VLG-CBM: Training Concept Bottleneck Models with Vision-Language Guidance

Divyansh Srivastava*, Ge Yan*, Tsui-Wei Weng (*Equal contribution)
UC San Diego

NeurIPS 2024

⭐ **Paper:** https://arxiv.org/pdf/2408.01432 ⭐ **Code**: https://github.com/Trustworthy-ML-Lab/VLG-CBM

⭐ **Project website:** https://lilywenglab.github.io/VLG-CBM/

Concept Bottleneck Model (CBM)

Concept Bottleneck Models (CBMs) [1] provide final interpretable predictions based on human-understandable concepts c

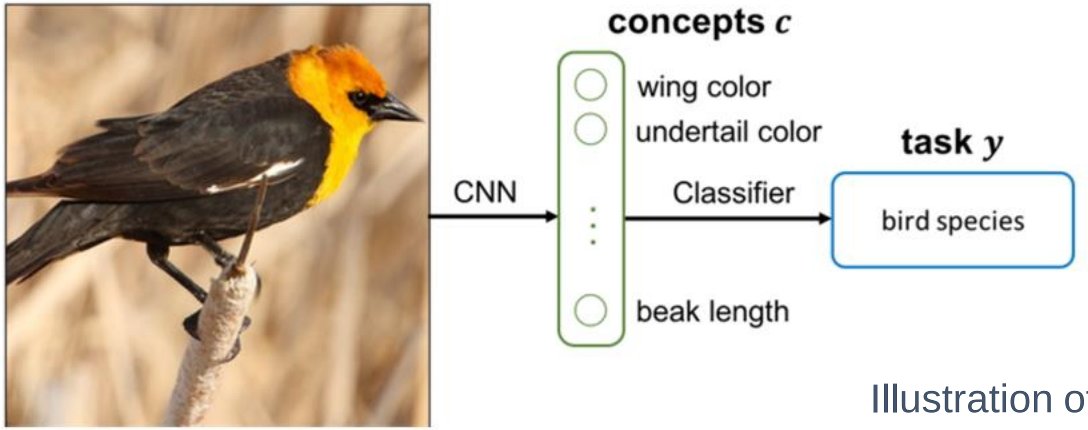


Illustration of CBM, fig from [1]

[1] Koh et al, Concept bottleneck models, ICML 2020

Critical Challenges in current CBMs

Existing CBMs in prior work suffer from two major issues:

- **Challenge #1: Inaccurate concept prediction**
  Inaccurate or wrong explanations which do not match the input images
- **Challenge #2: Information Leakage**
  The concept prediction encodes unintended information for downstream tasks, even if the concepts are irrelevant to the task (e.g. random concepts can still get high acc.)

