Coarse-to-Fine Concept Bottleneck Models

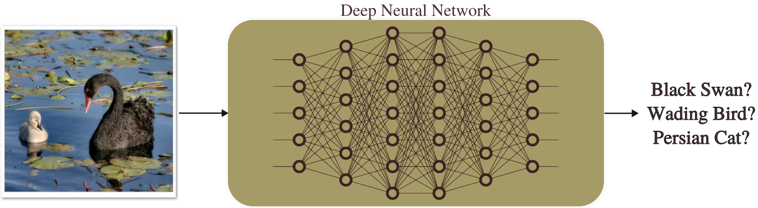
Konstantinos P. Panousis, Dino Ienco, Diego Marcos

konstantinos.panousis@inria.fr
dino.ienco@inrae.fr
diego.marcos@inria.fr

NeurIPS 2024

Motivation: Interpretable Models

- DNNs are treated like *black-box models*:
  - Given an input, the model takes a decision via an **un-interpretable decision process.**
  - The model complexity, generally, **hinders any potential examination** of the underlying process.

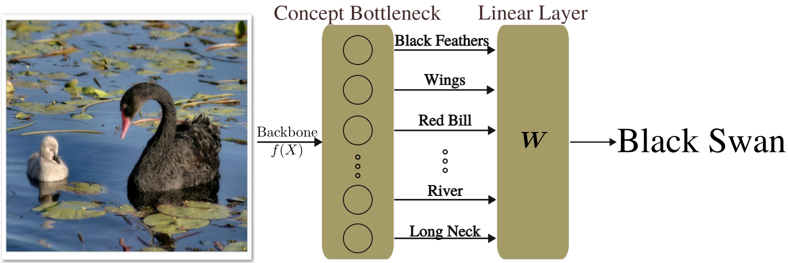


Lack of interpretability

Undesired property, especially in safety- or bias- aware applications $\Rightarrow$ Crucial research and societal challenge

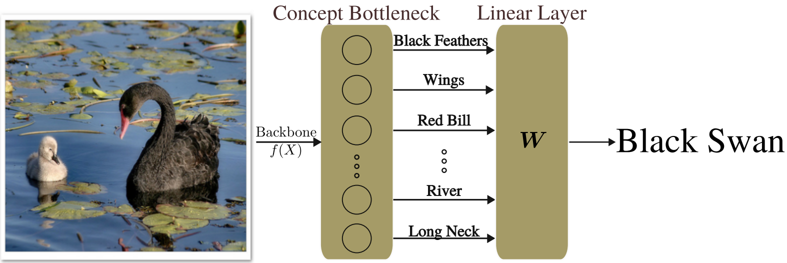
Related Work: Concept Bottleneck Models

- Ante-hoc methods: Design models that are inherently *interpetable*, e.g., Concept Bottleneck Models (CBMs) [1].



Related Work: Concept Bottleneck Models

- Ante-hoc methods: Design models that are inherently *interpetable*, e.g., Concept Bottleneck Models (CBMs) [1].



- Several drawbacks:
  1. **Performance degradation** compared to standard backbones,
  2. Use of **dense concept sets**, all potentially contributing to the final decision,
  3. Not suited for tasks that could **exploit multi-granularity** information.

Coarse-to-Fine Concept Bottleneck Models

Our proposal: Coarse-to-Fine Concept Bottleneck Models

- A **hierarchical** approach to concept discovery.
- We consider a **per-example discovery mechanism** to limit concepts associated to each example, and
- Leverage the notion of **concept hierarchy** to uncover both high and low level image information.

Our proposal: Coarse-to-Fine Concept Bottleneck Models

- A **hierarchical** approach to concept discovery.
- We consider a **per-example discovery mechanism** to limit concepts associated to each example, and
- Leverage the notion of **concept hierarchy** to uncover both high and low level image information.

Proposed Approach: Concept Discovery Model Block

- Extract image and concept embeddings, $E_I(\boldsymbol{X}) \in \mathbb{R}^{N \times K}$ and $E_T(\boldsymbol{A}) \in \mathbb{R}^{H \times K}$ with CLIP,
- Compute :

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

- Adopt data-driven binary indicators $\boldsymbol{Z} \in \{0,1\}^{N \times H}$ to select a concept subset.
- Classify using $Y = (\boldsymbol{Z} \cdot \boldsymbol{S})\boldsymbol{W}_c^T$

Proposed Approach: Concept Discovery Model Block

- Extract image and concept embeddings, $E_I(\boldsymbol{X}) \in \mathbb{R}^{N \times K}$ and $E_T(\boldsymbol{A}) \in \mathbb{R}^{H \times K}$ with CLIP,
- Compute :

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

- Adopt data-driven binary indicators $\boldsymbol{Z} \in \{0, 1\}^{N \times H}$ to select a concept subset.
- Classify using $Y = (\boldsymbol{Z} \cdot \boldsymbol{S}) \boldsymbol{W}_c^T$

