predictive performance of our method as well as the baseline methods with CIFAR10, ImageNet, AwA2, and CUB datasets. As CBM requires concept supervision, we can't use this method for CIFAR10 and ImageNet. An unsupervised version of our method outperforms SENN significantly for all the datasets. CBM, being a method with concept supervision, performs slightly better than our unsupervised version. Our approach, with concept supervision, beats CBM by a large margin. Please note that the predictive performance by our method, reported in table 1, is solely based on the backbone network $f_\theta(.)$. We decoupled the main prediction task and concept extraction so that our model doesn't sacrifice much of the predictive performance and still can produce meaningful explanations.

	Baselines		OURS	
Dataset	SENN	CBM	w/o sup	w sup
CIFAR10	84.50	NA	**91.68**	NA
ImageNet	58.55	NA	**65.09**	NA
AwA2	76.41	81.61	81.04	**85.70**
CUB-200	58.81	64.17	63.05	**65.28**

Table 1. *Accuracy* (in %) of different methods on CIFAR10, ImageNet, AwA2 and CUB-200 datasets using ResNet18 architecture as concept (or base) encoder. (w=with, w/o=without)

4.1. Quantitative Evaluation

We evaluate and compare the concept-based explanations generated by our method with other state-of-the-art frameworks like SENN and CBM. We consider metrics of interpretability that assess the effectiveness of additional losses we use in our framework. Apart from the existing metrics such as faithfulness, fidelity, and explanation error, we also perform interventions on the generated concepts to illustrate their meaningfulness. Fig. 3 shows examples of interventions that lead to the model changing its prediction when we intervene on the top concept. Besides the predictive performance, our method consistently outperforms the baseline methods in all the other explainability metrics as explained below.

Faithfulness Metric: In practice, we want the concepts learned to be meaningful and faithfully explain the model's predictions. To evaluate how faithful the explanations generated by different frameworks are, we measure the predictive capacity of the generated concepts, i.e., from the output of $s_{\theta_{cce}}$, in our case. This metric represents the capability of the overall concept vector to predict the ground truth task label. It is similar to other measures such as explicitness [24] and informativeness [7] used to measure feature disentanglement.

	Baselines		OURS	
Dataset	SENN [1]	CBM [14]	w/o sup	w sup
CIFAR10	84.50	NA	**90.86**	NA
ImageNet	58.55	NA	**59.73**	NA
AwA2	76.41	81.61	79.29	**83.30**
CUB-200	58.81	64.17	61.49	**62.59**

Table 2. Comparison of *faithfulness* (in %, predictive performance solely based on concepts) of concepts generated by different methods on CIFAR10, ImageNet, AwA2 and CUB-200 data sets. (w=with, w/o=without)

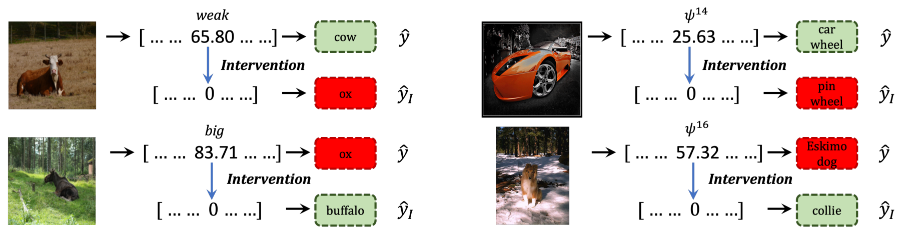

Figure 3. Successful examples of test-time intervention, where intervening on a single concept changes the model prediction (from $\hat{y}$ to $\hat{y}_I$) to the correct label on bottom examples and to the incorrect label on top examples, for AwA2 (left) and ImageNet (right).

Fidelity Metric: Fidelity measures the fraction of the data points where the model prediction matches the prediction from the interpretation. It is widely used to measure how well the generated explanations approximate the model [16]. This metric does not apply to methods where the interpreter is directly used to provide model prediction, such as SENN and CBM. Table 3 reports all the comparative results for all the datasets. We use fidelity loss $\mathcal{L}_F$ during training which justifies the high fidelity score for our models.

