Sparse Linear Concept Discovery Models

Konstantinos P. Panousis[1,3,4,5] Dino Ienco[2,3,5] Diego Marcos[1,3,4,5]

[1]Inria [2]Inrae [3]University of Montpellier [4]LIRMM [5]UMR-Tetis

{konstantinos.panousis@inria.fr, diego.marcos}@inria.fr dino.ienco@inrae.fr

Abstract

The recent mass adoption of DNNs, even in safety-critical scenarios, has shifted the focus of the research community towards the creation of inherently intrepretable models. Concept Bottleneck Models (CBMs) constitute a popular approach where hidden layers are tied to human understandable concepts allowing for investigation and correction of the network's decisions. However, CBMs usually suffer from: (i) performance degradation and (ii) lower interpretability than intended due to the sheer amount of concepts contributing to each decision. In this work, we propose a simple yet highly intuitive interpretable framework based on Contrastive Language Image models and a single sparse linear layer. In stark contrast to related approaches, the sparsity in our framework is achieved via principled Bayesian arguments by inferring concept presence via a data-driven Bernoulli distribution. As we experimentally show, our framework not only outperforms recent CBM approaches accuracy-wise, but it also yields high per example concept sparsity, facilitating the individual investigation of the emerging concepts. Our code and models are available at: https://github.com/konpanousis/ConceptDiscoveryModels.

1. Introduction

Deep Neural Networks (DNNs) have been established as the de-facto state-of-the-art approach for a variety of domains and applications. Their performance has facilitated their widespread adoption, especially in CV and NLP, including safety-critical tasks such as autonomous driving and healthcare. However, due to their highly complex structure, DNNs are considered *black-box* models: they map an input to an output via an *un-interpretable* computation process. This constitutes a highly undesirable property, especially in safety- or bias-aware domains, where trustworthiness via the interpretation of the decision making process is key. Thus, conceiving *inherently* interpretable networks constitutes a crucial research and societal challenge.

One of the best known approaches in this context, is Concept Bottleneck Models (CBMs) [7]. CBMs commonly comprise two basic structures: (i) a Concept Bottleneck Layer (CBL) trained to tie its neurons to human interpretable *concepts*, e.g., textual descriptions, followed by (ii) a *linear* decision layer that facilitates the interpretability of the decision process since it is now based on an affine combination of the learned concepts. Despite this more interpretable mode of operation, CBMs suffer from three significant drawbacks: (i) need for labeled data for the predefined concepts, (ii) performance degradation compared to a standard neural backbone, and (iii) rely on *implicit* interpretation of the contribution of each concept, to the final decision, through the analysis of the last linear layer weights.

In this work, we aim to address the limitations of current CBMs by introducing a novel framework for interpretable neural networks based on: (i) recent advances in CLIP-based models, (ii) a single linear decision layer, and (iii) a novel per example *explicit* concept discovery and sparsification mechanism that builds upon solid Variational Bayesian arguments. We dub our approach Concept Discovery Models (CDMs). As we experimentally show, our framework significantly outperforms recent CBM-based SOTA alternatives accuracy-wise, while giving rise to a principled data-driven mechanism for discovering a *highly flexible* and *highly sparse* set of concepts for each example.

2. Related Work

Concept Bottleneck Models & Sparsity. The most similar approach to ours is Concept Bottleneck Models (CBMs) [7]. CBMs have facilitated recent developments towards interpretable architectures, with many methods aiming to alleviate their drawbacks, e.g., performance degradation. Post-hoc CBMs [16] constitute such an extension: any backbone is made interpretable through training a single FC layer, while optionally performing residual fitting to restore any performance loss. More recently, Label-Free (LF) CBM [10] was introduced, an "automated" CBM with a sparse linear prediction layer, in which four steps are considered: (i) automatic concept creation using GPT and fil-

tering, (ii, iii) computation of the CBL projection weights through the CLIP-based concept similarity matrix [13], and (iv) training the sparse final layer. Despite these advances, most works [15, 10, 9] resort to complicated schemes for separately training the CBL and the linear layer, relying on impromptu constraints that may harm interpretability. Specifically, the sparse layer is trained post-hoc, using custom solvers that require ad-hoc application-specific sparsity or accuracy thresholds, despite the existence of more data-driven approaches in the literature.

In this context, recent advances in Variational Bayesian methods towards pruning or component omission [1, 11, 12] have paved the way for a principled data-driven and end-to-end trainable sparsity mechanism: auxiliary binary latent variables are introduced to explicitly model the presence or absence of network components in an "on"-"off" fashion, without requiring any ad-hoc thresholds. We exploit this rationale and construct a principled framework that explicitly *infers* the per-example concept presences, allowing for varying and unrestricted flexibility.

