Supplemental Material: 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

A. Limitations & Future Work

One limitation of the proposed framework is the dependence on a CLIP-like backbone to obtain the images-concepts similarities. On this basis: (i) there is no "easy" way to recover from the backbone's *concept omissions*. Indeed, if the image-text model assigns a large similarity value to a particular unrelated concept, this can be removed via the concept discovery mechanism. However, if the backbone assigns zero similarity between an image and a given concept, despite the latter being present in the image, it will not contribute to the downstream task. (ii) The results depend on the suitability of the backbone to the considered application; thus, if the backbone can not adequately model the underlying data due to either its architecture or concepts missing from (or biases contained in) the data used for pretraining, the final performance will reflect that, even if the introduced CDM framework somewhat alleviates this issue via the concept discovery mechanism. In this context, even though the experimental results suggest that using the ViT-B CLIP backbone can yield significant performance, it may not work in all cases. However, the proposed framework constitutes a general proposal: any future advances on multi-modal models can be easily incorporated by changing the projection backbone. In our future work, we aim to lessen the dependence on the pretrained backbones and find ways to either adjust the arising similarities or combine different or multiple image and text encoders to match the downstream task.

B. Bernoulli Relaxation & Inference

Training. As already noted in the main text, to estimate the ELBO in Eq. (5):

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

we perform Monte-Carlo sampling, with a single reparameterized sample. However, the Bernoulli distribution is not amenable to the reparameterization trick [2]. To this end, we resort to its continuous relaxation[3, 1].

Let us denote by $\tilde{\boldsymbol{z}}_{\boldsymbol{i}}$, the probabilities of $q(\boldsymbol{z}_i)$, $i = 1, \dots N$. We can directly draw reparameterized samples $\hat{\boldsymbol{z}}_{\boldsymbol{i}} \in (0,1)^M$ from the continuous relaxation as:

$$\hat{\boldsymbol{z}}_{\boldsymbol{i}} = \frac{1}{1 + \exp\left(-(\log \tilde{\boldsymbol{z}}_{\boldsymbol{i}} + L)/\tau\right)} \quad (2)$$

where $L \in \mathbb{R}$ denotes samples from the Logistic function, such that:

$$L = \log U - \log(1-U), \quad U \sim \text{Uniform}(0,1) \quad (3)$$

where τ is called the *tempetature* parameter; this controls the degree of the approximation: the higher the value the more uniform the produced samples and vice versa. We set τ to 0.1 in all the experimental evaluations.

Inference. During inference, and for each test example $\boldsymbol{X}$, we draw sample(s) from the Bernoulli distribution defined in Eq. (4):

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

to obtain the binary indicator vector $\boldsymbol{z} \in \{0,1\}^M$: each entry therein denotes the presence or absence of a concept for the given example. This is used to: (i) compute the output of the network according to Eq. (3):

$$\boldsymbol{Y} = (\boldsymbol{Z} \cdot \boldsymbol{S})\boldsymbol{W}_c^T \quad (5)$$

and subsequently the loss function (in our case the cross-entropy), and (ii) examine each concept activated for the given example.

C. Ablation Study.

For learning the auxiliary binary latent variables $\boldsymbol{Z}$, we introduced appropriate prior and posterior distributions and constructed the ELBO. In this context, we introduced two additional hyperparameters: (i) the prior parameter α and (ii) the scale of the KL divergence, β. Here, we examine

the effect of these parameters of the final performance using the ViT-B/16 backbone and the CUB dataset and two different learning rates 10^{-2} and 10^{-3}. In Table 1, we report the performance of the framework in terms of accuracy and sparsity for different values of α, β.

α	β	Accuracy (%)	Sparsity (%)
10^{-2}	10^{-4}	80.67	23.38
10^{-4}	10^{-4}	80.00	16.12
10^{-4}	10^{-5}	79.70	14.07
10^{-3}	10^{-4}	82.23	37.7
10^{-3}	$5 \cdot 10^{-4}$	81.40	20.93
$5 \cdot 10^{-4}$	10^{-3}	81.07	17.61

Table 1: Ablation results on the impact of the hyperparameters on: (i) the resulting accuracy and (ii) the emerging sparsity using the ViT-B/16 backbone for CLIP and CUB200 as the training dataset. The learning rate in this study was set to $5 \cdot 10^{-3}$ for the top table and 10^{-3} for the bottom table respectively.

References

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumbel-softmax. In *Proc. ICLR*, 2017. 1

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 1

[3] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. ICLR*, 2017. 1