On the Concept Trustworthiness in Concept Bottleneck Models

Qihan Huang[1], Jie Song[1*], Jingwen Hu[1], Haofei Zhang[1], Yong Wang[2], Mingli Song[1]

[1] Zhejiang University
[2] State Grid Shandong Electric Power Company
{qh.huang, sjie, jw_hu, haofeizhang, brooksong}@zju.edu.cn, wangyong@sd.sgcc.com.cn

Abstract

Concept Bottleneck Models (CBMs), which break down the reasoning process into the *input-to-concept* mapping and the *concept-to-label* prediction, have garnered significant attention due to their remarkable interpretability achieved by the interpretable concept bottleneck. However, despite the transparency of the concept-to-label prediction, the mapping from the input to the intermediate concept remains a black box, giving rise to concerns about the trustworthiness of the learned concepts (*i.e.*, these concepts may be predicted based on spurious cues). The issue of concept untrustworthiness greatly hampers the interpretability of CBMs, thereby hindering their further advancement. To conduct a comprehensive analysis on this issue, in this study we establish a benchmark to assess the trustworthiness of concepts in CBMs. A pioneering metric, referred to as *concept trustworthiness score*, is proposed to gauge whether the concepts are derived from relevant regions. Additionally, an enhanced CBM is introduced, enabling concept predictions to be made specifically from distinct parts of the feature map, thereby facilitating the exploration of their related regions. Besides, we introduce three modules, namely the *cross-layer alignment* (CLA) module, the *cross-image alignment* (CIA) module, and the *prediction alignment* (PA) module, to further enhance the concept trustworthiness within the elaborated CBM. The experiments on five datasets across ten architectures demonstrate that without using any concept localization annotations during training, our model improves the concept trustworthiness by a large margin, meanwhile achieving superior accuracy to the state-of-the-arts. Our code is available at https://github.com/hqhQAQ/ProtoCBM.

Introduction

Concept Bottleneck Models (CBMs) have recently emerged as self-explanatory models that begin by predicting the high-level concepts present in the input data (*input-to-concept mapping*), and subsequently make class label predictions based on these inferred concepts (*concept-to-label prediction*). CBMs have garnered substantial attention from researchers due to their inherent interpretability and comparable performance when compared to non-interpretable models. Following the pioneering work of the first CBM (Koh et al. 2020), a plethora of CBM variants have emerged in the research landscape. These variants include, but are not limited to, PCBM (Yüksekgönül, Wang, and Zou 2023), Label-free CBM (Oikarinen et al. 2023), Hard CBM (Havasi, Parbhoo, and Doshi-Velez 2022), and Interactive CBM (Chauhan et al. 2023).

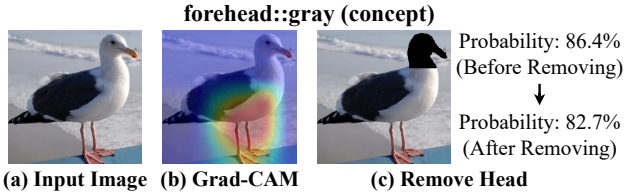

Figure 1: Untrustworthiness of a concept named "forehead::gray" in the vanilla CBM. (a) The input image. (b) The localization map of this concept (generated by Grad-CAM) concentrates on the underpart of the bird. (c) After removing the head part in the image, the prediction probability for this concept changes very slightly.

Despite the remarkable strides made in the field, the interpretability of existing CBMs primarily stems from the transparency of the concept-to-label prediction. The input-to-concept mapping, in contrast, is still a black box and thus remains elusive, which significantly undermines the trustworthiness of the subsequent concept-to-label prediction. Figure 1 illustrates an example highlighting this issue. Specifically, Figure 1 (b) demonstrates the attribution map of a head concept prediction using Grad-CAM (Selvaraju et al. 2017), indicating that the prediction of this head concept is based on the underpart rather than the head part. Besides, Figure 1 (c) further presents that after removing the head part in the input image, the classification probability of this head concept only changes slightly. While some recent studies have shed light on the concept untrustworthiness in CBMs (Margeloiu et al. 2021; Heidemann, Monnet, and Roscher 2023; Furby et al. 2023), these works have been limited to either visualization examples or a simple metric applied to a specific CBM on a single backbone. A more comprehensive investigation into concept trustworthiness is urgently needed to drive the progress of CBMs.

In this work, we establish a systematic benchmark on

*Corresponding author.

concept trustworthiness with a proposed evaluation metric termed *concept trustworthiness score*. This metric assesses to what extent the concepts predicted by CBMs are align with the annotated object parts in the dataset. To ensure a comprehensive evaluation, our benchmark encompasses various CBM variants implemented across multiple architectures, including a diverse range of CNNs and vision transformers. The experimental results obtained from this benchmark reveal a significant insufficiency in the concept trustworthiness of previous CBMs, thereby highlighting the limitations in their interpretability.

To further enhance the concept trustworthiness, we propose an elaborated CBM framework that deviates from the conventional approach of utilizing average pooling on the final feature map extracted from the backbone network. The average pooling operation mixes the features of different image regions together and misleads the concept predictions, thus undermining the overall concept trustworthiness. Therefore, we draw inspiration from the ProtoPNet (Chen et al. 2019) framework and eliminate the pooling operation altogether. In our proposed approach, we define multiple part-prototypes as trainable vectors to represent different object parts. We then iterate over different regions of the feature map to identify and evaluate the presence of these part-prototypes. The activation value of each part-prototype, which signifies the existence of the represented object part in the feature map, is calculated accordingly. Subsequently, a fully-connected layer is employed to make concept predictions based on the activation values of the part-prototypes.

