Faithful CBM - Building Faithful and Interpretable Deep Neural Network models

Henry Xu
jix049@ucsd.edu

Lily Weng
lweng@ucsd.edu

Abstract

Deep neural networks' black-box nature raises concerns about interpretability in critical applications. While recent Concept Bottleneck Models (CBMs) using Vision-Language Models have improved transparency, they struggle with concept prediction accuracy and faithfulness. We propose an expert-in-the-loop framework with three key innovations: a refined concept generation method with standardized cross-class concepts, a novel faithfulness measurement comparing spatial attribution maps with GroundingDINO bounding boxes, and fine-tuning concept neuron with saliency loss. Our approach enhances both the accuracy and interpretability of CBMs while ensuring faithful concept detection.

1 Introduction

With deep neural networks being increasingly deployed in critical real-world applications such as healthcare, autonomous driving, and finance, the need for interpretability has become essential. While these models often achieve remarkable predictive performance, their complex decision-making processes—frequently described as ”black boxes”—make it challenging to understand why a particular decision was made. This lack of transparency raises concerns about trust, accountability, and fairness, especially in high-stakes domains where incorrect predictions can have significant consequences.

The Concept Bottleneck Models (CBMs) address these concerns by introducing a concept bottleneck layer (CBL) before the final output layer that captures human-interpretable concepts Oikarinen et al. (2023). Instead of making predictions directly from raw data, the model identifies familiar concepts first—such as “wheels” and “headlights” in an image of a car—and then uses these concepts as building blocks to reach its final decision. This approach not only enhances transparency but also allows end-users, including non-experts, to understand and validate the reasoning behind a model’s predictions, fostering greater trust and reliability in AI systems.

Recent research has explored using Vision-Language Models (VLMs) as an alternative to human-created annotations. In these approaches, Large Language Models (LLMs) are used to generate concept sets, and Vision-Language Models are used to annotate dataset with concepts activations Oikarinen et al. (2023); Srivastava, Yan and Weng (2024). This elimination of manual labeling has allowed some Concept-Based Models (CBMs) to be applied to extensive datasets like ImageNet. Despite this progress, these approaches still face the limitation of inaccurate concept prediction: the models frequently make mistakes in identifying concepts, producing predictions that don’t accurately reflect what’s in the image, therefore hurting the faithfulness of these methods and lowering interpretability of final prediction. Additionally, the technique of Automatic Concept Correction is being used to improve performance of CBMs, which involves setting concept activation values to zero if the concept is not associated with the true class of an image Sun, Oikarinen and Weng (2024). While this is helpful in boosting the accuracy of final-class predictions, it risks concept neuron fitting to the actual class instead of the true concept, therefore raising faithfulness concerns.

We propose a more faithful expert in the loop framework to address the limitations of previous works. Our contributions are summarized as follows:

1. We propose to use a refined method for concept set generation which produces visual concepts that are well understood by both human and VLMs, results in more effective model training and accurate final concepts set. Additionally, we will standardize concepts shared across classes to enhance compatibility with Automatic Concept Correction (ACC), aiming to boost faithfulness while maintaining performance.
2. We propose a novel faithfulness measurement that evaluates concept neurons by comparing spatial information from attribution maps with GroundingDINO bounding boxes Liu et al. (2023). This approach helps determine whether a concept neuron genuinely detects the intended concept rather than fitting to the relevant classes.
3. We propose a saliency loss to be used during fine-tuning that improves the spatial

faithfulness of concept neurons.

2 Methods

2.1 Concept Set Refinement

Previous methods have developed a pipeline for automatic concept set generation given a list of target classes. In short, this pipeline works by prompting Large Language Models (LLMs) such as GPT-3 to ”list the most important features for recognizing something as a class.” Although this approach was scalable and produced some high-quality concepts, we identified several issues:

1. Some non-visual concepts were incorrectly included, such as ”fast, erratic flight patterns” for the CUB200 dataset.
2. The concept lists produced for certain classes were not comprehensive enough. For example, ”black head” is a common feature found in many bird species, but it was only identified as a concept for 10 out of 200 bird species. In contrast, our refined method identified it for 30 species.
3. Semantically identical concepts were generated for different classes with inconsistent wording, leading them to be treated as different concept neurons and each fitting only to the class that generated them. For example, ”black body” and ”all black body” represent the same concept but were treated as distinct concepts.

When ACCSun, Oikarinen and Weng (2024) is used, the concept activation of an image is set to zero whenever the concept is not listed for the ground truth class. Consequently, both issues 2 and 3 cause concept neurons to fit to specific classes rather than the true underlying concepts.

For CUB200, we developed an automatic attribute framework that provides GPT-4OpenAI (2023) with a list of 22 bird attributes, such as ”head color,” ”head pattern,” and ”bill color.” We also provided several completed examples to guide the model toward the desired concept types through few-shot learning. This approach resulted in more comprehensive and consistently worded concept lists for each bird class.

