Towards Better Generalization and Interpretability in Unsupervised Concept-Based Models

Francesco De Santis[1] (✉), Philippe Bich[1], Gabriele Ciravegna[12], Pietro Barbiero[3], Tania Cerquitelli[1], and Danilo Giordano[1]

[1] Politecnico di Torino, Torino, 10129, Italy {name.surname}@polito.it
[2] CENTAI Institute, Torino, 10138, Italy {name.surname}@centai.eu
[3] Universita' della Svizzera Italiana, Lugano, 6900, Switzerland.
name.surname@usi.ch

Abstract. To increase the trustworthiness of deep neural networks, it is critical to improve the understanding of how they make decisions. This paper introduces a novel unsupervised concept-based model for image classification, named Learnable Concept-Based Model (LCBM) which models concepts as random variables within a Bernoulli latent space. Unlike traditional methods that either require extensive human supervision or suffer from limited scalability, our approach employs a reduced number of concepts without sacrificing performance. We demonstrate that LCBM surpasses existing unsupervised concept-based models in generalization capability and nearly matches the performance of black-box models. The proposed concept representation enhances information retention and aligns more closely with human understanding. A user study demonstrates the discovered concepts are also more intuitive for humans to interpret. Finally, despite the use of concept embeddings, we maintain model interpretability by means of a local linear combination of concepts.

Keywords: CBM · XAI · Interpretable AI.

1 Introduction

Understanding the *reason* why Deep Neural Networks (DNNs) make decisions is critical in today's society, as these models are increasingly deployed and affect people's lives. This concern has also led regulatory institutions to mandate interpretability and the possibility of challenging the decisions of deep neural networks as prerequisites for Artificial Intelligence (AI)

systems [38,27]. EXplainable AI (XAI) methods have emerged to address this challenge [30,2,12]. However, several papers argue that feature importance explanations (such as saliency maps [45,33]) have failed to achieve this goal, since showing where a network is looking is insufficient to explain the reasons behind its decisions [31,1]. To truly explain what the network has seen, many XAI methods are shifting toward explanations in terms of human-understandable attributes, or *concepts* [14,1,10,28]. Concepts can be either extracted post-hoc [11] or directly inserted within the network representation to create a so-called concept-based model (CBMs, [16]).

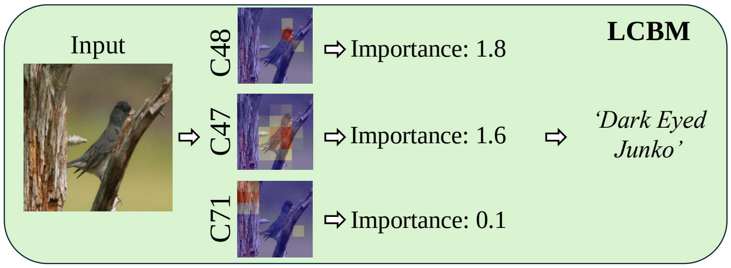


Fig. 1: Learnable Concept-Based Model (LCBM) learns a dictionary of unsupervised concepts. Unlike black-box models, LCBM classifies images interpretably using these concepts. Here, the image is correctly classified as *Dark-eyed junco* by leveraging concepts C48 (eyes/beak), C47 (wings), and C71 (trunk/tree). Notably, C71, while present, is less relevant to bird species classification.

CBMs can be created in a supervised way [16,4,8] or through a dedicated unsupervised learning process [3,6,40]. The latter approach enables the use of CBMs in contexts where concept annotations are unavailable and large language models (LLMs) [41,26] lack sufficient knowledge. Yet, a fundamental challenge persists: standard unsupervised approaches rely on single-neuron activations to represent each concept, thereby limiting the amount of information that can be captured. This limitation creates a trade-off between interpretability and accuracy. The issue becomes even more pronounced when concise explanations are needed to avoid cognitive overload for users [24,19,7]. As demonstrated in our experiments, standard unsupervised approaches exhibit a significant performance gap compared to end-to-end methods in such scenarios, making unsuitable their effective deployment. In this paper, we demonstrate that by using unsupervised concept embeddings, we can create a highly effective Learnable Concept-Based Model (LCBM) employing a limited number of concepts. Our experiments show that this approach: i) overcomes the limited generalization of compared models, almost matching the performance of black-box models; ii) increases the representation capability of standard unsupervised concept layers in terms of information retention and alignment with human representation; iii) ensures that the extracted concepts are more interpretable, as highlighted by a user study; and iv) by providing the task prediction through a local linear combination of concepts, it retains task interpretability[§].

[§] Code to reproduce the proposed model is available at *https://github.com/LCBM*.

2 Related Work

Concept-based XAI (C-XAI) aims to provide human-understandable explanations by using concepts as intermediate representations [14,28,31]. While supervised approaches [16,8] rely on predefined symbols, unsupervised models autonomously extract concepts by modifying a network's internal representation through unsupervised learning, prototypical representations, or hybrid techniques [3,16].