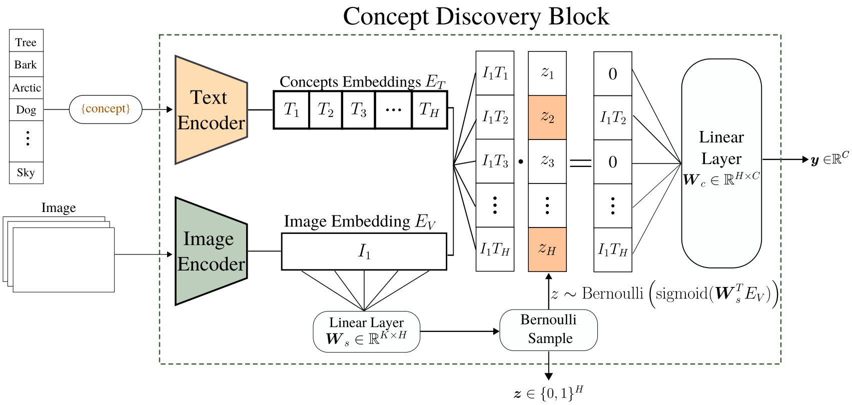
Implementation:

- $\boldsymbol{Z}$ are obtained via a data-driven random sampling procedure:
  - *Amortized formulation*: introduce a learnable weight matrix $\boldsymbol{W}_s \in \mathbb{R}^{K \times H}$ and use the image embeddings $E_I(\boldsymbol{X}_i)$ to drive the process:

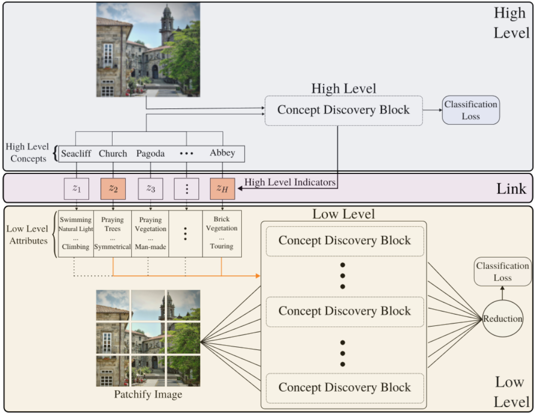
$$q(\boldsymbol{z}_i) = \text{Bernoulli}\Big(\boldsymbol{z}_i | \text{sigmoid}\Big(E_I(\boldsymbol{X}_i) \boldsymbol{W}_s^T\Big)\Big) \tag{2}$$

  - Advantages: (i) we only store $\boldsymbol{W}_s$, and (ii) can generalize to unseen examples.

Concept Discovery Model Block



Coarse-to-Fine Concept Bottleneck Models



- Discover concepts that describe the whole image, while exploiting information residing in patch-specific regions.

Coarse-to-Fine Concept Bottleneck Models

1. Discover **high-level concepts for the whole image** using a concept set A_H and indicators $\boldsymbol{Z}_H \in \{0,1\}^{N \times H}$. Each high level concept is described by L low-level attributes.
2. Discover the essential **low-level concepts in the context of sub-regions** of the image using the concept set A_L and indicators $\boldsymbol{Z}_L \in \{0,1\}^{N \times P \times H \cdot L}$.
3. Having discovered which high level concepts are active, we can now **further mask** the low-level concepts, i.e., **zero-out** the ones that are irrelevant, in a top-down way.
4. To formalize the linkage:

$$[Z]_{n,p} \propto \sum_h [Z_H]_{n,h} \cdot [Z_L]_{n,p,h,:} \in \{0,1\}^L \tag{3}$$

CF-CBM: Experimental Results - Accuracy

- **Training**:
  - Evidence Lower Bound (ELBO) via Stochastic Gradient Variational Bayes.
  - A_H equals the set of classes and A_L the available per-class attributes.
- **Inference**: Draw samples from the learned posterior and investigate the values of $\boldsymbol{Z}$.
- **Evaluation metrics**: Accuracy and Sparsity (Average Percentage of Activated concepts)

CF-CBM: Experimental Results - Accuracy

- **Training**:
  - Evidence Lower Bound (ELBO) via Stochastic Gradient Variational Bayes.
  - A_H equals the set of classes and A_L the available per-class attributes.
- **Inference**: Draw samples from the learned posterior and investigate the values of $\boldsymbol{Z}$.
- **Evaluation metrics**: Accuracy and Sparsity (Average Percentage of Activated concepts)