3. Proposed Approach

Let us denote by $\boldsymbol{X} \in \mathbb{R}^{N\times H\times L\times c}$, a dataset comprising N images, each with height H, width L and c channels, and by $\boldsymbol{A} = \{a_1, a_2, \ldots, a_M\}$ a predefined set of attributes/concepts, where $M = |\boldsymbol{A}|$ denotes the dimensionality of the set, i.e., the number of concepts.

Image-Language models, e.g., CLIP [13], typically comprise an image encoder $E_I(\cdot)$ and a text encoder $E_T(\cdot)$; these are jointly trained in a contrastive manner to learn a *common embedding space* [2, 14]. During inference, we first project images and text to this common space; therein, we can compute their similarity using these (ℓ_2-normalized) *embeddings*. The cosine similarity is usually considered, computed via the inner product:

$$\text{Cos Similarity} \triangleq \boldsymbol{S} \propto E_I(\boldsymbol{X})E_T(\boldsymbol{A})^T \in \mathbb{R}^{N\times M} \quad (1)$$

Since $\boldsymbol{S}$ is computed between all possible images-concepts pairings, it yields a unique representation for each image, encoded via the similarity with each distinct concept; thus, these per-image representations can be naturally employed to support a downstream task.

We consider classification using a single linear layer comprising a weight matrix $\boldsymbol{W}_c \in \mathbb{R}^{C\times M}$, where C is the number of classes. During training and since we use the similarity value as an input, $\boldsymbol{W}_c$ will learn to encode how each concept relates to each particular class. The output of the network $\boldsymbol{Y} \in \mathbb{R}^{N\times C}$ yields:

$$\boldsymbol{Y} = \boldsymbol{S}\boldsymbol{W}_c^T \propto \left(E_I(\boldsymbol{X})E_T(\boldsymbol{A})^T\right)\boldsymbol{W}_c^T \quad (2)$$

In Eq. (2), we *linearly* combine all concept-related information and compute a class probability *for each image*. Conversely to related CBMs approaches that rely on complicated projections, we posit that the CLIP similarity vector presents a sufficient images-concepts representation without the need for any additional computational overheads. We then use the standard cross-entropy loss for classification. A graphical illustration is depicted in Fig. 1 (Left).

However, the commonly used linear decision layer has a significant drawback: the relation between images and concepts is *implicit*. Indeed, most approaches rely on the magnitudes of $\boldsymbol{W}_c$ and the projections to assess the effect of *all* concepts, without considering information redundancy. Recent approaches [15, 10], aim to alleviate this issue by sparsifying the linear layer *for each class*, leveraging however complicated solvers that require tuning of an ad-hoc task specific cut-off thresholds. We posit that restricting the concepts per class to a fixed set, greatly limits the flexibility of appropriate per-example concept representation.

To bypass these limitations, and in stark contrast to other sparsity-inducing approaches, we propose a novel, data-driven formulation for inferring the essential number of concepts present on a per-example basis. To this end, we introduce a set of auxiliary *binary latent indicators* $\boldsymbol{Z} \in \{0,1\}^{N\times M}$; these denote whether a particular concept considered for each example in an "on"-"off" fashion; that is, $z_{n,m} = 1$ if concept m is active for example n, $z_{n,m} = 0$ otherwise. Thus, instead of only relying on *implicit* measures, we have now defined an *explicit* mechanism of concept presence. The output (Eq.(2)) now reads:

$$\boldsymbol{Y} = (\boldsymbol{Z}\cdot\boldsymbol{S})\boldsymbol{W}_c^T \quad (3)$$

In this case, the output for each image is facilitated via the inner product computation between the weights of the linear layer and the *effective concepts* as dictated by the introduced latent indicators $\boldsymbol{Z}$.

A naive construction of the latent indicators $\boldsymbol{z}_i \in \{0,1\}^M$, $i = 1, \ldots, N$, for each example would result in: (i) significant computational overhead for storing each indicator, and (ii) no evident way to generalize the learned indicators to unseen examples. On this basis, we draw inspiration from recent Variational Bayesian advances [1, 11] and postulate that the indicators $\boldsymbol{Z}$ are obtained via a data-driven random sampling procedure. This translates to drawing samples from Bernoulli distributions, where the probability of concept presence is driven from a separate linear computation between the image embedding and a learnable weight matrix $\boldsymbol{W}_s \in \mathbb{R}^{K\times M}$, K being the dimensionality of the embedding space; thus, for each input $\boldsymbol{X}_i$, this yields:

$$q(\boldsymbol{z}_i) = \text{Bernoulli}\Big(\boldsymbol{z}_i|\text{sigmoid}\Big(E_I(\boldsymbol{X}_i)\boldsymbol{W}_s^T\Big)\Big) \quad (4)$$

where the $\text{sigmoid}(x) = 1/(1+e^{-x})$ nonlinearity is applied to convert the linear computation to probability.