Unsupervised Concept Basis. These methods learn disentangled representations in the model's latent space by grouping samples based on fundamental characteristics. They typically achieve this via input reconstruction [3,40] or unsupervised losses [46,40]. In [46], convolutional filters act as unsupervised concepts, maximizing mutual information between images and filter activations. SENN [3] employs an autoencoder to derive clustered representations and generate class-concept relevance scores. BotCL [40] enhances SENN with attention-based concept scoring and contrastive loss. Compared to these, LCBM introduces concept embeddings for richer representations, improving the generalization-interpretability trade-off.

Prototype Concepts. This approach encodes training example traits as prototypes within the network, comparing them to input samples for prediction. [21] explains predictions via prototype similarity, using an autoencoder for dimensionality reduction. ProtoPNet [6] extracts prototypes representing image subparts, while HPNet [13] organizes prototypes hierarchically for classification across taxonomy levels. Despite providing useful example-based explanations, these models constrain representation capacity. Our approach enhances performance by leveraging richer representations while retaining prototype-based interpretability, as shown in concept dictionaries.

Hybrid Approaches. Recent research explores hybrid models that integrate supervised and unsupervised concepts [23,32] or leverage pre-trained LLMs [26,41,44]. The variational approach in [23] shares similarities with ours but relies on single neurons and partial supervision, limiting scalability. Methods leveraging LLMs assume sufficient knowledge for zero-shot concept annotations, yet this depends on the underlying model [35]. For instance, CLIP [29], despite its popularity, exhibits low concept accuracy even in contexts similar to its pre-training, as confirmed by our experiments.

3 Methodology

In an unsupervised concept-based setting, the objective is to make predictions using a set of abstract, human-interpretable concepts that are not predefined but must be directly inferred from the data. To address this challenge, we propose a set of desiderata that define the required properties of the learned concepts:

- *Representativity* [5]: Concepts should capture key features of the input data.
- *Completeness* [43]: Concepts should support strong task generalization.

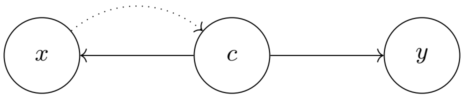


Fig. 2: Probabilistic Graphical Model. Solid arrows represent the data generating process. Dotted arrows represent inference.

- *Alignment* [28]: Concepts should correspond to human-understandable constructs.

If concepts do not accurately represent the input data, they cannot be trusted for either inference or explainability. As a result, *Representativity*, a common requirement in representation learning [5], is a necessary but not sufficient condition for an interpretable unsupervised concept-based model: *Completeness* is also essential to ensure the concepts are useful for making task-specific predictions. Ultimately, interpretability is achieved only with an *Alignment* with a human-defined representation. While this property is always met in supervised CBMs, in unsupervised contexts it is a major challenge.

3.1 Learnable Concept-Based Model

In a supervised learning context, a CBM is trained to approximate the joint distribution $p(x, c, y)$, where x, c, and y correspond to realizations of the random variables X (images), C (concepts), and Y (class labels), respectively. In contrast to the supervised setting, where these variables are fully observable during training, the unsupervised scenario lacks knowledge about C. Therefore, it is only through marginalizing over C that we can account for the combined effect of all possible values of C on the relationship between X and Y:

$$p(x, y) = \int_c p(x, c, y)\, dc \tag{1}$$

In order to address this problem, we introduce the Learnable Concept-Based Model (LCBM), a latent variable model enabling explanations and interventions in terms of a set of unsupervised concepts. Following [25], LCBM considers a data generating process in which concepts C represent latent factors of variation for both X and Y, as shown in the probabilistic graphical model in Figure 2. Thus, the joint distribution factorizes as:

$$p(x, y) = \int_c p(x, c, y)\mathrm{d}c = \int_c p(x \mid c)p(y \mid c)p(c)\mathrm{d}c \tag{2}$$

where $p(y \mid c)$ is modelled as a categorical distribution parametrized by the task predictor f; $p(x \mid c)$ as a Gaussian distribution parametrized by the concept decoder ψ. Finally, $p(c)$ is a prior distribution over a set of unsupervised concepts.

During training, LCBM assumes to observe realizations of the random variables X and Y which hold new evidence we can use to update the prior $p(c)$.