For critical applications and real-world scenarios with a small number of classes, expert knowledge can be utilized to generate high-quality concept sets or refine LLM-generated ones. The attribute framework helps by classifying raw concepts into categories, making it easier for human experts to unify similar concepts that have different wordings.

2.2 Faithfulness Measurement

We propose to measure faithfulness with attribution map and GroundingDINO bounding boxes alignment.

The concept activation for an input image x can be represented as:

$$G(x) = g(\phi(x)) = \begin{bmatrix} s_{c_1} \\ s_{c_2} \\ \vdots \end{bmatrix}$$

Where g is the concept bottleneck layer. $\phi : x \rightarrow \mathbb{R}^d$ is the backbone model that generates a d-dimensional embeddings for an input image x_i, s_{c_j} is the activation for jth concept.

Start with a validation dataset, for a concept neuron with target concept c_j, find top k activating images. Then, for each image, use GroundingDINO to annotate it with prompt "bird . target concept", obtain the GroundingDINO bounding box. Additionally, use attribution method like GradCAM to obtain a saliency map of that neuronSelvaraju et al. (2019).

We compute three statistics:

- IoU: Inersection over Union score between bounding box and area with top 30% saliency
- Saliency ratio: average saliency inside bounding box vs average saliency inside bounding box
- % Saliency captured: sum of saliency inside bounding box as a percentage of total sum.

2.3 Fine-tuning with saliency loss

In addition to the Binary Cross Entropy (BCE) loss for multi-label prediction, we defined a saliency loss to make sure the concept neurons pay attention to the correct parts of an image:

$$L_{sal} = RELU(\text{saliency outside bboxes} - \text{saliency inside bboxes} + 0.5)$$

3 Results

3.1 Concept set refinement results

We generated new concept set with the new prompt and measured the faithfulness and accuracy of resulted CBM, results are shown in Table 1. We also compared the metrics among mutual concepts, results are shown in Table 2. We also plotted the relationship between number of images assigned to concept and faithfulness score in Figure 1.

3.2 Faithfulness measurement results

Example visualizations of faithfulness measurement and corresponding scores are shown in Figure 2.

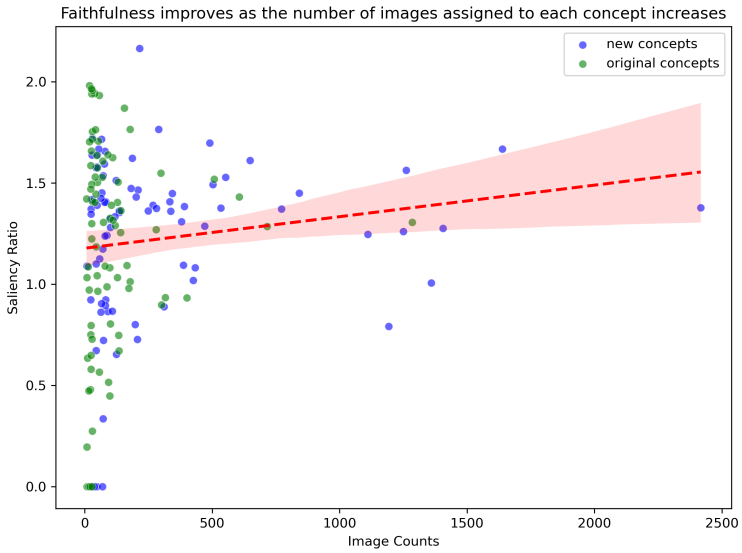
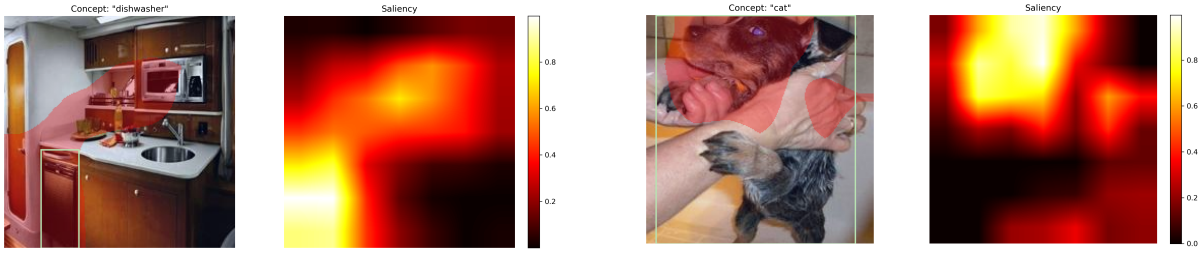
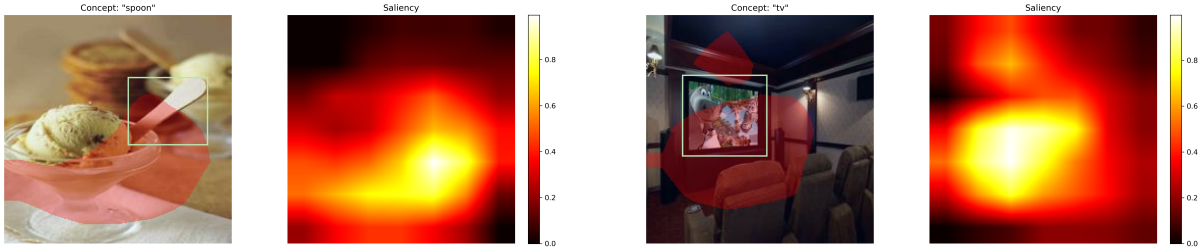