Besides, we propose three modules into the above CBM framework to further improve the concept trustworthiness: a cross-layer alignment (CLA) module, a cross-image alignment (CIA) module, and a prediction alignment (PA) module. The above CBM framework locates the related image region of each predicted concept in two steps: (1) The prototypes faithfully locate their corresponding object parts in the last feature map; (2) Each concept accurately matches the prototypes that represent its related image regions. The first step requires that the last feature map is spatially aligned with the input image, and the CLA & CIA modules are proposed to approach this requirement. Specifically, the CLA module adopts a multi-scale mechanism to facilitate the spatial alignment between the feature maps in the deep and shallow layers. The CIA module promotes the spatial alignment between the feature maps of the original image and the augmented image. Furthermore, to improve the second step, the PA module is proposed to constrain the localization regions (generated with the matched prototypes) of concepts to be consistent.

We perform comprehensive experiments to validate the performance of our proposed model. Experiment results demonstrate that without using any concept localization annotations during training, our proposed CBM framework and three modules significantly improve the concept trustworthiness. Besides, with the decoupled concept learning mechanism that concentrates the learning of concepts, our model achieves state-of-the-art accuracy on five datasets across ten architectures.

To sum up, the key contributions of our work can be listed as follows:

- We establish a systematic benchmark to evaluate the concept trustworthiness with a proposed evaluation metric (concept trustworthiness score) across multiple backbones, unveiling the cons of various CBMs.
- We introduce an elaborated CBM model that decouples the feature map for concept prediction, with three proposed modules to further improve the concept trustworthiness: a cross-layer alignment (CLA) module, a cross-image alignment (CIA) module, and a prediction alignment (PA) module.
- Experiment results show that our model achieves state-of-the-art performance, in both concept trustworthiness and accuracy, on five datasets across ten architectures.

Related Work

Concept Bottleneck Models. Concept bottleneck models (CBMs) are self-explainable models that make class predictions based on the prior predicted concepts. With the intermediate human-understandable concepts inside the model, CBMs enable humans to interpret the original black-box model and quickly intervene on the concept prediction for higher accuracy. After the emergence of the vanilla CBM (Koh et al. 2020), various variants of CBMs have been proposed. Label-free CBM utilizes large-scale language model and cross-modal model, GPT-3 and CLIP (Radford et al. 2021), to automatically generate concept embeddings, thus reducing the cost concept labeling in CBMs. PCBM aims to turn any pre-trained deep neural network into a CBM, by leveraging CAVs (Concept Activation Vectors) (Kim et al. 2018) to represent concepts and projecting the image features into these learned CAVs for concept prediction. However, recently some works (Margeloiu et al. 2021; Heidemann, Monnet, and Roscher 2023; Furby et al. 2023) point out that many learned concepts of CBMs are not predicted from the related image regions, thus weakening the interpretability of CBMs. Therefore, our work aims to establish a systematic benchmark to evaluate the concept trustworthiness of CBMs and propose an enhanced CBM to improve the concept trustworthiness.

Part-Prototype Networks. Part-prototype networks are self-explainable models that make class predictions through intermediate interpretable part-prototypes. Specifically, they define multiple part-prototypes to represent object parts, and mimic humans to make predictions by comparing object parts across different images. ProtoPNet (Chen et al. 2019) is the first part-prototype network, with many follow-up part-prototype networks: ProtoTree (Nauta, van Bree, and Seifert 2021), Deformable ProtoPNet (Donnelly, Barnett, and Chen 2022), TesNet (Wang et al. 2021), ProtoPFormer (Xue et al. 2022), ProtoPShare (Rymarczyk et al. 2021), and ProtoPool (Rymarczyk et al. 2022). Our work follows part-prototype networks to decouple the feature map and make concept predictions from different parts of the feature map.

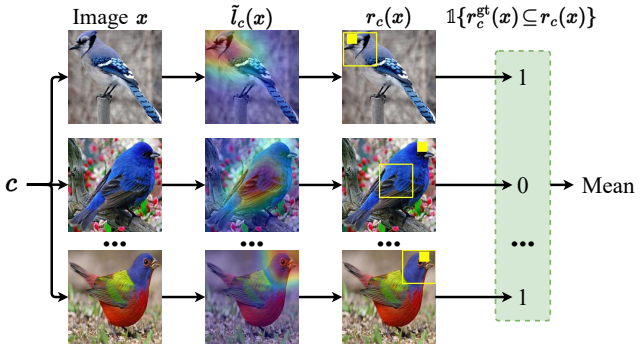

Figure 2: The calculation of concept trustworthiness score of a concept c about the forehead of the bird.

Concept Trustworthiness Benchmark

Preliminaries

The vanilla CBM consists of a feature extractor f, a concept predictor g, and a category predictor h. Specifically, given an input image x with height H and width W, f extracts the feature map $z \in \mathbb{R}^{H_z \times W_z \times D}$ (H_z, W_z, D denote the height, width and dimension of the feature map z). Next, CBM employs average pooling on z to obtain $\bar{z} \in \mathbb{R}^D$, then feeds $\bar{z}$ into the concept predictor g to generate the concept classification probabilities $g(\bar{z}) \in \mathbb{R}^C$ (C denotes the total number of concepts). Finally, CBM feeds $g(\bar{z})$ into the category predictor h to generate the classification probabilities $h(g(\bar{z})) \in \mathbb{R}^K$ (K denotes the total number of categories).

