DCBM: Data-efficient Concept Bottleneck Models

Katharina Prasse* [1] Patrick Knab* [2] Sascha Marton [2] Christian Bartelt [2] Margret Keuper [1,3]

*Equal contribution [1]Data and Web Science Group, University of Mannheim [2]Clausthal University of Technology [3]Max-Planck-Institute for Informatics, Saarland Informatics Campus

Paper

Code

Motivation

- Trust in applications requires **interpretability of neural networks**.
- **C**oncept **B**ottleneck **M**odels (**CBMs**) learn a linear mapping from concept activations to classes that are inherently interpretable.
- CBMs main objectives:
  - → Meaningful **human-interpretable concepts**.
  - → Concepts are sufficiently **specific for the given task**.
  - → **Efficient extraction of concepts** from training images/classes.

Framework: Data-efficient CBMs

Step 1: Concept proposals are created using **foundation models** for segmentation / detection.

Step 2: Concepts are generated by **clustering concept proposals** to remove redundancies.

Step 3: **CBM is trained** to map concept activations to class labels.

Step 4: **Visual concepts are mapped to text** within CLIP space.

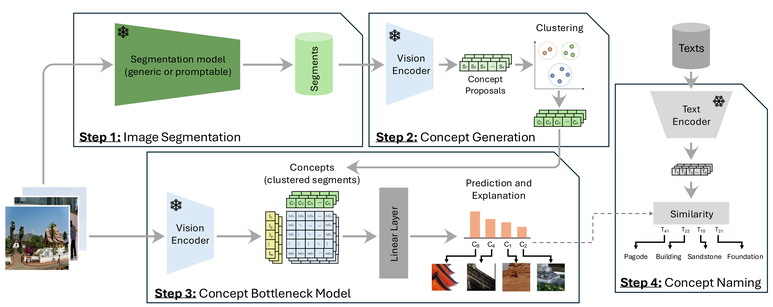

Figure 1. The **DCBM framework** generates concept proposals through foundation models (Step 1). These proposals are then clustered, each represented by its centroid (Step 2). Finally, the concepts are utilized to train a sparse CBM (Step 3). We leverage the image-text alignment to map the visual concept to the corresponding textual concept (Step 4). We can remove undesired concepts after Step 2.

No Description,
No Supervision,
No External Data.
—
Extract Concepts from
YOUR Data.

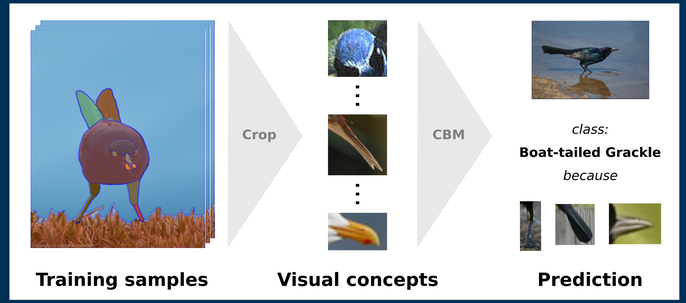

Figure 2: Using vision foundation models, we use cropped image regions as concepts for CBM training. Based on few concept samples (50 imgs/class), DCBMs offer interpretability even for fine-grained classification.

Qualitative & Quantitative Results

Key Insights:

- DCBMs **perform within at most 6% of the linear probe for all datasets (9).**
- Mask-RCNN concept proposals slightly outperform SAM2 and GDINO.
- DCBM **excels in domain specific tasks** (CUB).
- DCBM concepts can be used in OOD settings.
- DCBMs achieve competitive performance based on 50 imgs/class as concept samples.

Table 1. Top-1 **accuracy comparison** across CBM models.

Model	CLIP ViT L/14				
	IMN	Places	CUB	Cif10	Cif100
Linear Probe ↑	83.9*	55.4	85.7	98.0*	87.5*
Zero-Shot ↑	75.3*	40.0	62.2	96.2*	77.9*
LF-CBM [3] ↑	-	49.4	80.1	97.2	83.9
LaBo [6] ↑	**84.0***	-	-	97.8*	86.0*
CDM [4] ↑	83.4*	55.2*	-	95.9	82.2
DCLIP [2] ↑	75.0*	40.5*	63.5*	-	-
DN-CBM [5] ↑	83.6*	**55.6***	-	**98.1***	**86.0***
DCBM-SAM2 (Ours) ↑	77.9	52.1	**81.8**	97.7	85.4
DCBM-GDINO (Ours) ↑	77.4	52.2	**81.3**	97.5	85.3
DCBM-MASK-RCNN (Ours) ↑	77.8	52.1	**82.4**	97.7	85.6

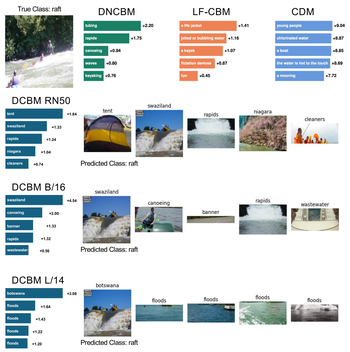

Figure 3. **CBM concept explanation comparison**. DCBM explanations contain no abstract concepts, e.g., fun, chlorinated water. The visual concepts in DCBM should be prioritized over the textual ones, as they originate from the image modality.

Table 2. **OOD performance.** Error rate changes compared between visual CBMs (CLIP ViT-L/14) on ImageNet-R.

	IN-200	IN-R	Gap(%)
DN-CBM [5] ↓	16.4	55.2	38.8
DCBM-SAM2 (Ours) ↓	21.1	**48.5**	**27.4**
DCBM-GDINO (Ours) ↓	22.6	**47.2**	**24.6**
DCBM-MASK-RCNN (Ours) ↓	22.2	**44.6**	**22.4**

Table 3. **Data-efficiency.** DCBM concept proposals are generated based on 50 images/class.

	DN-CBM [5]	DCBM - ImageNet
Dataset	CC3M	50k images (50 imgs/class)
Mem	850 GB (assuming 256x256px)	6 GB
No extra data	×	✓

References

[1] M. Bohle, M. Fritz, and B. Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *CVPR*, 2021.
[2] S. Menon and C. Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.
[3] T. Oikarinen, S. Das, L. Nguyen, and L. Weng. Label-free concept bottleneck models. In *ICLR*, 2023.
[4] K. P. Panousis, D. Ienco, and D. Marcos. Sparse linear concept discovery models. In *ICCV*, 2023.
[5] S. Rao, S. Mahajan, M. Böhle, and B. Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *ECCV*, 2024. First 2 authors contribute equally.
[6] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.

ICML International Conference On Machine Learning
UNIVERSITY OF MANNHEIM
TU Clausthal Clausthal University of Technology
MAX PLANCK INSTITUTE FOR INFORMATICS
Bundesministerium für Forschung, Technologie und Raumfahrt
Funded by the European Union NextGenerationEU