Dataset	OURS w/o sup	OURS w sup
CIFAR10	**99.11**	NA
ImageNet	**90.22**	NA
AwA2	**97.84**	97.19
CUB-200	**97.52**	95.87

Table 3. Comparison of *fidelity* (the % of match between model prediction and the prediction through the interpretation.) of concepts generated by different methods on CIFAR10, ImageNet, AwA2 and CUB-200 data sets. (w=with, w/o=without)

Explanation Error: In data sets like CUB and AwA2, where ground truth concepts are available, we also measure how close are the concepts learned to the ground truth. We compute the L_2 distance between the concepts learned and the ground truth concepts to measure the alignment. From table 4, we can observe that the concepts generated by our method is most aligned to the ground truth concepts. While this should be the case for methods with concept supervision, our method without concept supervision also performs better than SENN, which illustrates our method's effectiveness in learning concept-based explanation even when annotations for concepts (or attributes) aren't available.

Dataset	AWA2	CUB
SENN	0.99	1.34
CBM	0.91	1.17
OURS (w/o sup)	0.97	1.29
OURS(w sup)	**0.89**	**1.14**

Table 4. Comparison of *explanation error* (we measure the mismatch using L_2 distance, hence lower the better) between concepts generated by different methods and the ground truth concepts (or attributes) on AwA2 and CUB-200 data sets. (w=with, w/o=without)

Intervention on Concepts: To study the concepts' usefulness, we scale their values in the $[0, 1]$ range, select those above the threshold value ω, set the concepts to 0, and then predict the label solely based on the intervened concept vector. A change in the prediction means that the concepts zeroed are essential for explaining the model's decision. We repeat this procedure for all the instances in the test set and measure the predictive performance solely based on the generated concepts. A lower value indicates that the concepts generated are faithfully explaining the models' decisions. Ideally, the predictive ability of the concepts generated by methods like SENN and CBM should be higher. Since, in their cases, the interpreter (or the explainer) is directly used to generate the model prediction, the predictive performance based on concepts generated should be lower. But, you can observe from table 5 the predictive performance after the intervention is the lowest for the proposed framework.

Dataset	Baselines SENN	Baselines CBM	OURS w/o sup	OURS w sup
CIFAR10	66.57	NA	**43.19**	NA
ImageNet	43.91	NA	**34.52**	NA
AwA2	61.39	40.29	37.61	**35.92**
CUB-200	47.22	36.11	34.38	**32.59**

Table 5. The effect of interventions (accuracy in % after intervention, lower the better) on concepts generated by different methods for CIFAR10, ImageNet, AwA2 and CUB-200 data sets. (w=with, w/o=without)

4.2. Qualitative Results

Qualitative results are significant for methods that explain models through concept-based representations. We generate explanations corresponding to every concept as the most representative images from the dataset. We present results from CIFAR10 and ImageNet datasets in the main paper and move the rest to the Appendix due to space constraints. The top concept activations generated for ImageNet are presented in Fig.4. We can observe that every concept captures homogeneous characteristics from the dataset that mostly corresponds to a class or similar class type. For example, ψ^7 represents concepts of faces for cheetah and some other similar types of cat species.

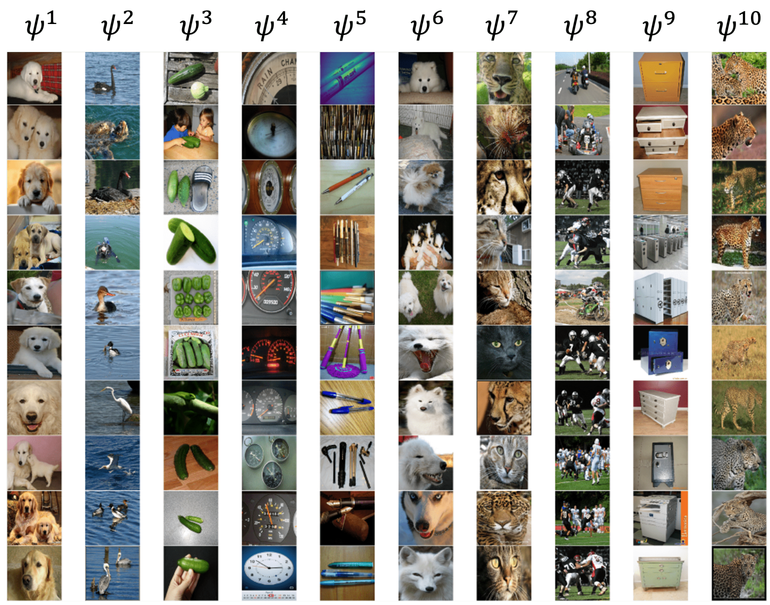

Figure 4. A subset of 10 concept activations learnt by our framework on ImageNet. All these examples were correctly predicted by the model, and it can be seen that the each concept captures a certain set of homogeneous properties corresponding to a class. For ImageNet, we observe that the learned concepts are shared across the classes. For instance, ψ^7 is shared between tiger, cheetah, and different types of cat classes, and ψ^6 is shared among different forms of wolf and dog classes.