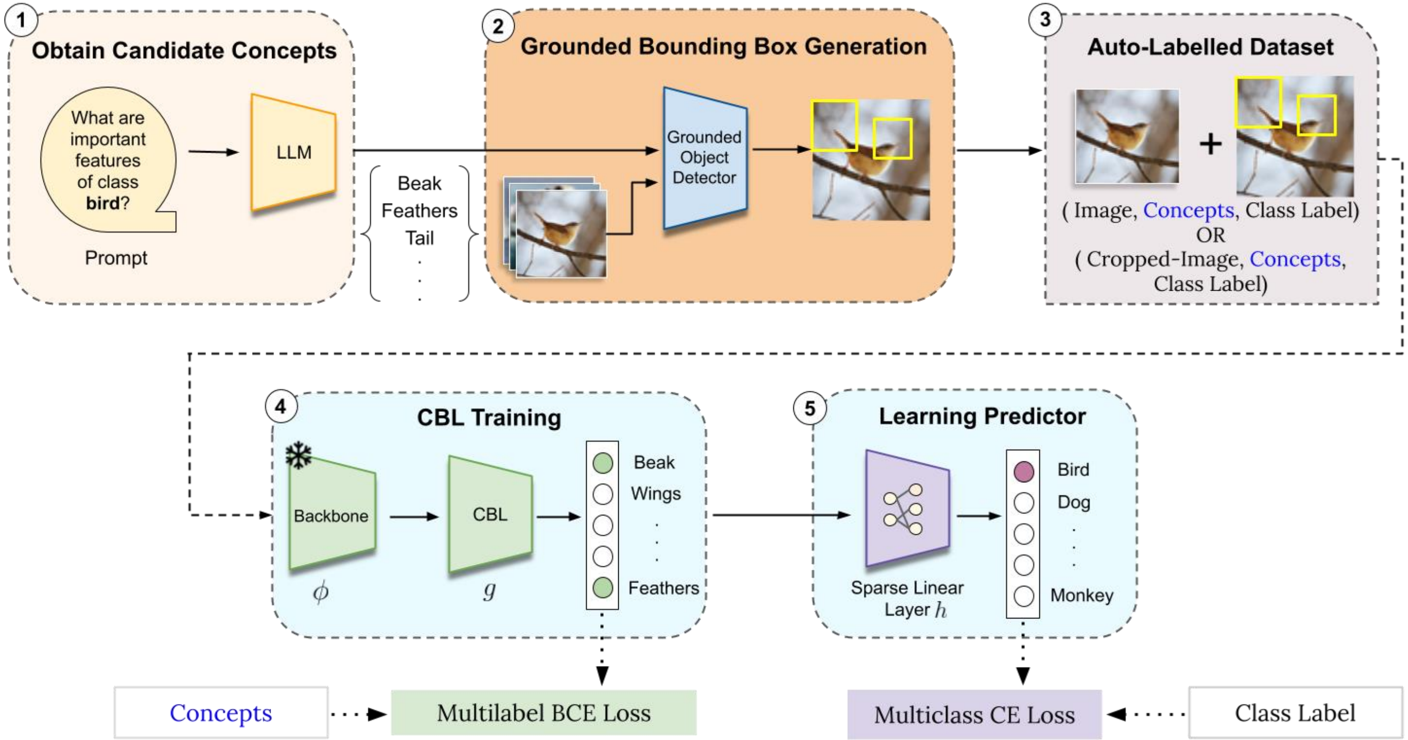
Explanations of why this bird is a *painted bunting*:

1. Color - Blue head, olive back, yellow underparts (101.08)
2. grayish head, back, wings and tail with blue highlights (94.03)
3. bright blue and orange plumage (91.44)
4. large red bill with a slightly hooked tip (89.09)
5. distinctive white throat (-76.91)

Sum of other concepts (-34.89)

Our contribution #1: a new pipeline VLG-CBM

VLG-CBM address Challenge #1 by automatically grounding concepts



Our contribution #2.1: New theory

To explain Challenge #2 information leakage, we prove that
“a random CBL could approximate any linear classifier (w) when the number of concepts (k) is greater or equal to the embedding dimension (d)”

weight vector of linear classifier

approx. error

$$E(k) \leq \begin{cases} \lambda_{max}(1 - \frac{k}{d})\|w\|_2^2, & k < d; \\ 0, & k \geq d. \end{cases}$$

\# of concepts

embedding dim of backbone

Our contribution #2.2: New evaluation metric

Inspired by our theory, we proposed to use the Number of Effective Concepts (NEC) to control information leakage in Challenge #2.

$$NEC(W_F) = \frac{1}{C} \sum_{i=1}^{C} \sum_{j=1}^{k} \mathbf{1}\{(W_F)_{ij} \neq 0\}$$

# of concepts

# of classes

final weight matrix of the predictor

Results

Accuracy on 5 datasets under (1) NEC=5 (2) average accuracy. Our VLG-CBM outperforms all baselines [2-4] under both metrics.

