Concept Bottleneck with Visual Concept Filtering for Explainable Medical Image Classification

Injae Kim⋆, Jongha Kim⋆, Joonmyung Choi, Hyunwoo J. Kim

Department of Computer Science and Engineering, Korea University
{dna9041, jonghakim, pizard, hyunwoojkim}@korea.ac.kr

Abstract. Interpretability is a crucial factor in building reliable models for various medical applications. Concept Bottleneck Models (CBMs) enable interpretable image classification by utilizing human-understandable concepts as intermediate targets. Unlike conventional methods that require extensive human labor to construct the concept set, recent works leveraging Large Language Models (LLMs) for generating concepts made automatic concept generation possible. However, those methods do not consider whether a concept is *visually* relevant or not, which is an important factor in computing meaningful concept scores. Therefore, we propose a visual activation score that measures whether the concept contains visual cues or not, which can be easily computed with unlabeled image data. Computed visual activation scores are then used to filter out the less visible concepts, thus resulting in a final concept set with visually meaningful concepts. Our experimental results show that adopting the proposed visual activation score for concept filtering consistently boosts performance compared to the baseline. Moreover, qualitative analyses also validate that visually relevant concepts are successfully selected with the visual activation score.

Keywords: Medical Image Classification · Explainable AI · Concept Bottleneck Models · Large Language Models.

1 Introduction

Deep Neural Networks (DNNs) have addressed many problems in various fields, including the medical domain [1–4]. For instance, [3] diagnoses breast lesions on dynamic contrast-enhanced MRI with deep learning, and [4] diagnoses hip osteoarthritis using convolutional neural networks (CNNs). Despite the huge success of deep learning based models in the field of medical analysis and diagnosis, such models innately lack a crucial capability required in the medical domain - *interpretability*. Therefore, to tackle the difficulty in interpreting the model decision, a series of research have been presented to enhance the interpretability and the explainability of deep learning models.

Concept Bottleneck Models (CBMs) [5] is one of the works that make image classification more interpretable. Instead of directly predicting the target

---

⋆ Equal contribution

from non-interpretable latent representations, CBMs first predict concept scores, which measure the degree of an image matching to the human-understandable concepts. Then, the final prediction is done based on the predicted concept scores, which makes the prediction process interpretable.

Applying CBMs requires the construction of a set of concepts that well-describe images and are discriminable. Conventional approaches [5, 6] try manually defining concepts, which requires extensive labor of skilled expert that is familiar with the target domain. However, such an approach largely hinders the scalability and generalizability of CBMs, since building large-scale concept set manually is costly, especially in the medical domain.

To address the issue, recent works [7, 8] propose generating concepts automatically by prompting Large Language Model (LLM) that contains rich information across various subjects to generate the concepts describing target classes. Although such methods remove the need for manual concept generation, LLM turns out to generate non-visual concepts that do not align with images, therefore providing a noisy signal that hinders proper training.

To this end, we propose a *visual activation score*, which measures whether a concept contains visual information or not. By taking account of the visual activation score in the concept filtering phase via a submodular optimization [9], concepts that do not contain any visual cue useful for classifying the image are successfully removed from the initial concept set.

Experimental results demonstrate the effectiveness of the proposed visual activation score with consistent gains in terms of accuracy under multiple experimental settings. Also, further analyses again validate that the visual activation score successfully discriminates the visual and non-visual concepts, therefore helping the model better classify the image with refined concepts.

2   Related Works

2.1   Concept Bottleneck Models

Concept Bottleneck Models (CBMs) [5] aim to make the process of an image classification more interpretable, by implementing a concept bottleneck layer before the final classification layer. A concept bottleneck layer outputs a score of an image corresponding to multiple interpretable concepts. Calculated concept scores are fed into a final linear layer to classify an image. Recent works [7, 8] propose leveraging information learned by Large Language Models (LLMs) to automatically extract concepts instead of manually constructing the concept set.