Figure 1: Concept faithfulness increases as number of images assigned to each concept increases



(a) Dishwasher IoU:0.24,Saliency ratio:2.56)

(b) Cat IoU:0.338, Saliency ratio:2.260)



(c) Spoon IoU:0.150, Saliency ratio:1.420

(d) TV IoU:0.251, Saliency ratio:1.846)

Figure 2: Example visualizations of faithfulness measurements with their corresponding scores.

Table 1: Comparison of metrics between CBM trained with original concepts and CBM trained with refined concepts.

Metric	Original concepts	Refined concepts
IoU	0.202	0.186
Saliency ratio	1.222	1.217
%Saliency capture	0.246	0.213
%No matching bbox	0.071	0.052
Acc @NEC=5	0.7527	0.7504

Table 2: Comparison of metrics between mutual original concepts and new cleaned concepts.

Metric	Original concepts	Refined concepts
IoU	0.204	0.207
Saliency ratio	1.185	1.235
%Saliency capture	0.232	0.254

3.3 Fine-tuning with saliency loss results

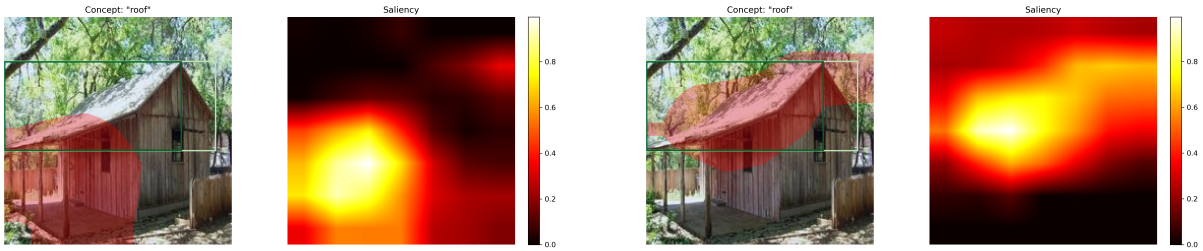
We performed fine-tuning experiments using both the CUB200 and Places365 datasetsWah et al. (2011); Zhou et al. (2017). For CUB200, we tested two settings: fine-tuning a single concept neuron and fine-tuning all concept neurons. For Places365, we only tested the single concept neuron setting due to computational limitations. Quantitative results are presented in Table 3, while qualitative visualizations are shown in Figure 3.

Table 3: Comparison of model performance across different datasets and training configurations, original scores are shown in parenthesis

Model	Saliency ratio	Acc @NEC=5	Training time
CUB fine-tune ”black head”	1.599 (1.252)	0.7503 (0.7504)	16 min
CUB fine-tune all	1.2629 (1.217)	0.7333 (0.7504)	200 min
Places365 finetune ”roof”	3.102 (1.219)	Should be close	2h per epoch

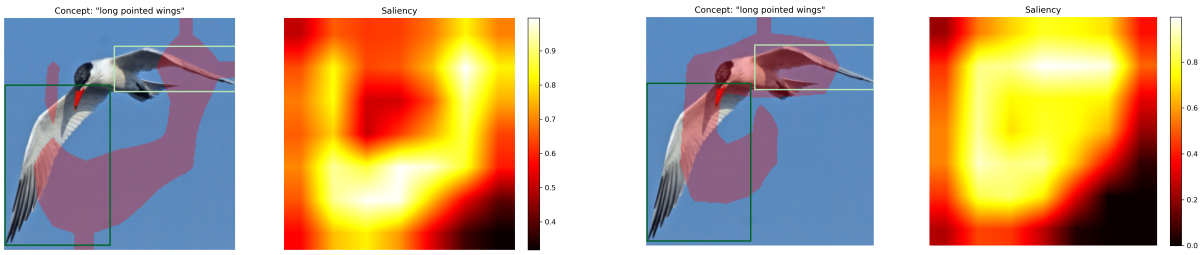
4 Discussion

With our concept refine method, the metrics actually worsened slightly. This occurred because some general concepts that tend to get large bounding boxes and higher scores were excluded in the new concept set. When we compared metrics among mutual concepts, we found that faithfulness did improve, as shown in Table 2. Therefore, our method helped make per-class concepts more comprehensive. We discovered a correlation between the number of training images and concept faithfulness, as illustrated in Figure 1. In the refined concept set, each concept was trained with 144 images on average (compared to 60