Dataset	CIFAR10		CIFAR100		CUB200		Places365		ImageNet	
Metrics	Acc@5	Avg. Acc.	Acc@5	Avg. Acc.	Acc@5	Avg. Acc.	Acc@5	Avg. Acc.	Acc@5	Avg. Acc.
Random	67.55%	77.45%	29.52%	47.21%	68.91%	73.44%	17.57%	28.62%	41.49%	61.97%
LF-CBM	84.05%	85.43%	56.52%	62.24%	53.51%	69.11%	37.65%	42.10%	60.30%	67.92%
LM4CV	53.72%	69.02%	14.64%	36.70%	N/A	N/A	N/A	N/A	N/A	N/A
LaBo	78.69%	82.05%	44.82%	55.18%	N/A	N/A	N/A	N/A	N/A	N/A
VLG-CBM (Ours)	**88.55%**	**88.63%**	**65.73%**	**66.48%**	**75.79%**	**75.82%**	**41.92%**	**42.55%**	**73.15%**	**73.98%**

(LM4CV [3] / LaBo [4] only supports CLIP-Backbone, thus some entries are marked as N/A)

[2] LF-CBM: Oikarinen etal, Label-free concept bottleneck models, ICLR 2023.
[3] LM4CV: Yan etal, Learning concise and descriptive attributes for visual recognition, ICCV 2023.
[4] LaBo: Yang etal, Language model guided concept bottlenecks for interpretable image classification, CVPR 2023

Results: CLIP backbone

VLG-CBM outperforms all baselines by a large margin under both metrics:

(i) Acc@NEC = 5 & (ii) Average Acc

Dataset	ImageNet		CUB	
Metrics	Acc@5	Avg. Acc	Acc@5	Avg. Acc
LF-CBM	52.88%	62.24%	31.35%	52.70%
LM4CV	3.77%	26.65%	3.63%	15.25%
LaBo	24.27%	45.53%	41.97%	59.27%
VLG-CBM(Ours)	**59.74%**	**62.70%**	**60.38%**	**66.03%**

Results: Decision Explanation

Our method provide accurate explanations while prior work (LF-CBM, LM4CV) provide inaccurate/wrong/less useful explanations

Our Method:
1. short pointed beak (0.65)
2. blue head (0.21)
3. green back (0.09)
4. short stout bill (0.01)
5. small songbird (0.01)

Sum of other concepts (**0.00**)

LF-CBM:
1. NOT a brown and white color scheme (1.77)
2. NOT white and black coloration (1.66)
3. iridescent feathers (1.37)
4. NOT a black bib with white stripes (1.29)
5. NOT a black and white color scheme (1.01)

Sum of other concepts (4.22)

LM4CV:
1. Color - Blue head, olive back, yellow underparts (101.08)
2. grayish head, back, wings and tail with blue highlights (94.03)
3. bright blue and orange plumage (91.44)
4. large red bill with a slightly hooked tip (89.09)
5. distinctive white throat (-76.91)

Sum of other concepts (-34.89)

Our Method:
1. black and white coloration (6.09)
2. long face (5.34)
3. black brindle or fawn coat (0.09)
4. droopy lips and ears (0.09)

Sum of other concepts (**0.00**)

LF-CBM:
1. black pepper (1.04)
2. a Belgian Malinois (0.90)
3. a giraffe (0.90)
4. a big dog (0.89)
5. a large, rocky mass (0.77)

Sum of other concepts (8.14)

LM4CV:
1. English setters are bred in England (37.18)
2. shaggy, long fur (18.31)
3. large quantities of baked goods (9.55)
4. typically has a "snow nose" (pinkish or black skin on the muzzle that is exposed due to cold weather) (7.27)
5. red and white stripes on the front (6.11)

Sum of other concepts (-34.68)

Conclusion

In this paper, we have 2 main contributions:

1. We proposed **VLG-CBM**, a novel framework to address inaccurate concept prediction (challenge #1) of previous CBMs.
2. We provided the **first theoretical analysis** for information leakage (challenge #2) and proposed a new metric **NEC** to control it, allowing fair comparison between CBMs.

For more details, please see:

⭐ **Paper:** https://arxiv.org/pdf/2408.01432 ⭐ **Code**: https://github.com/Trustworthy-ML-Lab/VLG-CBM

⭐ **Project website:** https://lilywenglab.github.io/VLG-CBM/