CBM adopts two losses for model training: a concept loss $\mathcal{L}_{\text{concept}}(g(\bar{z}), c^{\text{gt}})$ to supervise the concept prediction, and a task loss $\mathcal{L}_{\text{task}}(h(g(\bar{z})), y^{\text{gt}})$ to supervise the category prediction. Here, $c^{\text{gt}} \in \mathbb{R}^C$ denotes the concept label, *i.e.*, $c_i^{\text{gt}} = 1$ if the i-th concept exists in the image and $c_i^{\text{gt}} = 1$ otherwise, and y^{gt} denotes the category label. In the later sections, this paper uses $\mathcal{L}_{\text{concept}}$ and $\mathcal{L}_{\text{task}}$ instead of $\mathcal{L}_{\text{concept}}(g(\bar{z}), c^{\text{gt}})$ and $\mathcal{L}_{\text{task}}(h(g(\bar{z})), y^{\text{gt}})$ for simplicity.

Concept Trustworthiness Score

Our work establishes a systematic concept trustworthiness benchmark with an evaluation metric named *concept trustworthiness score*. (**First step**) To calculate this evaluation metric, our work first generates the corresponding region of each concept on the input image (*i.e.*, the image region that the concept is predicted from). (**Second step**) Next, this evaluation metric estimates the concept trustworthiness according to whether the corresponding region of the concept is consistent with the ground-truth (*i.e.*, the object part annotations in the dataset).

In the first step, our work first calculates the localization map $l_c(x) \in \mathbb{R}^{H_l \times W_l}$ of each concept c on the input image x (H_l & W_l denote the shape of $l_c(x)$, and $l_c(x)$ indicates the important regions for concept prediction in the input image). For example, for the previous CBMs which lack intrinsic concept localization ability, our work calculates $l_c(x)$ using the attribution methods (*e.g.*, Grad-CAM (Selvaraju et al. 2017), Grad-CAM++ (Chattopadhyay et al. 2018)). Next, our work follows ProtoPNet (Chen et al. 2019) to resize $l_c(x)$ to be $\tilde{l}_c(x) \in \mathbb{R}^{H \times W}$ with the same shape as x, then calculates the corresponding region $r_c(x)$ as a fix-sized bounding box (with a pre-determined shape $H_b \times W_b$) whose center is the maximum element in $\tilde{l}_c(x)$.

In the second step, our work determines the trustworthiness of concept c according to whether the ground-truth region $r_c^{\text{gt}}(x)$ of concept c is inside $r_c(x)$, which can be described as $\mathbb{1}\{r_c^{\text{gt}}(x) \subseteq r_c(x)\}$ ($\mathbb{1}\{\cdot\}$ denotes the indicator function). Our work estimates the trustworthiness of each concept c on all the images that contain c, then calculates the averaged results over total C concepts. Let $\mathcal{I}_c$ denote the images that contain c, the concept trustworthiness score S_{concept} is finally defined as (note that $\|\cdot\|$ denotes cardinality of a set, and $S_{\text{concept}} \in [0, 1]$):

$$S_{\text{concept}} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{\|\mathcal{I}_c\|} \sum_{x \in \mathcal{I}_c} \mathbb{1}\{r_c^{\text{gt}}(x) \subseteq r_c(x)\}. \quad (1)$$

With this proposed evaluation metric, a systematic benchmark can be established for various CBM variants across multiple backbones. Besides, the proposed benchmark normalizes the object sizes (*i.e.*, by cropping the objects and resizing them to be the same size), which eliminates the evaluation distractors for a more normative benchmark (*e.g.*, if the object size is too small in the image, the corresponding regions of different concepts are easily confounded).

Enhanced CBM

The experiments on this concept trustworthiness benchmark demonstrate that previous CBMs have insufficient concept trustworthiness. Our work speculates that their concept untrustworthiness results from the average pooling operation on the last feature map. Specifically, the average pooling operation mixes the features of different image regions together for the further concept prediction, thus easily misleading concepts into learning from their unrelated image regions. Therefore, our work establishes an elaborated **CBM framework**, following ProtoPNet (Chen et al. 2019) to eliminate the pooling operation and make predictions from specific parts of the feature map using the learnable prototypes. Furthermore, we propose three modules into this framework for higher concept trustworthiness: a **cross-layer alignment (CLA) module**, a **cross-image alignment (CIA) module**, and a **prediction alignment (PA) module**.

CBM Framework

The proposed CBM framework contains M learnable prototypes $\mathbf{P} = \{\mathbf{p}_j \in \mathbb{R}^{1 \times 1 \times D}\}_{j=1}^{M}$. Given the feature map $z \in \mathbb{R}^{H_z \times W_z \times D}$ extracted from the input image x, our work generates the localization map $l_{\mathbf{p}_j}(x) \in \mathbb{R}^{H_z \times W_z}$ of each prototype $\mathbf{p}_j$ on z by calculating and concatenating the similarity scores between $\mathbf{p}_j$ and each element $\tilde{z}$ of z (z consists of $H_z \times W_z$ elements, and the shape of each element is D). Next, the activation value $a_{\mathbf{p}_j}(x)$ of $\mathbf{p}_j$ on z is calculated as the maximum value in $l_{\mathbf{p}_j}(x)$ ($\text{Sim}(\cdot, \cdot)$ is the similarity score between two vectors):