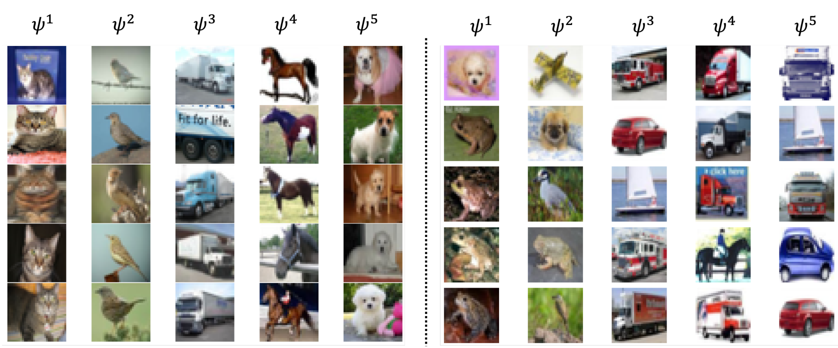

Figure 5. Effect of decoder for CIFAR10 dataset. We can see the without the decoder & the corresponding reconstruction loss the concept-based explanations (on the right) doesn't capture any homogeneous property and are hard to understand, unlike the case where decoder is present (on the left)

For data sets like CIFAR10, where there isn't much intersection of higher-level attributes across classes, we observe that each learned concept only corresponds to characteristics (or features) from a single class. In comparison, for data sets like ImageNet, where there is a lot of intersection between the higher-level attributes of different classes, we observe that the learned concepts are shared across the classes. For instance, ψ^7 is shared between tiger, cheetah, and different types of cats classes, and ψ^6 is shared among different types of wolf and cat classes (refer Figs 4 & 5).

4.3. Global Explanations

An advantage of concept-based explanation methods compared to others is that they provide local as well global explanations. We identify class-concept (or attribute) pairs with a high proportion of co-occurrence to generate global explanations. We consider CIFAR10 and AwA2 for our experiments to explain the effectiveness of our method in generating such global explanations on datasets without and with ground truth concepts. Simply analyzing these can reveal helpful information about the generated concepts. For instance, based on samples, we can see that (from Fig.6) concept 'ocean' is a distinguishing attribute for class killer+whale of AwA2. Similarly, ψ^1 represents a distinctive concept for cat class of CIFAR10 (from ψ^1 to ψ^5 of Fig.5 on the left).

5. Ablation Studies

Importance of Self-supervision: As discussed in Sec.3, our framework enables us to incorporate self-supervision on

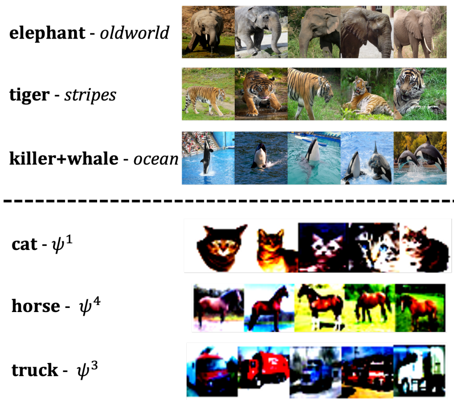

Figure 6. Example class-attribute pair analysis on AwA2 and CIFAR10 datasets with high global relevance (proportion of co-occurrence)

Dataset	Baseline SENN	OURS w/o sup	OURS w self-sup
CIFAR10	84.5	91.68	91.28
ImageNet	58.55	65.09	64.84
AwA2	76.41	81.04	79.89
CUB-200	58.81	63.05	61.93