Thus, instead of only relying on the implicit relation between images and concepts learned using the distance of

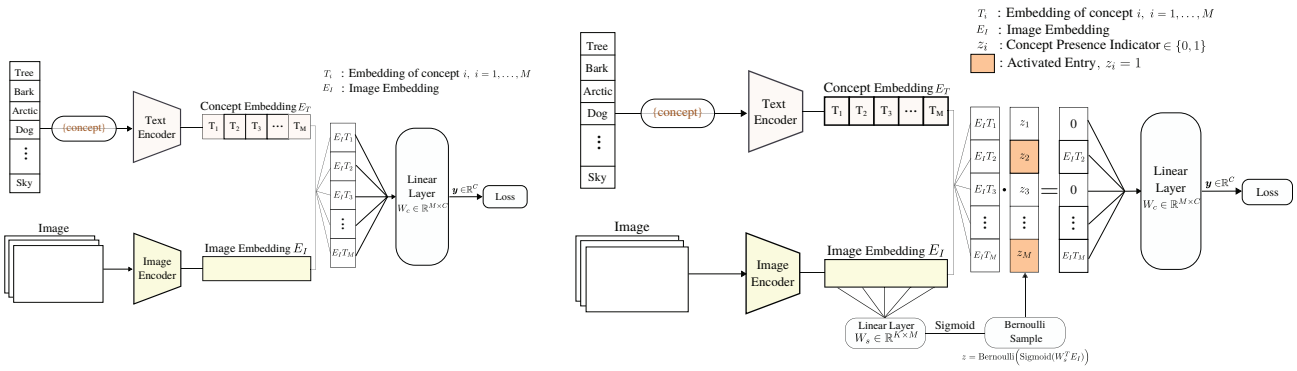

Figure 1: (Left) The base model utilizes the similarities between the images-concepts CLIP embeddings to perform classification with a single linear layer. (Right) The proposed data-driven concept discovery framework. In this case we exploit the information of the image embeddings to devise a mechanism for explicit concept presence indication per example.

their embeddings, i.e., the similarity, and the linear layer, we exploit a separate source of information to devise an *explicit* mechanism indicating concept presence in the context of the downstream task. This *amortized* formulation bypasses both the aforementioned complications: we only need to store a single matrix with dimensions $K \times M$, while at the same time allowing for generalization to unseen examples. A graphical illustration of the envisioned architecture is depicted on Fig. 1 (Right).

Training & Prediction. Assuming a dataset $\mathcal{D} = \{(\boldsymbol{X}_i, \hat{Y}_i)\}_{i=1}^N$ and a concept set $\boldsymbol{A} = \{a_1, \ldots, a_M\}$, the core training objective is the cross-entropy loss, denoted as $\text{CE}(\hat{Y}_i, f(\boldsymbol{X}_i, \boldsymbol{A}))$, where $f(\boldsymbol{X}_i, \boldsymbol{A}) = \text{Softmax}(\boldsymbol{y}_i)$ are the class probabilities; $\boldsymbol{y}_i$ is computed via Eqs.(2), (3). Since the base model comprises only the weight matrix $\boldsymbol{W}_c$, it can be trained only via the cross entropy signal.

When using the concept presence mechanism, the introduction of the binary latent indicators $\boldsymbol{Z}$, necessitates a different treatment of the training objective. In line with related literature [11], we adopt the stochastic gradient Variational Bayes (SGVB) framework [6] for scalability. In this context, we impose an appropriate prior distribution for the latent indicators $\boldsymbol{z}_i$, i.e., a Bernoulli distribution, s.t., $p(\boldsymbol{z}_i) = \text{Bernoulli}(\alpha)$, $\forall i$, where α is a fixed non-negative constant. The resulting objective takes the form of an Evidence Lower Bound (ELBO) [4]:

$$\mathcal{L} = \sum_{i=1}^N \text{CE}(\hat{Y}_i, f(\boldsymbol{X}_i, \boldsymbol{A}, \boldsymbol{z}_i)) - \beta \text{KL}\left(q(\boldsymbol{z}_i) || p(\boldsymbol{z}_i)\right) \quad (5)$$

where we augmented the notation to reflect the dependence on the binary indicators $\boldsymbol{Z}$ and β is a scaling factor[3]. The second term is the Kullback-Leibler divergence; this encourages the posterior to be close to the prior. Thus, by setting α to a value close to zero, we can effectively enforce a sparsity-inducing behavior in the learning process, while striking a balance between accuracy and sparsity without relying on any ad-hoc application or task specific thresholds.