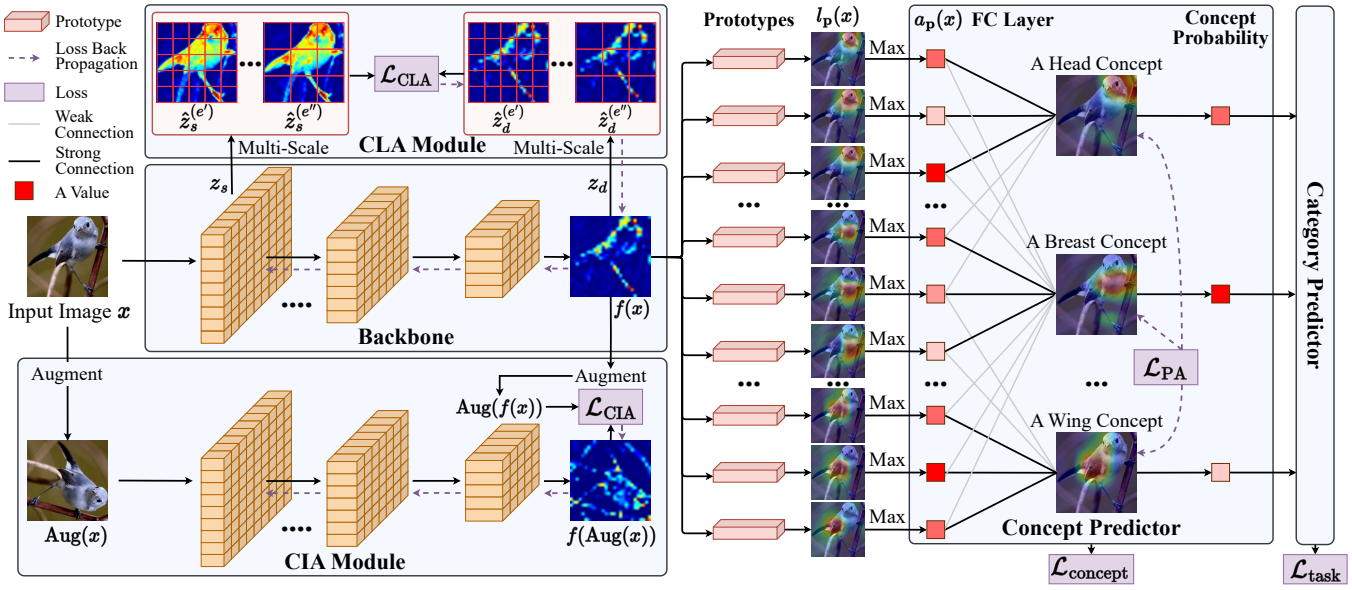

Figure 3: Overview of our proposed model (only three concepts are presented for brevity). Given the input image x, the "Backbone" extracts different layers of features for x. The CLA module spatially aligns the last feature map $f(x)$ (also denoted as z_d) with the input image according to the shallow feature map z_s in a multi-scale manner. Meanwhile, the CIA module spatially aligns $f(x)$ with x by aligning $f(x)$ with the feature map of another augmented image. Next, $f(x)$ is aggregated with multiple prototypes, generating the localization maps $l_{\mathbf{p}}(x)$ and activation values $a_{\mathbf{p}}(x)$ of prototypes. Finally, the activation values $a_{\mathbf{p}}(x)$ are fed into "Concept predictor" and "Category Predictor" for concept prediction and category prediction, respectively. The PA loss is used to supervise the localization maps of the learned concepts. Note that the loss back propagation of $\mathcal{L}_{\text{concept}}$ and $\mathcal{L}_{\text{task}}$ is omitted for simplicity in the figure.

$$
\begin{aligned}
a_{\mathbf{p}_j}(x) &= \max\ l_{\mathbf{p}_j}(x) \\
&= \max_{\tilde{z}\ \in\ \text{elements}(z)} \text{Sim}(\tilde{z}, \mathbf{p}_j).
\end{aligned} \tag{2}
$$

In this manner, each $a_{\mathbf{p}_j}(x)$ is calculated from a specific part of the feature map, thus the learned prototypes can represent the features of the local part instead of the global object. Therefore, our work makes concept predictions according to these learned prototypes, facilitating the concepts to explore their related prototypes and the corresponding image regions. Specifically, with the calculated activation values $\{a_{\mathbf{p}_j}(x)\}_{j=1}^{M}$ of M prototypes, the concept predictor g can be simply implemented as a fully-connected layer with a weight matrix $\omega^g \in \mathbb{R}^{C\times M}$, which takes $\{a_{\mathbf{p}_j}(x)\}_{j=1}^{M}$ as input. Finally, the concept classification probabilities $g(\{a_{\mathbf{p}_j}(x)\}_{j=1}^{M})$ output by the concept predictor are fed into the category predictor to generate the classification probabilities. This CBM framework is also trained with the concept loss $\mathcal{L}_{\text{concept}}$ and the task loss $\mathcal{L}_{\text{task}}$.