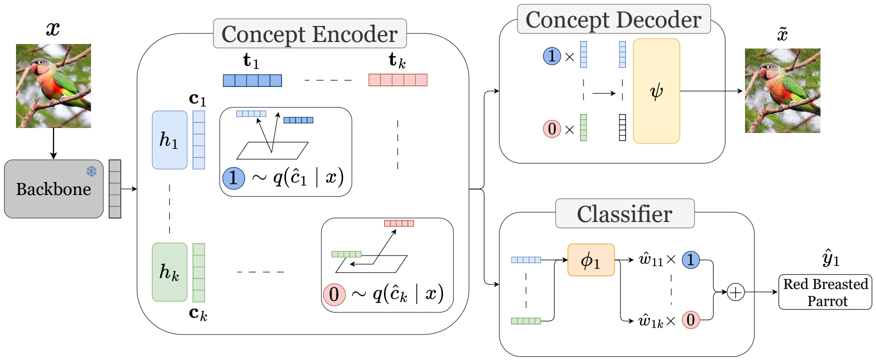


Fig. 3: LCBM schema. The concept encoder $q(c \mid x)$ provides the probability for each learnt concept $\hat{c}_j$ and the associated embeddings $\mathbf{c}_j$. Both concept scores and embeddings are used to predict the output class $p(y \mid c)$ and to reconstruct the input $p(x \mid c)$.

Since the computation of the true posterior $p(c \mid x, y)$ is intractable, LCBM amortizes inference needed for training by introducing an approximate posterior $q(c \mid x)$ parametrized by a neural network. Since at test time we can only observe X, we condition the approximate only on this variable.

Optimization problem. LCBMs are trained to optimize the log-likelihood of tuples (x, y). Following a variational inference approach, we optimize the evidence lower bound (ELBO) of the log-likelihood, which results as follows:

$$\text{ELBO} = \overbrace{E_q[\log\, p(x|c)]}^{\text{Representativity}} + \overbrace{E_q[\log\, p(y|c)]}^{\text{Completeness}} - \overbrace{KL(q(c|x)\; ||\; p(c))}^{\text{Alignment}} \tag{3}$$

This likelihood has three components: a reconstruction term $p(x|c)$, whose maximization ensures concepts' *Representativity*; a classifier $p(y \mid c)$, which quantifies concepts' *Completeness*; and a Kullback–Leibler divergence term that encourages the approximate posterior $q(c \mid x)$ to remain close to a defined prior $p(c)$, promoting an *Alignment* to human representations. To achieve all the desired properties, we must define a sufficiently rich concept representation. We describe the latter together with its prior in Section 3.2, the classifier in Section 3.3 and the decoder in Section 3.4. For an overall visualization of LCBM, see Figure 3.

3.2 Unsupervised Concept Representation

To model each concept in the concept representation c in an unsupervised way, we define it as following a Bernoulli distribution. This choice reflects a discrete, binary nature of a 'concept' as an atomic unit of knowledge, inducing *Alignment* and facilitating its comprehension and modification through human interventions. However, Bernoulli distributions may not be able to represent both the input and output distributions, ultimately creating a bottleneck in the representation of the model.

To solve this issue, we associate each concept with a corresponding unsupervised concept embedding $\mathbf{c}_j \in \mathbf{C} \subseteq \mathbb{R}^d$. This embedding provides a richer representation of the concept, capturing further nuances (e.g., ‘color‘ and ‘size‘ if the concept represents a ‘vehicle‘). The concept embedding $\mathbf{c}_j$ is derived as a composition of two neural modules $h \circ g$. The latter is a frozen pre-trained backbone $g : X \rightarrow E$ mapping input X into an intermediate embedding space $E \subseteq \mathbb{R}^s$, while the first module, $h_j : E \rightarrow \mathbf{C}$, is a per-concept MLP producing the concept embedding $\mathbf{c}_j = h_j(g(x))$. To ensure that this embedding is representative of the intended concept and allows interventions, we assign each concept a prototype $\mathbf{t}_j \in \mathbb{R}^d$, which serves as a learned reference within the embedding space of $\mathbf{c}_j$. To compute the concept score $\hat{c}_j$, we first calculate the alignment between the concept embedding $\mathbf{c}_j$ and its prototype $\mathbf{t}_j$ through their dot product $\mathbf{c}_j \cdot \mathbf{t}_j$, and transform it into a probability $\pi_j = \sigma(\mathbf{c}_j \cdot \mathbf{t}_j) \in [0, 1]$ via a sigmoid function σ. Using π_j as the probability parameter, we sample from a Bernoulli distribution applying the reparametrization trick [22] to obtain the concept score $\hat{c}_j$. Thus, the final concept score is sampled from the following distribution:

$$\hat{c}_j \sim q(\hat{c}_j \mid x) = \text{Bern}(\hat{c}_j; \pi_j) \cdot p(\mathbf{c}_j \mid x), \tag{4}$$

where $p(\mathbf{c}_j \mid x)$ is the probability distribution parametrized by $h(g(x))$ and can be modelled either via Gaussian distributions [15], or through a degenerate Dirac delta distribution, without assuming any uncertainty. For the sake of simplicity, we choose the second approach.