For training the model, we perform Monte Carlo sampling to estimate Eq. (5) using a single *reparameterized* sample. Since the Bernoulli distribution is not amenable to the reparameterization trick[6], we turn to its continuous relaxation [5, 8] for training. During inference, we can directly draw samples from the trained posteriors $q(\boldsymbol{Z})$ to investigate the per-example effect of a concept. For computing the per-class relevance, one can average the concept presence probability of all the examples of the class.

At this point, it is important to highlight the flexibility of the proposed framework; instead of forcing an artificial sparsity on each class, we enable the model to learn the relevant concepts for each image in a data-driven manner, allowing for a varying number of activated concepts.

4. Experimental Setup

Datasets, Concepts Sets & CLIP Backbones. For thoroughly evaluating the proposed framework, we consider 5 datasets with varying characteristics: (i, ii) CIFAR10/100, (iii) CUB, (iv) Places365, and (v) ImageNet-1k. CIFAR-10/100 are standard recognition benchmarks, comprising 32×32 images, while CUB comprises higher resolution images focusing on fine-grained bird species identification. Their sizes also greatly vary, with CUB comprising 5900 training examples, and Places365/ImageNet up to $1 - 2$ million. This highly diverse set of tasks will serve as an important showcase of the performance of the introduced mechanism. We consider the same concepts sets as in [10] for comparability, comprising 128, 824, 211, 2202 and 4505 concepts for CIFAR-10, CIFAR-100, CUB, Places365 and ImageNet, respectively. Finally, for the backbones of CLIP, we use the most common, i.e., ResNet50 (RN50) and ViT-B/16, which are frozen when computing the similarities. For our experiments, we set $\alpha = \beta = 10^{-4}$; we select the best performing learning rate among $\{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$ for the linear layer. We set a higher learning rate for $\boldsymbol{W}_s$ ($10\times$) to facilitate learning of the discovery mechanism

Model	Dataset (Accuracy (%) ‖ Sparsity (%))				
	CIFAR10	CIFAR100	CUB200	Places365	ImageNet
Standard [10]†	88.80	70.10	76.70	48.56	76.13
Standard (sparse) [10]†	82.96	58.34	75.96	38.46	74.35
Label-Free [10]†	86.37	65.27	74.59	43.71	71.98
CDM (RN50, w/o $\boldsymbol{Z}$)	81.90 ‖ −−	63.40 ‖ −−	64.70 ‖ −−	52.90 ‖ −−	71.20 ‖ −−
CDM (RN50, w/ $\boldsymbol{Z}$)	86.50 ‖ 2.55	67.60 ‖ 9.30	72.26 ‖ 21.3	**52.70** ‖ 8.28	72.20 ‖ 8.53
CDM (ViT-B/16, w/o $\boldsymbol{Z}$)	94.45 ‖ −−	79.00 ‖ −−	75.10 ‖ −−	54.40 ‖ −−	77.90 ‖ −−
CDM (ViT-B/16, w/ $\boldsymbol{Z}$)	**95.30** ‖ 1.69	**80.50** ‖ 3.38	**79.50** ‖ 13.4	52.58 ‖ 8.00	**79.30** ‖ 6.96

Table 1: Accuracy and Sparsity Results. By bold we note the best performing *sparse* model. †indicates the reported performance. "Standard" models correspond to the non-interpretable backbones used in [10].