After training, our work calculates the localization map $l_c(x)$ of each concept c by averaging the localization maps of the N most important prototypes for c. Specifically, let $\text{Top}_N(\omega_c^g)$ denote the indexes of the N most important prototypes for c (*i.e.*, these N prototypes contribute most positively to the prediction of c in the concept predictor, which is determined by their weights in $\omega_c^g \in \mathbb{R}^M$), $l_c(x)$ is calculated as in Equation 3 and it is used for concept trustworthiness evaluation of this CBM framework ($l_c(x) \in \mathbb{R}^{H_z\times W_z}$, note that $H_l = H_z$ here).

$$
l_c(x) = \frac{1}{N} \sum_{j\in \text{Top}_N(\omega_c^g)} l_{\mathbf{p}_j}(x). \tag{3}
$$

The proposed CBM framework significantly improves the concept trustworthiness by decoupling the feature map for concept prediction, however, it is still restricted. In this framework, the $l_c(x)$ locates the related image region of c in two steps: (1) The localization map $l_{\mathbf{p}_j}(x)$ of each prototype $\mathbf{p}_j$ faithfully locates its corresponding object part in the last feature map z; (2) The the concept predictor assigns higher positive values to the prototypes in ω_c^g that represent the related object part for c, *i.e.*, c is predicted mainly from these prototypes. Specifically, the first step requires that the last feature map z is spatially aligned with the input image x, *i.e.*, each pixel of the feature map represents the information of the image region with the same position in the input image. However, this requirement is not guaranteed in the deep neural networks, and some works (Hoffmann et al. 2021; Huang et al. 2023) doubt the interpretability of ProtoPNet due to this reason. Therefore, our work proposes the cross-layer alignment module and the cross-image alignment module to tackle this problem. Furthermore, our work proposes the prediction alignment module to improve the second step.

Cross-Layer Alignment Module

To promote the spatial alignment between the last feature map with the input image, the cross-layer alignment (CLA) module adopts a multi-scale mechanism to align the last feature map $z_d \in \mathbb{R}^{H_d \times W_d \times D_d}$ with the shallow feature map $z_s \in \mathbb{R}^{H_s \times W_s \times D_s}$ (because z_s is closer to the input pixel space). Specifically, the CLA module aligns the pair-wise element similarities within z_d and z_s, according to that intrinsic similarities between elements within the feature maps can be used for comparison (Kornblith et al. 2019; Huang et al. 2023). Let $\hat{z}_d \in \mathbb{R}^{H_d W_d \times D_d}$ and $\hat{z}_s \in \mathbb{R}^{H_d W_d \times \hat{D}_s}$ (note that $\hat{D}_s = D_s \cdot \frac{H_s}{H_d} \cdot \frac{W_s}{W_d}$) denote the resized z_d and z_s, then each element $\phi_{i,j}(\hat{z}_d)$ of the pair-wise element similarities $\phi(\hat{z}_d) \in \mathbb{R}^{H_d W_d \times H_d W_d}$ is calculated as $\mathrm{Sim}(\hat{z}_d[i], \hat{z}_d[j])$, and $\phi(\hat{z}_s)$ is calculated likewise.

However, current form of $\hat{z}_d$ and $\hat{z}_s$ only excavates the object similarities at a single scale and neglects necessary object similarities at other scales. For instance, it is essential to consider both the similarity between the left eye and right eye at a smaller scale, as well as the dissimilarity between the head part and the wing part at a larger scale of a bird object. Therefore, our work enriches the feature map to be multi-scale to tackle this problem. Detailedly, let E denote the total levels of scales, then for each level $e \in \{1, 2, ..., E\}$, our work concatenates the feature vectors within every window (with size $e \times e$) in the original feature map z to generate the feature map $z^{(e)}$. With the enriched feature maps $\{\hat{z}_d^{(e)}\}_{e=1}^{E}$ and $\{\hat{z}_s^{(e)}\}_{e=1}^{E}$, the alignment loss $\mathcal{L}_{\mathrm{CLA}}$ is calculated as (note that $\|\cdot\|_2$ denotes the L2 norm, and $\mathrm{Detach}(\cdot)$ is the detach operation to stop the gradients):

$$\mathcal{L}_{\mathrm{CLA}} = \frac{1}{E} \sum_{e=1}^{E} \|\phi(\hat{z}_d^{(e)}) - \mathrm{Detach}\big(\phi(\hat{z}_s^{(e)})\big)\|_2^2. \quad (4)$$

Cross-Image Alignment Module

The cross-image alignment (CIA) module spatially aligns the last feature map with the input image in a self-supervised manner. Specifically, given the input image x, the CIA module employs spatial augmentation Aug (*e.g.*, horizontal flip, rotation) on x to generate the augmented image $\mathrm{Aug}(x)$. Next, the CIA module employs the same spatial augmentation on the feature map $f(x)$ of the original image x, and constrains the augmented feature map $\mathrm{Aug}(f(x))$ to be consistent with the feature map $f(\mathrm{Aug}(x))$ of the augmented image. In this training manner, the feature map is promoted to perceive the spatial information of the images, thus better aligning with the input image. In practice, the CIA module freezes the feature map $f(x)$ of the original image x to prevent it from slipping into a trivial solution (*e.g.*, the parameters of feature extractor f become zero). Therefore, the alignment loss $\mathcal{L}_{\mathrm{CIA}}$ is finally calculated as:

$$\mathcal{L}_{\mathrm{CIA}} = \|f(\mathrm{Aug}(x)) - \mathrm{Detach}\big(\mathrm{Aug}(f(x))\big)\|_2^2. \quad (5)$$

Prediction Alignment Module

With the spatially aligned feature map, each concept c is promoted to be predicted from its related image region with the concept loss $\mathcal{L}_{\mathrm{concept}}$, because the features of the related object part of c lead to higher prediction accuracy of c. However, some concepts are still predicted wrongly from the unrelated image regions if they are confused with other concepts (*i.e.*, two concepts are easily confused if they often appear simultaneously in the same images). Therefore, the prediction alignment (PA) module adopts a concept grouping loss $\mathcal{L}_{\mathrm{grp}}$ and a concept division loss $\mathcal{L}_{\mathrm{div}}$ to align the concept prediction. The concept grouping loss $\mathcal{L}_{\mathrm{grp}}$ facilitates the concepts with the same related region to be predicted from the same region, and the concept division loss $\mathcal{L}_{\mathrm{div}}$ encourages the concepts with different related regions to be predicted from different regions.