Batch Prior Regularization. The parameter α parameterizes the Bernoulli prior in the KL term of Eq. 3, and determines the activation probability of each concept for every sample. It is fundamental to optimize KL divergence over a batch of sample rather than for each sample. Indeed, by setting $\alpha = 0.2$, and performing the optimization for each sample, we force each concept to activate for each sample with 20% confidence. This behaviour is far from optimal as we want LCBM to be sure about the presence or absence of a certain concept in a specific sample, i.e., producing $\pi_j \approx 1$ for a sample which contains a specific concept and $\pi_j \approx 0$ otherwise. To address this, we shift the KL divergence optimization at the batch level by averaging the activation probabilities for a concept j across the batch: $\bar{\pi}_j = \frac{1}{B} \sum_{z=1}^{B} \pi_{jz}$, where B is the batch size.

3.3 Interpretable Classifier

The classifier $f(\mathbf{c}, \hat{c})$ leverages both concept embeddings and concept scores to boost prediction accuracy without sacrificing interpretability. Specifically, each class prediction $\hat{y}_i$ is represented as a linear combination of concept scores $\hat{c}_j$ and associated weights $\hat{w}_{ij} \in \mathbb{R}$, where the weights are predicted over the concept embedding, i.e., $\hat{w}_{ij} = \phi_i(\mathbf{c}_j)$, and ϕ_i is a class-specific function parameterized by a neural network. The output prediction $\hat{y}$ is then computed as

$\hat{y} = \underset{i}{\operatorname{argmax}}\ p(y_i|c) = \underset{i}{\operatorname{argmax}}\ \sum_j \hat{w}_{ij} \cdot \hat{c}_{ij}$. Note that $\hat{w}_{ij}$ depends on the concept embedding $\mathbf{c}_j = g(x)$ predicted for a specific sample. This means that while the final prediction is provided by means of a linear classification over the concept scores, thus preserving locally the interpretability of standard CBMs, the network ϕ can predict different weights $\hat{w}_{mj}$ for different samples, thus overcoming the representation bottleneck of standard CBMs.

3.4 Concept Decoder

As previously introduced, we parametrize the decoding function with a neural network ψ. To improve the image reconstruction capabilities, also in this case we rely on the concept embeddings $\mathbf{c}$. However, to still take into account the associated concept predictions $\hat{c}$, we multiply the embeddings $\mathbf{c_j}$ by the corresponding concept prediction $\hat{c}_j$ before feeding them to the concept decoder. As a result, the input is reconstructed as $\hat{x} = \psi(\hat{c} \cdot \mathbf{c})$.

4 Experiments

In this section, we present the experiments conducted to evaluate our proposed methodology. The experiments are designed to address the following key research questions:

1. **Generalization:** How effectively does the model generalize for classification tasks? Is the concept representation *complete*?
2. **Concept Representation Evaluation:** How much information is captured from c with respect to both the input image x and the label y? Are the learnt concepts *representative* of the data?
3. **Concept Interpretability:** How interpretable are the learnt concepts produced? Are they *aligned* with human representations?
4. **Model Interpretability:** Are the final predictions interpretable in terms of the discovered concepts? Can a user modify the concept prediction to extract counterfactual predictions?

4.1 Experimental setting

In the following we report the dataset, metrics and baselines that we consider for evaluating and comparing our model. We conducted experiments using two backbones g: ResNet-18 and ViT-base-patch32. Instead, we always use a decoder ψ composed by five transposed convolutional layers.

Dataset. This study uses seven image classification datasets of varying complexity. We employ two MNIST [20] variants, *Even/Odd* (digit parity) and *Addition* (paired digits summed as labels). *CIFAR-10* and *CIFAR-100* contain 10 and 100 natural image classes, with models extracting 15 and 20 macro-class concepts, respectively [17]. *Tiny ImageNet* includes 200 classes but is tested on

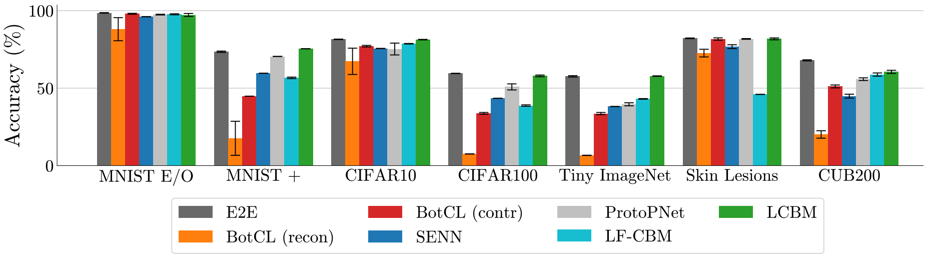


Fig. 4: Comparison of the generalization performance across the evaluated datasets. LCBM consistently provides the highest generalization accuracy across concept-based models, closing the gap with end-to-end black box ones.

30 concepts for added challenge [42]. *Skin Lesions* classifies dermatoscopic images into 4 macro categories [37]. Finally, *CUB-200* [39] covers 200 bird species with species and attribute annotations.

Metrics. We use specific metrics to address each research question. All results are reported with the mean and standard deviation, computed over the test sets by repeating the experiments with three different initialization seeds.