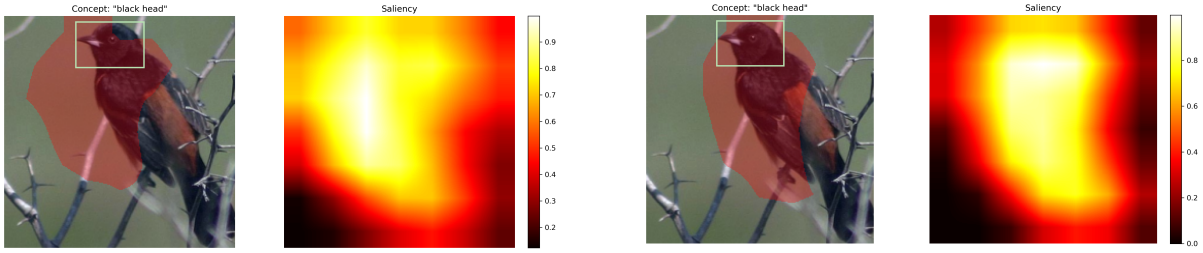
(a) Original Roof (b) Fine-tuned Roof

IoU: 0.111→0.580, saliency ratio: 0.600→2.866



(c) Original Long Pointed Wings (d) Fine-tuned Long Pointed Wings

IoU: 0.294→0.368, saliency ratio: 1.115→1.467



(e) Original Black Head (f) Fine-tuned Black Head

IoU: 0.150→0.176, saliency ratio: 1.372→1.769

Figure 3: Comparison between original and fine-tuned models showing improvements in IoU and saliency ratio.

(a) Original ”brown wings” concept neuron primarily activating for a single bird class

(b) Refined ”brown wings” concept neuron activating for multiple bird classes with brown wings

Figure 4: Comparison of concept neuron activations before and after concept refinement.

in the original set), leading to more faithful learning of the part. For example, the concept ”brown wings” was originally fitting to a single bird class, but after refinement, the top activating images contain birds with brown wings from multiple classes, as shown in Figure 4.

From our saliency measurement results, we found that GroundingDINO works well in most cases but occasionally has glitches. For example, in Figure 2b, although the image actually contains a dog, a bounding box labeled ”cat” was still detected. There were also cases where it failed to capture concepts present in the image. Future work could improve GroundingDINO’s effectiveness by rephrasing complex concepts in ways more understandable by the model. We may also adopt more powerful object grounding models in the future or use densely labeled datasets directly.

The results of our fine-tuning experiments are promising. The concept neurons now direct their attention more accurately to the concept areas, as shown in Figure 3, suggesting improved spatial faithfulness. However, a limitation is that prediction accuracy dropped slightly. This occurs because some of the original concepts were directly fitting to the class, which boosted performance compared to faithfully detecting the concepts. Another limitation is the high computational cost when fine-tuning all neurons for large-scale datasets with millions of images, such as Places365 and ImageNet. One direction for future work is to optimize the pipeline to reduce this computational cost.

5 Conclusion

We have presented a framework for enhancing CBM faithfulness through three key innovations: refined concept generation with standardized cross-class concepts, a novel faithfulness measurement using attribution maps with GroundingDINO bounding boxes, and

fine-tuning with saliency loss. Our experiments on CUB200 and Places365 datasets demonstrate improved concept detection faithfulness while maintaining reasonable classification performance. These advancements make CBMs more trustworthy for critical applications where understanding model reasoning is essential, paving the way for AI systems that are both powerful and transparently interpretable.

References

Liu, Shilong, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu et al. 2023. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." *arXiv preprint arXiv:2303.05499*

Oikarinen, Tuomas, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. "Label-Free Concept Bottleneck Models." [Link]

OpenAI. 2023. "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774*. [Link]

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision* 128 (2), p. 336–359. [Link]

Srivastava, Divyansh, Ge Yan, and Tsui-Wei Weng. 2024. "VLG-CBM: Training Concept Bottleneck Models with Vision-Language Guidance." [Link]

Sun, Chung-En, Tuomas Oikarinen, and Tsui-Wei Weng. 2024. "Crafting Large Language Models for Enhanced Interpretability." [Link]

Wah, C., S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. "CUB200." Technical Report CNS-TR-2011-001, California Institute of Technology

Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. "Places: A 10 million Image Database for Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*