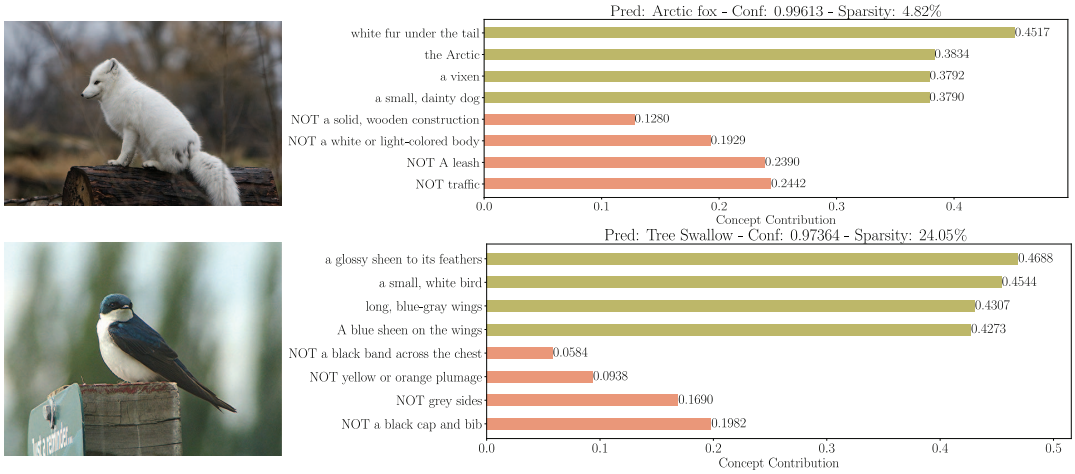

Figure 2: Concept Discovery investigation for an example from ImageNet-1k (Up) and CUB200 (Down). Khaki denotes positive contributions to the decision, while red negative.

and train the models for a maximum of 2000 epochs for CIFAR10-100/CUB and 300 for ImageNet/Places365.

Our main competitor is Label Free-CBMs [10]. Even though it constitutes a highly different approach, in all settings, CLIP is used to compute similarities between images and concepts; then, instead of using them directly as in our framework, they are used to train a CBL. To achieve this, they use as backbones: (i) for CIFAR10/100, the CLIP RN50 encoder, (ii) for CUB, a CUB-trained RN18, and (iii) for Places/ImageNet, an ImageNet-trained RN50. After learning the projections, the GLM-SAGA solver [15] is used for learning a sparse linear layer, reporting $0.7 - 15\%$ non-zero weights without specifying the per class results.

Results. The obtained comparative results are depicted in Table 1. Therein, the sparsity values for our framework denotes the dataset-wise sparsity computed by averaging the per-example number of activated concepts over all samples. We observe that our concept discovery framework allows for extremely low concept retention while often it *improves* the accuracy compared to the base model defined in Eq.(2). Compared to related methods where sparsity is arbitrarily enforced on a class-wise level in an ad-hoc manner, in CDMs, the presence of a concept is inferred end-to-end on a data-driven per example basis. This facilitates balancing the trade-off between accuracy and sparsity during the learning process for each different example, greatly enhancing the flexibility of the framework. In Fig. 2, graphical illustrations of per-example discovered concepts are presented. The upper figure corresponds to an ImageNet *test* sample and the lower from CUB. For the former, we observe a concept retention rate (sparsity) of 4.82%, translating to approximately 217 active concepts out of the 4505 potential concepts, with the top four most contributing being semantically similar to the example image. On the other hand, for the CUB example, we observe that the framework inferred a potentially necessary higher retention rate of 24.05%, translating to around 53 out of 211 active concepts, nevertheless exhibiting semantically similar most contributing concepts.

5. Conclusions

In this work, we proposed a novel framework towards intepretable networks based upon: (i) image-concept similarities arising from CLIP models, (ii) a linear layer for classification, and (iii) a novel data-driven mechanism for per-example concept discovery. The experimental results vouch for the efficacy of the approach. Our CDMs retain or even improve classification performance, while at the same time enabling very high per-example sparsity without limiting the flexibility of the concept selection mechanism.

References

[1] Sotirios P. Chatzis. Indian buffet process deep generative models for semi-supervised classification. In *Proc. ICASSP*, 2018. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 2020. 2

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proc. ICLR*, 2017. 3

[4] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013. 3

[5] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumbel-softmax. In *Proc. ICLR*, 2017. 3

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 3

[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proc. ICML*, 2020. 1

[8] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. ICLR*, 2017. 3

[9] Diego Marcos, Ruth Fong, Sylvain Lobry, Rémi Flamary, Nicolas Courty, and Devis Tuia. Contextual semantic interpretability. In *Proc. ACCV*, 2020. 2

[10] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *Proc. ICLR*, 2023. 1, 2, 3, 4

[11] Konstantinos Panousis, Sotirios Chatzis, and Sergios Theodoridis. Nonparametric Bayesian deep networks with local competition. In *Proc. ICML*, 2019. 2, 3

[12] Konstantinos P Panousis, Anastasios Antoniadis, and Sotirios Chatzis. Competing mutual information constraints with stochastic competition-based activations for learning diversified representations. In *Proc. AAAI*, 2022. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 2

[14] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. NIPS*, 2016. 2

[15] Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *Proc. ICML*, 2021. 2, 4

[16] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. 1