Specifically, let $\{\mathcal{C}_i\}_{i=1}^{T}$ denote T groups of concepts, where each $\mathcal{C}_i$ contains concepts with the same related region, and $\mathcal{C}_i$ and $\mathcal{C}_j$ have different related regions for $\forall i \neq j$. Besides, for a localization map $l_c(x) \in \mathbb{R}^{H_z \times W_z}$, the coordinates $k_c(x) \in \mathbb{R}^2$ of the center in $l_c(x)$ are calculated as the weighted average of the multiplication of each element in $l_c(x)$ and its coordinates. Finally, the alignment loss $\mathcal{L}_{\mathrm{PA}}$ in the PA module is calculated as below:

$$\begin{cases} \mathcal{L}_{\mathrm{PA}} = \mathcal{L}_{\mathrm{grp}} + \mathcal{L}_{\mathrm{div}}. \\ \mathcal{L}_{\mathrm{grp}} = \frac{1}{T} \sum_{i=1}^{T} \sum_{c \in \mathcal{C}_i} \sum_{c' \in \mathcal{C}_i, c' \neq c} \|k_c(x) - k_{c'}(x)\|_2^2. \\ \mathcal{L}_{\mathrm{div}} = -\frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1, j \neq i}^{T} \sum_{c \in \mathcal{C}_i} \sum_{c' \in \mathcal{C}_j} \|k_c(x) - k_{c'}(x)\|_2^2. \end{cases}$$

Loss Function

With the above modules, the total loss function $\mathcal{L}_{\mathrm{total}}$ of our proposed CBM model is finally calculated as below:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{concept}} + \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{CLA}} + \mathcal{L}_{\mathrm{CIA}} + \mathcal{L}_{\mathrm{PA}}. \quad (6)$$

Experiments

Experimental Settings

Datasets. We follow previous CBMs to conduct experiments on CUB-200-2011 (Wah et al. 2011), HAM10000 (Tschandl, Rosendahl, and Kittler 2018), and CIFAR10 & CIFAR100 (Krizhevsky, Hinton et al. 2009). In particular, CUB-200-2011 contains concept annotations and location annotations of the object parts for each image, including 312 concepts and 15 object parts that cover the whole body of the bird. Therefore, our proposed concept trustworthiness benchmark is mainly established on CUB-200-2011. Specifically, we select the concepts that relate to object parts (*e.g.*, head shape, wing color) for the concept trustworthiness benchmark. Besides, we crop the bird part in the images according to the bounding box annotations in CUB-200-2011, thus eliminating the evaluation distractors to establish a normative benchmark.

Benchmark Setup. Our benchmark evaluates five CBMs: vanilla CBM (Koh et al. 2020), Hard CBM (Havasi, Parbhoo, and Doshi-Velez 2022), Label-free CBM (Oikarinen et al. 2023), PCBM (Yüksekgönül, Wang, and Zou 2023), and ours. The concept labels are pre-processed following the

Method	ResNet18		ResNet34		ResNet152		Dense121		Dense161		DeiT-Ti		DeiT-S		Swin-S	
	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.	Loc.	Acc.
Baseline	N/A	78.8	N/A	82.3	N/A	81.5	N/A	80.5	N/A	82.2	N/A	81.2	N/A	82.2	N/A	83.4
vanilla CBM	22.6	77.2	21.1	79.1	18.9	79.5	24.6	78.8	26.2	80.4	13.7	78.3	12.4	79.6	11.7	81.5
Hard CBM	26.4	76.9	25.2	80.4	20.6	81.0	28.1	77.9	20.7	78.9	15.8	79.4	12.9	80.6	10.4	82.0
Label-free CBM	29.5	77.8	32.3	79.8	28.1	78.7	31.6	78.4	28.9	79.3	18.4	77.6	14.4	78.4	12.0	81.2
PCBM	34.1	77.3	36.9	80.5	40.1	80.8	46.4	79.3	40.5	80.7	16.2	78.6	10.7	79.1	12.7	82.2
Ours	44.6	77.2	43.2	80.3	47.2	**82.2**	45.9	**81.3**	45.1	82.2	34.7	81.4	34.5	81.1	30.2	83.5
Ours + CLA	51.2	77.8	52.4	80.0	50.4	82.1	49.6	80.8	48.6	82.4	36.5	81.2	37.4	81.0	35.3	82.8
Ours + CLA + PA	57.6	77.9	64.1	80.4	64.8	81.9	65.5	80.6	63.2	**82.7**	44.7	81.0	43.1	80.8	38.9	83.5
Ours + CLA + PA + CIA	**62.8**	**78.6**	**70.5**	**81.1**	**67.6**	82.1	**70.2**	81.0	**65.4**	82.6	**47.2**	**81.4**	**46.3**	**81.5**	**42.4**	**83.7**

Table 1: The comprehensive evaluation of concept trustworthiness and accuracy of CBMs on CUB-200-2011 dataset. The results are over eight backbones pre-trained on ImageNet. Loc. and Acc. denote concept trustworthiness score and accuracy, respectively. Our results are averaged over 4 runs with different seeds. Bold font denotes the best result in CBMs.

Method	CUB	HAM10000	CIFAR10	CIFAR100
PCBM	58.8	94.7	77.7	52.0
Ours	**68.3**	**96.6**	**81.3**	**53.8**
PCBM-h	61.0	96.2	87.1	68.0
Ours-h	**69.4**	**96.4**	**88.2**	**69.1**

Table 2: Accuracy comparison with PCBM & PCBM-h on four datasets. Bold font denotes the best result.

vanilla CBM. The results are re-implemented faithfully following their released codes.