2.2   Large Language Models

Recently, Large Language Models (LLMs) based on Transformer [10] architecture trained with the large-scale text corpus have shown to be effective in various downstream tasks including zero-shot and few-shot learning [11, 12], multi-task [13], visual question answering (VQA) [12]. Moreover, leveraging knowledge of LLMs for computer vision tasks, such as generating candidate concepts to classify an image [7, 8], is being studied recently.
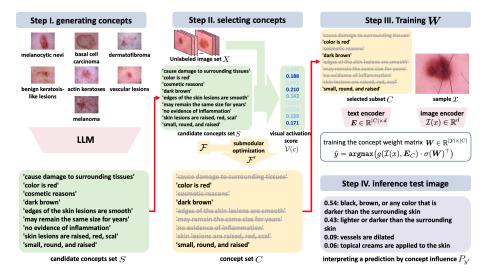
Fig. 1: **Method Overview. Step 1**: Generate candidate concepts set S by prompting the large language model (LLM); **Step 2**: Select visually relevant concepts via submodular optimization with the score function $\mathcal{F}'$; **Step 3**: Train a concept weight matrix W which projects concept scores into prediction logits; **Step 4**: Interpret inference results with the concept influence $P_{y'}$.

3   Method

3.1   Preliminary

Image classification is the task of predicting the class $y \in \mathcal{Y}$ an input image $x \in \mathcal{X}$ belongs to, where $\mathcal{Y}$ and $\mathcal{X}$ are sets of target classes and images, respectively. In order to make the classification process more interpretable, Concept Bottleneck Models [5, 7, 8] firstly compute similarities between the image and a concept c in the concept set C, which indicates how well the image and pre-defined concepts are aligned. Computed similarities are then fed into the final classification layer to determine the class an image belongs to.

Generating Candidate Concept Set We adopt LaBo [8] as the baseline, which constructs the candidate concept set S using GPT-3 [11]. For each class y, LaBo prompts GPT-3 to retrieve sentences that describe the class. 500 sentences per class are retrieved from GPT-3 and split into shorter concepts to form a candidate concept set S_y for each class. Then, the whole concept set S is defined as a union of S_y as below:

$$S = \bigcup_{y \in \mathcal{Y}} S_y \tag{1}$$

Concept Selection via Submodular Optimization For a concept set S_y, submodular optimization [9] is applied to select concepts with desired property

using the score function $\mathcal{F}$, therefore resulting in a concept subset $C_y \subseteq S_y$, where $|C_y| = k$. Due to the limitation in space, we refer to details about the submodular optimization to [8]. The score function $\mathcal{F}$ that evaluates the utility of the subset C_y is defined as:

$$\mathcal{F}(C_y) = \alpha \cdot \sum_{c \in C_y} D(c) + \beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2), \tag{2}$$

where $D(c)$ is a discriminability score of the concept c, $\phi(c_1, c_2)$ is a concept similarity between two concepts c_1 and c_2, and α, β are controllable hyperparameters. Maximizing the discriminability score $D(c)$ encourages selecting concepts that are aligned only with images with specific labels but not with the other images. To do so, the conditional likelihood $\overline{\text{sim}}(y|c)$ of a similarity score $\text{sim}(y, c)$ given a concept c is defined as:

$$\overline{\text{sim}}(y|c) = \frac{\text{sim}(y, c)}{\sum_{y' \in \mathcal{Y}} \text{sim}(y', c)}, \quad \text{sim}(y, c) = \frac{1}{|\mathcal{X}_y|} \sum_{x \in \mathcal{X}_y} \mathcal{I}(x) \cdot \mathcal{T}(c)^\top, \tag{3}$$

where $\mathcal{X}_y$ is training image set labeled with y, and $\mathcal{I}(\cdot)$ and $\mathcal{T}(\cdot)$ are the CLIP [14] image encoder and text encoder, respectively. Finally, $D(c)$ is defined as its negative entropy to maximize:

$$D(c) = \sum_{y' \in \mathcal{Y}} \overline{\text{sim}}(y'|c) \cdot \log\left(\overline{\text{sim}}(y'|c)\right). \tag{4}$$

The second term of Equation 2 is a coverage score that aims to maximize the minimum similarity between each concept in the subset C_y and the overall set S_y. With the coverage score, the selection of concepts covering a wide range of meanings of a target class is enabled. Then, the whole concept set C is defined as the union of C_y, analogous to the definition of S in Equation 1.