1. **Generalization:** To assess the classification generalization performance, we compute the *Task Accuracy*.
2. **Concept Representation Evaluation:** We employ the *Information Plane* approach [36] to analyze the information retained in the different concept representations. The information plane reports the evolution of the mutual information between the concept representation and both the input x ($I(C, X)$) and the label y ($I(C, Y)$) as the training epoch increases. For models reconstructing the input from the concepts, we assess the *Input Reconstruction Error* by computing the Mean Squared Error (MSE) between the inputs x and their reconstructions $\hat{x}$.
3. **Concept Interpretability:**
For datasets with annotated concepts, we assess their alignment with the learnt concepts using the macro *Concept F1 Score* (best-match approach) and the *Concept Alignment Score (CAS)* [8] for concept representation alignment. Additionally, we conducted a user study with 72 participants, each answering 18 questions. The study evaluated *Plausibility* by asking users to (i) select an image that best represents a given concept and (ii) identify an intruder image among those representing a single concept. It also assessed *Human Understanding* by having participants assign a name to a set of images illustrating a concept. Finally, we provide qualitative insights through *Concept Dictionaries*, showcasing images with the strongest activations for each concept.
4. **Model Interpretability:** We perform *Concept Interventions* [16] to observe how model predictions change when concept predictions are modified. As positive concept interventions are non-trivial in unsupervised concept settings, we perform negative interventions [8]. Negative interventions involve

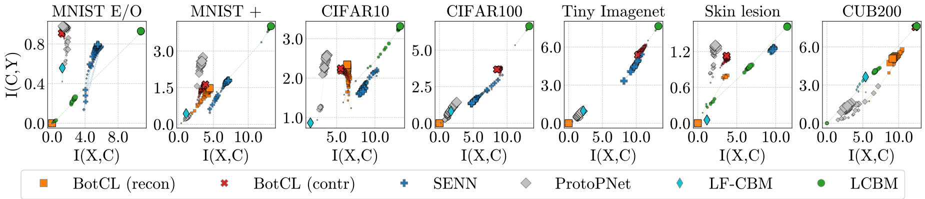


Fig. 5: Information Plane for the different models in terms of Mutual information between concept and input variables $I(X, C)$, and between the concept and output variables $I(C, Y)$. The size of the markers is proportional to the training epoch.

Table 1: We report the Input Reconstruction Error in terms of MSE for those methods that explicitly reconstruct the input.

	MNIST E/O	**MNIST Add.**	**CIFAR10**	**CIFAR100**	**Tiny ImageNet**	**Skin Lesion**	**CUB200**
BotCL	1.40 ± 0.09	1.57 ± 0.03	0.87 ± 0.02	0.82 ± 0.01	1.35 ± 0.06	0.72 ± 0.02	0.07 ± 0.01
SENN	0.62 ± 0.02	0.93 ± 0.03	0.81 ± 0.04	0.74 ± 0.01	1.10 ± 0.02	0.61 ± 0.04	0.05 ±≤0.01
LCBM	**0.32 ±≤0.01**	**0.71 ± 0.11**	**0.51 ±≤0.01**	**0.55 ±≤0.01**	**0.72 ±≤0.01**	**0.32 ±≤0.01**	**0.05 ±≤0.01**

randomly swapping the values of the concept scores with a given probability, expecting model accuracy to decrease as intervention probability increases. For LCBM, to switch a concept to inactive, we set $\hat{c}_j = 0$, while to activate it, we set $\hat{c}_j = 1$ and replace the concept embedding with the concept prototype $\bar{c}_j = \mathbf{t_j}$. Additionally, we provide *Qualitative Explanations* generated by LCBM using as concept importances the predicted weights multiplied by the concept predictions $w_{ij} \cdot \hat{c}_j$.

Baselines. To compare the performance of the proposed approach, we test it against unsupervised approaches like *SENN* [3] and two variants of a SOTA model BotCL [40]: *BotCL (Recon)*, which employs an autoencoder-based approach to reconstruct the input image from the concept bottleneck, and *BotCL (Contr)*, which applies a contrastive term to the loss to encourage distinct concept activations for different classes. Also, we consider a prototype based approach *ProtoPNet* [6] and Label-Free CBM (LF-CBM) [26] a recent hybrid approach. If the concepts were known (e.g., MNIST Addition), we used CLIP to align the model with concept captions (e.g., this image contains the digit 4). If the concepts were unknown, we used the LLM to generate a list of possible concepts. Finally, we compare with a standard black-box model trained end-to-end (E2E).