Details. In the proposed benchmark, H_b and W_b are both set to be 90 for all CBMs. In the proposed CBM framework, M, D, E, N are set to be 2000, 64, 2, 10. We train the proposed model for 18 epochs (5 epochs for warm-up) with Adam optimizer (Kingma and Ba 2015), and the learning rate of the model is set to be 1e-4. In the early stage training, we follow the paradigm of ProtoPNet to train the model.

Concept Trustworthiness Benchmark

With the proposed concept trustworthiness score, we demonstrate the evaluation results of the concept trustworthiness benchmark in Table 1 (over eight backbones pre-trained on ImageNet). In the table, "Baseline" is the simplest non-interpretable with a fully-connected layer on the last feature map for classification. We select Grad-CAM++ (Chattopadhyay et al. 2018) to generate the localization maps for previous CBMs, because it has the best performance (the ablation experiments on the choice of CAM methods are in **S2.1** of the appendix).

As shown in the table, the concept trustworthiness of previous CBMs are quite insufficient to support their interpretability. For example, the concept trustworthiness score of the vanilla CBM ranges from 11.7 to 26.2, meaning that most of its learned concepts are not trustworthy. Specifically, each learned concept in the vanilla CBM is not consistent with the human perception (because it is inferred from its unrelated image regions), which brings serious potential risks to the subsequent concept-based classification.

Comparisons with State-of-the-Art Methods

The performance of our proposed framework and modules is also demonstrated in Table 1 ("Ours" denotes the proposed CBM framework based on part-prototypes). As shown in the table, the CBM framework achieves superior performance to the previous CBMs in both concept trustworthiness score and accuracy, owing to the concept prediction mechanism that acquires prediction clues from specific parts of the feature map. Next, with the proposed CLA, CIA and PA modules, our model achieves significantly higher concept trustworthiness score (Averagely it is 40.2 points higher than the vanilla CBM).

Furthermore, we also validate the accuracy of our model on the benchmark proposed by the previous CBM. Table 2 demonstrates the performance comparison with PCBM on four datasets: CUB, HAM10000, CIFAR10, and CIFAR100. Specifically, the backbones for CUB, HAM10000, CIFAR10 & CIFAR100 are ResNet18, Inception-v3, and CLIP, respectively. Note that the performance on CUB is significantly lower than the results in Table 1 because the CUB dataset contains only partial training data and the images are not cropped here (following the setting in the released codes of PCBM). Compared to PCBM, PCBM-h adopts an auxiliary fully-connected layer on the image features for classification, which can also be utilized in our model ("Ours-h"). The comparison results indicate that our model achieves better accuracy than PCBM on these datasets.

Ablation Experiments

Table 1 demonstrates the effectiveness of our proposed CLA, CIA and PA modules. Specifically, the proposed modules significantly improve the concept trustworthiness score of CBM, while achieving comparable or even better accuracy. Furthermore, in **S2.2** & **S2.3** of the appendix, we conduct two more ablation experiments to verify that: (1) The CLA & CIA modules significantly improve the alignment between the deep feature maps and shallow feature maps under the spatial transformation on the original images; (2) The PA module effectively facilitates the concepts with the same related regions to be predicted from the same region, and encourages the concepts with different related regions to be predicted from different regions.

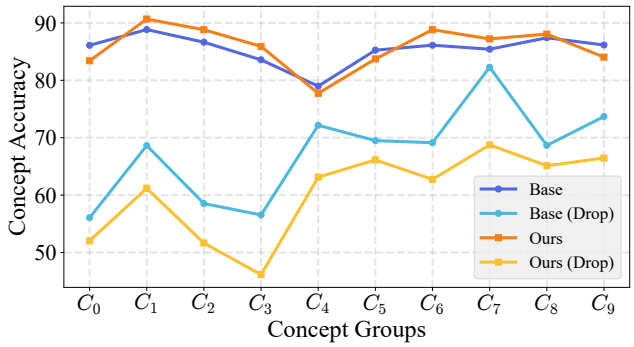

Figure 4: The concept prediction accuracy of our model is similar with the base model before dropping the related regions, while it significantly decreases after dropping the related regions ("Ours (Drop)").

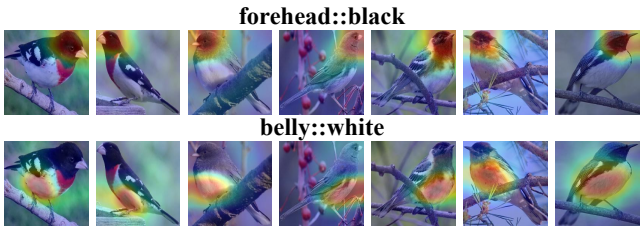

Figure 5: The localization maps of two concepts ("forehead::black" and "belly::white") generated by our model, which accurately locate the related regions of the concepts.

Patch Drop Experiment. We conduct a patch drop experiment to verify that the concepts are rescued from being predicted from the unrelated regions in our proposed model. Specifically, we evaluate the concept prediction accuracy before/after dropping the related image regions (by setting the pixels of these regions to be zero, according to the part localization annotations and object segmentation annotations, as described in **S2.4** of the appendix). The experiment is conducted over all T ($T = 10$) concept groups (each concept group $\mathcal{C}_i$ contains concepts with the same related region), and the CBM framework without/with the proposed modules ("Base", "Ours") are evaluated. As shown in Figure 4, our model has the similar concept prediction accuracy with the base model before dropping the related regions. However, after dropping the related regions, the concept prediction accuracy of our model decreases significantly, indicating that more concepts in our model are predicted from the related regions.