Optimizing Concept Weight Matrix After obtaining the concept set C, CLIP text features of concepts are stacked to form a concept embedding matrix $\boldsymbol{E}_C \in \mathbb{R}^{|C| \times d}$, where each row of $\boldsymbol{E}_C$ corresponds to a CLIP text embedding of a concept, $|C|$ is the size of the whole concept set C, and d is a CLIP feature dimension. With $\boldsymbol{E}_C$, a concept score $g(\mathcal{I}(x), \boldsymbol{E}_C) \in \mathbb{R}^{|C|}$ between $\boldsymbol{E}_C$ and an image $x \in \mathcal{X}$ is calculated as:

$$g(\mathcal{I}(x), \boldsymbol{E}_C) = \mathcal{I}(x) \cdot \boldsymbol{E}_C^\top. \tag{5}$$

Finally, a concept weight matrix $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{Y}| \times |C|}$ which maps concept scores into the final prediction logit is optimized, where the final prediction $\hat{y}$ is computed as $\texttt{argmax}\left(g(\mathcal{I}(x), \boldsymbol{E}_C) \cdot \sigma(\boldsymbol{W})^\top\right)$. $\sigma(\boldsymbol{W})$ denotes the softmax operation applied along the concept axis, where $\boldsymbol{W}_{y,c} = e^{\boldsymbol{W}_{y,c}} / \sum_{y' \in \mathcal{Y}} e^{\boldsymbol{W}_{y',c}}$. As of the initialization, $\boldsymbol{W}_{y,c}$ is set as 1 if $c \in C_y$ and 0 otherwise, in order to learn the weight $\boldsymbol{W}$ effectively in few-shot settings. Given an image x, the concept influence

$P_y \in \mathbb{R}^{|C|}$, which represents how much a concept influences the prediction of the class y can be calculated as below:

$$P_y = g(\mathcal{I}(x), \boldsymbol{E}_c) \odot \sigma(\boldsymbol{W}_{y,*}), \qquad (6)$$

where $\odot$ is an element-wise multiplication operation.

3.2   Concept Selection with Visual Activation Score

In order to obtain a reliable concept score in Equation 5, the concept $c \in C$ must include a *visual cue*, e.g., 'darker in color'. However, the candidate concept set S automatically extracted from LLM includes a lot of *non-visual concepts*, which do not contain any visual cue, e.g., 'most common type of precancerous lesion in the united states'. Those non-visual concepts hinder proper learning of the concept weight matrix $\boldsymbol{W}$ since scores of those concepts provide noisy signals. Therefore, an appropriate criterion to filter out the non-visual concepts is required when constructing the concept subset C via a submodular optimization.

To measure the amount of visual information a concept contains, we define a scalar visual activation score $\mathcal{V}(c)$ of a concept c defined as a standard deviation of CLIP scores between a concept c and images $x \in X$ as below:

$$\mathcal{V}(c) = \text{stdev}(\{\mathcal{T}(c) \cdot \mathcal{I}(x)^\top\}_{x \in X}), \qquad (7)$$

where X denotes an unlabeled target image set to calculate visual activation scores on, and $\mathcal{T}(c), \mathcal{I}(x) \in \mathbb{R}^d$ are the CLIP text embedding and the image embeddings of a concept c and an image x, respectively. As defined in the equation, the visual activation score of a concept c is calculated as a standard deviation of concept scores among every image in the target image set X. In other words, a concept that is activated differently depending on the image is regarded as a concept containing visual cues, since those concepts sensitively respond to visually distinct samples. Note that an arbitrary dataset can be set as X since it does not require labels. Further analyses regarding the utilization of various datasets as X are provided in Section 5.2. Calculated $\mathcal{V}(c)$ is then added to the original score function $\mathcal{F}$ to form a new score function $\mathcal{F}'$ as follows:

$$\mathcal{F}'(C_y) = \alpha \cdot \sum_{c \in C_y} D(c) + \beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2) + \gamma \cdot \sum_{c \in C_y} \mathcal{V}(c), \qquad (8)$$

With the new score function $\mathcal{F}'$, the subset $C_{y'}$ is obtained via a submodular optimization, and the following procedures are done analogously as described in Section 3.1.