Table 6. Comparison of predictive performance (accuracy in %) of models on different methods on CIFAR10, ImageNet, AwA2 and CUB-200 data sets with and without self-supervision. (w=with, w/o=without)

the set of concepts easily, and this helps us to improve the quality of concepts by leveraging the underlying structure of the data. Our experiments use rotation prediction as the auxiliary self-supervision task for CIFAR10 and ImageNet datasets due to the unavailability of ground truth concepts. To compare the quantitative and qualitative performance with unsupervised concept training, we train our model with self-supervision for AwA2 and CUB datasets too. Table 6 reports the predictive performances of our method (with self-supervision on concepts and without any supervision) and SENN, as these methods don't require any ground truth concepts. Please note that the self-supervision is performed only on the concepts. Hence, this doesn't improve the predictive performance but rather enhances the faithfulness of the learned concept-based explanations. From table 7, we can observe that self-supervision is improving the predictive performance through the concepts, and this, in turn, validates our hypothesis that leveraging the underlying structure of the data can the quality of concepts.

Importance of Reconstruction Error: A decoder improves the quality of the concepts by enforcing sufficiency, i.e., making them capable of faithfully reconstructing the image. In other words, this compels the set of concepts to capture all the image information and make the set of concepts complete. We measure the effect of the decoder by training our model for the CIFAR10 dataset without the decoder, keeping all the other model parts unchanged. We generate the explanations of the trained model and present them in fig. 5. For comparing with the complete model (i.e., our model with decoder), we add explanations generated by our complete model in the same figure. The first and the last five columns are explanations generated by our complete model and the model without a decoder, respectively. These examples support our claim about the importance of decoder for learning better concepts. Please note that the model without a decoder performs slightly better than our complete model, but sacrificing a little bit of predictive performance can be justified to gain trust in the model.

Dataset	Baseline SENN	OURS w/o sup	OURS w self-sup
CIFAR10	84.50	90.86	90.93
ImageNet	58.55	58.73	60.28
AwA2	76.41	79.29	79.77
CUB-200	58.81	61.49	61.81

Table 7. Comparison of *faithfulness (in %)* (predictive capacity of the generated concepts) of concepts generated by different methods on CIFAR10, ImageNet, AwA2 and CUB-200 data sets with and without self-supervision. (w=with, w/o=without)

6. Conclusion

In this work, we propose a new framework towards learning ante-hoc concept-based explanations that: (i) can be added easily to existing backbone classification architectures with minimal additional parameters; (ii) can provide explanations for model decisions in terms of concepts for an individual input image or groups of images; & (iii) can work with different levels of supervision, including no concept-level supervision at all. Even though our framework incorporates additional components to existing deep learning backbones (or pipelines), we can discard most of them after the training. We only retain the sub-network (or module) to generate explanations in addition to the ones on standard deep learning pipelines (i.e., feature extractor and the classifier function) during the prediction time. Hence, compared to existing self-explaining models, the additional cost incurred by our framework is relatively insignificant. We performed a comprehensive suite of experiments to study the accuracy and explainability of our method on multiple benchmark datasets both quantitatively and qualitatively. Our approach consistently outperforms the baseline methods in all the datasets. In addition to this, we also performed ablation studies to illustrate the importance of additional components added by our method.

Acknowledgements. This work has been partly supported by the funding received from MoE and DST, Govt of India, through the UAY and ICPS programs. We thank the anonymous reviewers for their valuable feedback that improved the presentation of this paper.

References

[1] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, 2018. 2, 4, 5, 11

[2] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[6] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2016. 3

[7] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. 5

[8] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 3

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 5

[11] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2019. 3

[12] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (cme): concept-based model extraction. In *AIMLAI workshop at the 29th ACM International Conference on Information and Knowledge Management*, 2020. 2

[13] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2

[14] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2, 4, 5, 11

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 4

[16] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5628–5638. PMLR, 2020. 6

[17] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013. 4

[18] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. 1

[19] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 527–544. Springer, 2016. 3

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016. 3

[21] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. 3

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, 2016. 1, 11

[24] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, 2018. 5

[25] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 05 2019. 1

[26] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2021. 1

[27] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In *ICCV'17*, 2017. 1, 11

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the*

25th international conference on Machine learning, pages 1096–1103, 2008. 3

[29] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 3

[30] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 3

[31] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 4, 5

[32] Han Yang, Xiao Yan, Xinyan Dai, and James Cheng. Self-enhanced gnn: Improving graph neural networks using model outputs. In *International Joint Conference on Neural Networks*, 2020. 3

[33] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. 2

[34] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 3

[35] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 3

[36] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR'16*, 2016. 1

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[38] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 2