4.2 Generalization

LCBM is the most accurate interpretable model (Fig. 4). The proposed methodology significantly outperforms the baselines. In the most challenging scenario (Tiny ImageNet with only 30 concepts), it achieves up to a 50% increase in task accuracy compared to the worst baseline (BotCL (recon)) and up to a 17% improvement over the runner-up model, ProtoPNet. Our model consistently delivers the best generalization accuracy across all datasets, with higher gaps in challenging gaps with lower concept-class ratios. This result is valid even when we compare LCBM with LF-CBM which exploit the pre-existing knowledge within a VLM to extract concept annotations. Only in the very simple MNIST Even/Odd dataset a few methods perform better, by a few decimals. We attribute this improvement to the unsupervised concept embeddings, which allows learning more *complete* representations for task prediction.

LCBM closes the gap with black-box models (Fig. 4). Figure 4 also shows that LCBM achieves results comparable to the E2E black-box model. The generalization loss is always less than 1-2%. Notably, on the MNIST addition dataset, a setting where reasoning capabilities over concepts are required, our approach outperforms the black-box model with a task accuracy improvement of 2%. Overall, LCBM demonstrates its capability to achieve high interpretability without sacrificing accuracy in unsupervised concept learning settings.

4.3 Concept Representation Evaluation

LCBM concept representation retains more information regarding both the input and the output (Fig. 5). The concept representation obtained through unsupervised concept embedding is significantly richer than that derived from simple concept scores. As training progresses, most baselines experience a reduction in mutual information with the input $I(X, C)$ while increasing the mutual information with the output $I(Y, C)$. This observation supports the conclusion that unsupervised CBM models tend to lose input-related information while attempting to optimize task performance [34], even for those models that explicitly require concepts to be representative of the input, such as SENN and BotCL (Recon). On the contrary, LCBM overcome this limitation by means of concept embeddings, which facilitate a better balance between competing objectives, as evidenced by the monotonic increase in mutual information with both the input and output during training.

LCBM allows better input reconstruction (Tab. 1). To understand why the mutual information $I(X, C)$ of LCBM is consistently higher compared to other reconstruction-based unsupervised CBMs, we assess the Input Reconstruction Error in terms of MSE. As shown in Table 1, LCBM achieves lower MSE in image reconstruction compared to the runner-up model (usually SENN), with values ranging from 0.18 to 0.38. We believe that concept embeddings facilitate more accurate and efficient reconstruction by allowing more information to flow to the decoder network when a concept is active ($\hat{c}_j = 1$). Unlike other unsupervised models that can only pass a single value ($\hat{c}_j$), LCBM passes the entire associated concept embedding ($\bar{c}_j$) to the decoder.

Table 2: Macro F1-score for candidate concepts with respect to existing human-representations for datasets on which the latter are available.

	MNIST E/O	**MNIST Add.**	**CIFAR100**	**Skin**	**CUB200**
BotCL (recon)	0.47 ± 0.01	0.41 ± 0.01	0.38 ± 0.03	0.47 ± 0.02	0.34± 0.01
BotCL (contr)	0.47 ± 0.02	0.45 ± 0.02	0.40 ± 0.04	0.44 ± 0.03	0.37± 0.02
SENN	0.61 ± 0.02	0.58 ± 0.01	0.44 ± 0.02	0.52 ± 0.02	0.41± 0.02
ProtoPNet	0.26 ± 0.01	0.24 ± 0.01	0.31 ± 0.01	0.16 ± 0.02	0.28± 0.03
LF-CBM	0.52± 0.01	0.50 ± 0.03	0.45 ± 0.01	0.58 ± 0.01	0.45 ± 0.01
LCBM (ours)	**0.88 ± 0.08**	**0.81 ± 0.04**	**0.60 ±≤0.01**	**0.58 ±≤0.01**	**0.55 ±≤0.01**

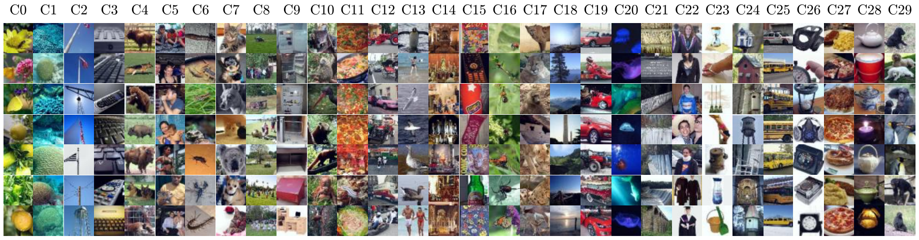


Fig. 6: Tiny-Imagenet dictionary produced by LCBM. Each column of images represents the set of 7 images that mostly activate each concept. Concept numbers are reported on top of each column.

4.4 Concept Interpretability

LCBM concepts are more aligned to human-defined representations (Tab. 2). For datasets with human-defined concept representations, we evaluate the alignment between these representations and those extracted by the compared concept learning methods. In Table 2, we observe that LCBM learns concepts that are significantly more aligned with human-defined representations than existing methods. After matching the concept predictions with the concept annotations using the Hungarian algorithm [18], LCBM achieves an F1 score that is up to +0.36 higher than the runner-up, which is always LF-CBM. We remind, however, that this model has a huge intrinsic advantage, as the employed VLM is prompted to predict the concepts of each datasets. The fact that LCBM without any concept supervision achieves higher concept F1 scores than LF-CBM is impressive, but consistent with recent literature reporting poor LF-CBM concept accuracy [35].