4   Experiments

We validate the effectiveness of the proposed method by applying the method to HAM-10000 [15], a skin disease dataset. The dataset consists of 10,015 dermatoscopic images collected from various patients. The target classes $\mathcal{Y}$ consist of 7

Table 1: **Performance on HAM-10000 dataset.** "Number of shots" denotes the number of labeled samples per target class, where 'Full' denotes that the model is trained with the whole dataset. * denotes reproduced results.

Method	Number of Shots					
	1	2	4	8	16	Full
Linear Probe*	44.4	58.5	44.9	49.0	61.5	82.5
LaBo[8]*	36.5	44.9	44.5	43.0	58.5	80.8
LaBo[8]* + Ours	53.2 (+16.7)	45.4 (+0.5)	47.4 (+2.9)	46.1 (+3.1)	61.4 (+2.9)	81.0(+0.2)

5 concepts with the highest $\mathcal{V}(c)$	5 concepts with the lowest $\mathcal{V}(c)$
- dark brown or black mole with irregular borders - central area of darker pigmentation - small, pearly-white or flesh-colored bump on the skin - dark brown or black lesion with irregular borders - nose, ears, lips, and hands	- others may require medical or surgical treatment - thought to be caused by a combination of genetic and environmental factors - may not become apparent for years - considered to be low-risk - at least one in their lifetime

Fig. 2: Examples of concepts that have the highest and the lowest visual activation score $\mathcal{V}(c)$ are listed.

types of skin problems, Melanocytic Nevi, Benign Keratosis-like Lesions, Dermatofibroma, Vascular Lesions, Actinic Keratoses, Basal Cell Carcinoma, and Melanoma. For all experiments, we follow the experimental settings of LaBo [8] except for the hyperparameters $\alpha, \beta,$ and γ. All experiments are done with a single NVIDIA RTX A6000 GPU.

4.1   Experimental Results

In Table 1, a consistent gain in accuracy under every single setting compared to LaBo is reported by applying the visual activation score. In terms of filtering out non-visual concepts, LaBo's discriminability score $D(c)$ also indirectly refines non-visual concepts by maximizing the negative entropy of $\overline{\text{sim}}(y|c)$. However, the fewer label data, the more inaccurate value of $\overline{\text{sim}}(y|c)$ is, and at 1-shot, the LaBo achieved a low accuracy of 36.7%. By using an unlabeled image set X, the visual activation score $\mathcal{V}(c)$ encourages to filter non-visual concepts effectively, therefore outperforming LaBo by 16.5% at 1-shot. Also, while maintaining the interpretability, in some cases, the proposed method is shown to even outperform performances of linear probing where the classification process is completely non-interpretable.

5   Analysis

5.1   Analysis on a Visual Activation Score $\mathcal{V}(c)$

In Figure 2, five concepts with the highest and the lowest visual activation score $\mathcal{V}(c)$ are listed, respectively. As shown in the figure, concepts with high $\mathcal{V}(c)$

Table 2: **Analysis on the image set X for visual activation score.** "Number of shots" denotes the number of labeled samples per target class, where 'Full' denotes that the model is trained with the whole dataset.

Image set X	Number of Shots					
	1	2	4	8	16	Full
w/o $\mathcal{V}(c)$	36.5	44.9	44.5	43.0	58.5	80.8
HAM10000	**53.2**	45.4	47.4	46.1	61.4	**81.0**
ImageNet	48.5	**46.2**	**50.3**	**56.8**	62.1	80.9
COCO	47.7	45.4	48.5	45.7	**62.3**	80.8

turn out to be visual concepts, such as "dark brown or black mole with irregular borders". On the other hand, concepts with low $\mathcal{V}(c)$ are mostly non-visual or meaningless concepts, including "others may require medical or surgical treatment". Such qualitative examples demonstrate that the visual activation score $\mathcal{V}(c)$ acts as an effective measure to detect non-visual or meaningless concepts.