Architecture Type	Model	Concepts	Sparsity	Dataset (Accuracy (%) || Sparsity (%)) CUB	SUN	ImageNet
Non-Interpretable	Baseline (Images)	✗	✗	76.70	42.90	76.13
	CLIP Embeddings^H	✗	✗	81.90	65.80	79.40
	CLIP Embeddings^L	✗	✗	47.80	46.00	62.85
Concept-Based Whole Image High Level	Label-Free CBMs [2]	✓	✓	74.59	–	71.98
	CDM^H[3]	✓	✗	80.30	66.25	75.22
	CDM^H[3]	✓	✓	78.90||19.00	**64.55**||13.00	76.55||14.00
	CF-CBM^H (Ours)	✓	✓	**79.50**||50.00	64.00||47.58	**77.40**||27.20
Concept-Based Patches Low Level	CDM^L	✓	✗	39.05	37.00	49.20
	CDM^L	✓	✓	59.62||58.00	42.30||67.00	58.20||25.60
	CF-CBM^L (Ours)	✓	✓	**73.20**||29.80	**57.10**||28.33	**78.45**||15.00

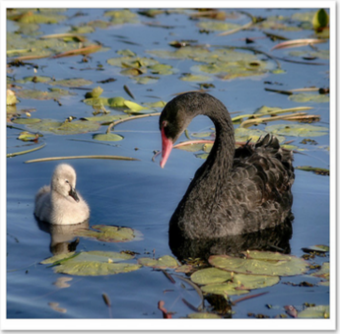
CF-CBM: Experimental Results - Attribute Matching

- Classification performance is not appropriate for measuring interpretability.
- A new metric for interpretability in the context of concept-based methods given ground truth attributes: Jaccard Index.

Attribute matching accuracy. We compare our approach to the recent CDM model trained with the considered A_L set. Then, we predict the matching between the inferred per-example concept indicators to: (i) class-wise and (ii) per-example ground truth attributes found in both SUN and CUB.

			Dataset (Matching Accuracy (%)|| Jaccard Index (%))	
Model	Attribute Set Train	Atrribute Set Eval	SUN	CUB
CDM[3]	whole set	class-wise	51.43||26.00	39.00||17.20
CDM^{L}	whole set	class-wise	30.95||26.70	25.81||19.60
CF-CBM (Ours)	hierarchy	class-wise	**53.10**||**28.20**	**79.85**||**32.50**
CDM[3]	whole set	example-wise	48.45||15.70	36.15||09.50
CDM^{L}	whole set	example-wise	20.70||15.00	17.65||10.40
CF-CBM (Ours)	hierarchy	example-wise	**49.92**||**16.80**	**81.00**||**17.60**

CF-CBM: Experimental Results - Qualitative Analysis



Original Concept Set

Black body/white feathers
bird revered and respected
Found in wetlands Australia/NZ
very elegant bird
Black feathers/white stripes
incubation period 32-34
National icon
magnificent bird/beautiful black feathers
black w/ white wingtips
become aggressive if threatened
kept in zoos
bill is red
very gentle
wingspan up to 2.5m
enjoyed by birdwatchers
fleeing from predators
swim up to 30km/h
perform acrobatic in water
graceful bird that can swim and object
good swimmer

Discovered Concepts Patches

light brown eyes
small head
small/delicate songbird
small wading bird
dark/brown color
head is small
and a black tail
spiral shaped with a pointy end
smooth, wet skin
long/slender body
with small white spots
light brown coat with darker spots
breeds in tundra
mouth light color
damp places
small and black
black

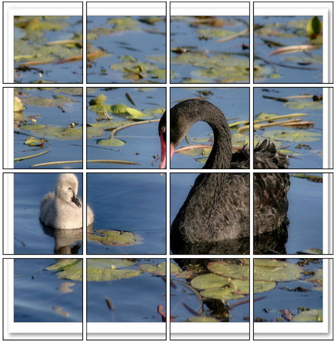


Figure: A random example from the *Black Swan* class of ImageNet-1k validation set. On the upper part, the original concept set corresponding to the class is depicted; on the lower, some of the concepts discovered via our novel CF-CBM.

CF-CBM: Experimental Results - Qualitative Analysis

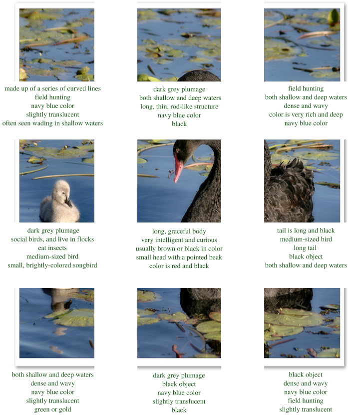


Figure: A random example from the *Black Swan* class of ImageNet-1k validation set. After training, we have access to the discovered concepts on both the image and the patch level.

The End

Thank you!

ArXiV:
arxiv.org/pdf/2310.02116.pdf

GitHub Repository:
github.com/konpanousis/Coarse-To-Fine-CBMs/

References

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

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

[3] Konstantinos P. Panousis, Dino Ienco, and Diego Marcos. Sparse linear concept discovery models. In *Proc. ICCCW CLVL*, 2023.