LCBM concepts are qualitatively distinguishable (Fig. 6). While we quantitatively demonstrated that the LCBM concepts align with those of datasets equipped with annotations, for datasets lacking annotations, we examine the dictionaries representing the images that most strongly activate each concept, as proposed in [3]. Fig. 6 presents the dictionary generated by the model for the Tiny ImageNet dataset. Each column (concept) exhibits a recurring and

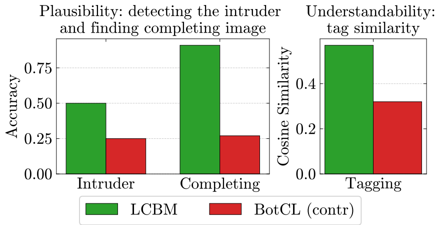


Fig. 7: User study results. Left, user accuracy in detecting the intruder image and the image completing a set of images representing a concept. Right, the similarity of the tags employed by users to describe an extracted concept.

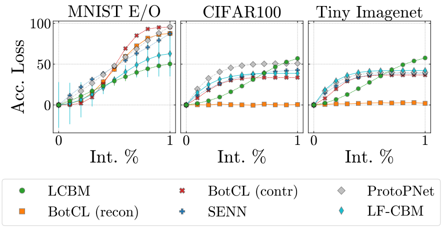


Fig. 8: Negative interventions: percentage accuracy loss when increasing the intervention probability. The higher the accuracy loss, the higher the sensitivity of the model to human interventions.

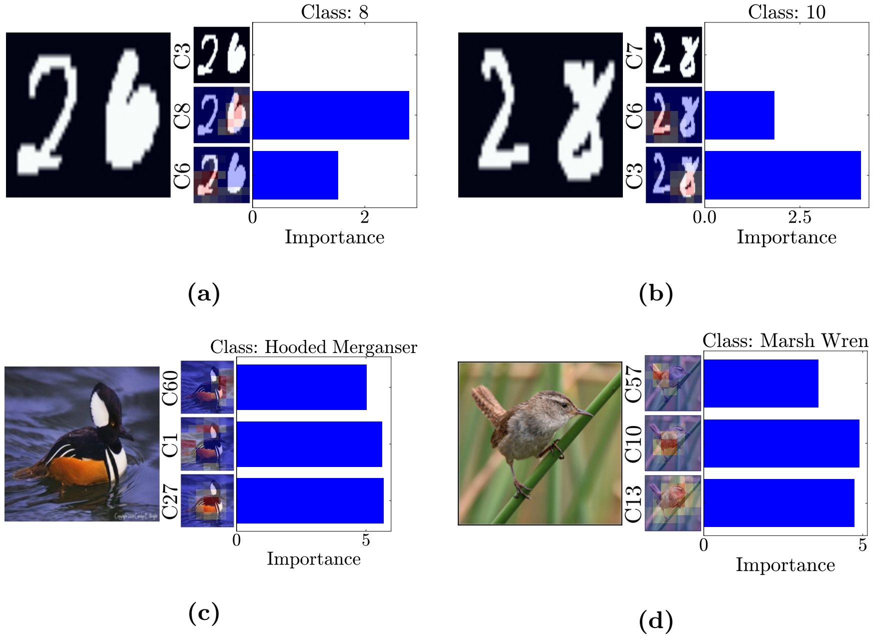
distinguishable pattern. For example, concept $C0$ encompasses images of flowers and plants, whereas $C2$ appears to correspond to long, slender objects. Concept $C4$ includes large mammals such as bison and bears, while $C7$ represents close-up images of small animals.

LCBM concepts are more plausible and understandable to humans (Fig.7). To quantitatively assess the quality of the representations, we conducted a user study comparing the plausibility and human-understandability of the concept extracted by our method and BotCL the SOTA baseline for unsupervised concept learning. Figure 7 shows that LCBM concepts enable users to find the intruder image and to complete the set of images much better than BotCL concepts, with an accuracy up to $+25\%$ in terms of finding the right intruder image and up to $+64\%$ in terms of selecting the completing image. Also, when assessing the understandability of the concepts we see a higher similarity up to $+.35$ of the embeddings of the terms employed to tag the concepts provided by LCBM than those of BotCL. The embeddings are generated using the multilingual sentence encoder “all-MiniLM-L6-v2”.

4.5 Model Interpretability

LCBM is sensitive to concept interventions (Fig. 8). Figure 8 shows that our methodology responds to interventions similarly to other baselines, except MNIST Even-Odd, where the precision drops only to 50%. In all other datasets, concept interventions are effective, with LCBM generally experiencing one of the highest accuracy losses when fully intervened, particularly on CIFAR100 and Tiny ImageNet. This is notable since embedding-based supervised CBMs typically resist interventions and require specialized training [9], whereas LCBM does not, suggesting its potential for more effective interventions in supervised settings.