5.2   Analysis on Target Image Set X

In Table 2, experimental results with multiple target dataset X to calculate the visual activation score $\mathcal{V}(c)$ in Equation 7 are provided. We conduct experiments under adopting train splits of HAM-10000 [15], ImageNet [16], and COCO [17] as X. For the ImageNet and COCO datasets, 10,000 images are randomly sampled to match the size of X to that of the HAM-10000 dataset. The result shows that overall performance gains are reported regardless of the target set X. Such results show the effectiveness of the proposed method in that it does not require a domain-specific dataset, but instead can be implemented using an arbitrary dataset. We anticipate that such property could facilitate the application of the proposed method to diverse medical domains, where acquiring large-scale domain-specific target images is expensive.

5.3   Qualitative Examples

In Figure 3, the final prediction of a model and three concepts with the highest influence P_y in predicting the ground-truth class y are illustrated. As depicted in the figure, the baseline fails correctly classifying the image since the concepts with the highest influences in predicting the ground-truth class are mostly non-visual concepts *e.g.*, "early detection and treatment of skin lesions can help prevent skin cancer", "lesions may be associated with other health problems" which does not help classify an image. In contrast, the prediction result on the same image is corrected when applying the proposed method since visually irrelevant concepts are removed from S_y, thus concepts with the highest influences are replaced with concepts with visual cues. The result validates that adding the

		baseline	**ours**
basal cell carcinoma	Top 1:	early detection and treatment of skin lesions can help prevent skin cancer	surrounded by **red, inflamed skin**
	Top 2:	more advanced lesions may be **darker**	in some cases, the skin lesions may be **darker in color**
	Top 3:	most common type of precancerous lesion in the united states	a type of skin cancer that typically appears as a **small, pearly-white** or
	Pred:	actinic keratoses **(X)**	**basal cell carcinoma (O)**
benign keratosis-like lesions	Top 1:	lesions may be associated with other health problems	**waxy or crusted surface**
	Top 2:	lesions may be **surrounded by an inflammatory halo**	**smooth, waxy surface**
	Top 3:	**fluid is clear**	**slightly rough surface**
	Pred:	vascular lesions **(X)**	**benign keratosis-like lesions (O)**
melanocytic nevi	Top 1:	some people with skin lesions may require lifelong treatment	**black, brown, or any color that is darker** than the surrounding skin
	Top 2:	lesions may be associated with other health problems	some nevi may also itch or **bleed**
	Top 3:	lesions may be **surrounded by a halo of lighter skin**	**lighter or darker** than the surrounding skin
	Pred:	vascular lesions **(X)**	**melanocytic nevi (O)**

Fig. 3: **Qualitative results on HAM-10000 dataset.** Several comparisons with the baseline, LaBo, are shown with the top-3 concepts which are ranked by their weights in the linear function. Parts that provide visual cues are colored **blue**.

visual activation score $\mathcal{V}(c)$ to the score function $\mathcal{F}$ helps filter visually irrelevant concepts, therefore contributing to better classification results by providing concepts that are rich in information.

6   Conclusion

In this paper, we propose a method to refine the non-visual concepts generated from large language models (LLMs) which hinder the training of Concept Bottleneck Models (CBMs). In order to filter out non-visual concepts, we propose the visual activation score which measures whether a concept contains visual information or not. With computed visual activation scores of concepts, non-visual concepts are filtered out via a submodular optimization. Quantitative and qualitative analyses demonstrate that the proposed visual activation score contributes to detecting and filtering out non-visual concepts, therefore resulting in consistent improvement in accuracy.

Acknowledgments

References

1. Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 2018.
2. Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 2018.
3. Jiejie Zhou, Yang Zhang, Kai-Ting Chang, Kyoung Eun Lee, Ouchen Wang, Jiance Li, Yezhi Lin, Zhifang Pan, Peter Chang, Daniel Chow, et al. Diagnosis of benign and malignant breast lesions on dce-mri by using radiomics and deep learning with consideration of peritumor tissue. *Journal of Magnetic Resonance Imaging*, 2020.
4. Yanping Xue, Rongguo Zhang, Yufeng Deng, Kuan Chen, and Tao Jiang. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLOS ONE*, 2017.
5. Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.
6. Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *NeurIPS*, 2022.
7. Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023.
8. Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.
9. Francis Bach. Convex analysis and optimization with submodular functions: a tutorial. *arXiv preprint arXiv:1010.4207*, 2010.
10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.
11. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.
12. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
13. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

14. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
15. Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 2018.
16. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
17. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.