LCBM provides interpretable predictions (Fig. 9). Finally, we present sample explanations generated by LCBM. As illustrated in Figure 9.a, the image



(a)

(b)

(c)

(d)

Fig. 9: Example of interpretable prediction on different datasets. We provide the concept importance together with the Grad-CAM for the most important concepts.

contains a "two" and a "six," with the model correctly predicting the sum as eight. The model identifies concepts $C6$ and $C8$ as important, which correspond to the learned concepts "two" and "six", and further validated by the Grad-CAM results (shown on the y-axis), highlighting the respective digits in the image. Concept $C3$, which does not appear in the image, has an importance value of 0. Figure 9.c illustrates how an image of a bird is classified as a Hooded Merganser based on a triplet of concepts: $C27$ focuses on the orange wing, $C1$ on the crest extending from the back of the head, and $C60$ on the black beak.

5 Conclusion

This paper introduced a novel unsupervised concept learning model that leverages unsupervised concept embeddings. This approach enables improved generalization accuracy compared to traditional unsupervised Concept-Based Models (CBMs), while also enhancing the representation of concepts. Our experiments demonstrate that the extracted concepts better represent the input data and align more closely with human representations, as evidenced by the F1-score metric, CAS, and the findings from the user study.

Limitations and Future work. The first limitation of this work lies in the employed CNN decoder. While it helps extract meaningful unsupervised concept representations, it struggles to effectively decode the learned concepts. While our model reduces the human effort in understanding learned concepts, some manual inspection is still required. Vision Language Models (VLMs) could help fully automate concept labeling by using representative images, but this approach may be less effective in contexts where VLMs lack knowledge, which is the primary area of application for unsupervised CBMs. Finally, our experiments have been limited to image classification tasks. Extending the model to generative tasks presents a challenge and could be explored in future work.

References

1. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence **5**(9), 1006–1019 (2023)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
3. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems **31** (2018)
4. Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., Melacci, S.: Entropy-based logic explanations of neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 6046–6054 (2022)
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
6. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
7. Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., Melacci, S.: Logic explained networks. Artificial Intelligence **314**, 103822 (2023)
8. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept embedding models: Beyond the accuracy-explainability trade-off. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 21400–21413. Curran Associates, Inc. (2022)
9. Espinosa Zarlenga, M., Collins, K., Dvijotham, K., Weller, A., Shams, Z., Jamnik, M.: Learning to receive help: Intervention-aware concept embedding models. Advances in Neural Information Processing Systems **36** (2024)
10. Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., Serre, T.: Craft: Concept recursive activation factorization for explainability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2711–2721 (2023)
11. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems **32** (2019)

12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
13. Hase, P., Chen, C., Li, O., Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 7, pp. 32–40 (2019)
14. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
15. Kim, E., Jung, D., Park, S., Kim, S., Yoon, S.: Probabilistic concept bottleneck models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
16. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
17. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
18. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
19. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 131–138 (2019)
20. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
21. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
22. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: International Conference on Learning Representations (2022)
23. Marconato, E., Passerini, A., Teso, S.: Glancenets: Interpretable, leak-proof concept-based models. Advances in Neural Information Processing Systems **35**, 21212–21227 (2022)
24. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review **63**(2), 81 (1956)
25. Misino, E., Marra, G., Sansone, E.: Vael: Bridging variational autoencoders and probabilistic logic programming. Advances in Neural Information Processing Systems **35**, 4667–4679 (2022)
26. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FlCg47MNvBA
27. Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., et al.: The role of explainable ai in the context of the ai act. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1139–1150 (2023)
28. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., Baralis, E.: Concept-based explainable artificial intelligence: A survey. arXiv preprint arXiv:2312.12936 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

30. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
31. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
32. Sawada, Y., Nakamura, K.: Concept bottleneck model with additional unsupervised concepts. IEEE Access **10**, 41758–41765 (2022)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810 (2017)
35. Srivastava, D., Yan, G., Weng, L.: Vlg-cbm: Training concept bottleneck models with vision-language guidance. Advances in Neural Information Processing Systems **37**, 79057–79094 (2024)
36. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
37. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
38. Veale, M., Zuiderveen Borgesius, F.: Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International **22**(4), 97–112 (2021)
39. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
40. Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Learning bottleneck concepts in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10962–10971 (2023)
41. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)
42. Yao, L., Miller, J.: Tiny imagenet classification with convolutional neural networks. CS 231N **2**(5), 8 (2015)
43. Yeh, C.K., Kim, B., Arik, S., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. Advances in neural information processing systems **33**, 20554–20565 (2020)
44. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data (2022), `https://openreview.net/forum?id=HAMeOIRD_g9`
45. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)

46. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8827–8836 (2018)