Visualization

The localization maps of concepts. Figure 5 shows the localization maps $l_c(x)$ of two concepts ("forehead::black" and "belly::white") generated by our proposed model, indicating that our proposed model accurately makes concept predictions from their related image regions.

Interpretable classification of the image. Figure 6 demonstrates the interpretable classification of a Glaucous Winged Gull in our proposed model. Specifically, our model first determines the localization map of a group of concepts, then predicts the concept probabilities according to the localization maps. Finally, the classification probability of this bird is calculated by aggregating the predicted concept probabilities through a fully-connected layer. In this figure, three concepts ("forehead::white", "wing::grey", "belly::pattern") with the highest classification weights (in the fully-connected layer) are listed.

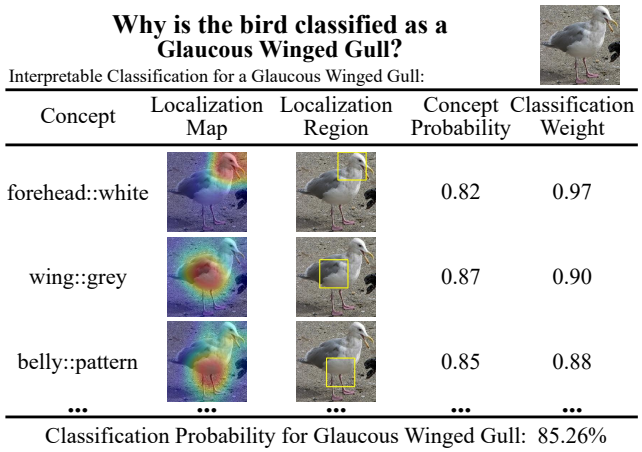

Figure 6: Our model first determines the localization map of the concepts, then makes concept predictions on them, and finally predicts the category from the concept probabilities.

We also provide more visualization analysis on our proposed model in **S3** of the appendix for a comprehensive understanding of our method.

Conclusion

Our work aims to address the concept untrustworthiness problem in concept bottleneck models (CBMs). First, we establish a concept trustworthiness benchmark to systematically evaluate current CBMs, based on an evaluation metric named *concept trustworthiness score*. Next, we propose a CBM framework that utilizes part-prototypes to make concept predictions from specific parts of the feature maps. Furthermore, we propose three modules (CLA module, CIA module, and PA module) into this CBM framework to improve the concept trustworthiness. The comprehensive experiments validate that our model achieves the state-of-the-art performance compared to previous CBMs, in both concept trustworthiness and accuracy. On the whole, our work conducts an in-depth analysis on the concept untrustworthiness problem in CBMs, towards reducing the potential risks of CBMs in the future.

Acknowledgements

This work is supported by the Science and Technology Project of SGCC: Research and Digital Application of High-precision Electric Power Super-scale Pre-trained Visual Model (5108-202218280A-2-395-XG).

References

Chattopadhyay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *WACV 2018*, 839–847. IEEE.

Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; and Dvijotham, K. 2023. Interactive Concept Bottleneck Models. In *AAAI 2023*, 5948–5955. AAAI Press.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *NIPS 2019*, 8928–8939.

Donnelly, J.; Barnett, A. J.; and Chen, C. 2022. Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In *CVPR 2022*, 10255–10265.

Furby, J.; Cunnington, D.; Braines, D.; and Preece, A. 2023. Towards a Deeper Understanding of Concept Bottleneck Models Through End-to-End Explanation. *arXiv preprint arXiv:2302.03578*.

Havasi, M.; Parbhoo, S.; and Doshi-Velez, F. 2022. Addressing Leakage in Concept Bottleneck Models. In *NIPS 2022*.

Heidemann, L.; Monnet, M.; and Roscher, K. 2023. Concept Correlation and Its Effects on Concept-Based Models. In *WACV 2023*, 4780–4788.

Hoffmann, A.; Fanconi, C.; Rade, R.; and Kohler, J. 2021. This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*.

Huang, Q.; Xue, M.; Huang, W.; Zhang, H.; Song, J.; Jing, Y.; and Song, M. 2023. Evaluation and Improvement of Interpretability for Self-Explainable Part-Prototype Networks. In *ICCV 2023*, 2011–2020.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F. B.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, 2673–2682. PMLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept Bottleneck Models. In *ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, 5338–5348. PMLR.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. E. 2019. Similarity of Neural Network Representations Revisited. In *ICML 2019*, 3519–3529.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; and Weller, A. 2021. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*.

Nauta, M.; van Bree, R.; and Seifert, C. 2021. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *CVPR 2021*, 14933–14943.

Oikarinen, T. P.; Das, S.; Nguyen, L. M.; and Weng, T. 2023. Label-free Concept Bottleneck Models. In *ICLR 2023*. OpenReview.net.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rymarczyk, D.; Struski, L.; Górszczak, M.; Lewandowska, K.; Tabor, J.; and Zielinski, B. 2022. Interpretable Image Classification with Differentiable Prototypes Assignment. In *ECCV 2022*, 351–368.

Rymarczyk, D.; Struski, L.; Tabor, J.; and Zielinski, B. 2021. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In *SIGKDD 2021*, 1420–1430.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV 2017*, 618–626.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, J.; Liu, H.; Wang, X.; and Jing, L. 2021. Interpretable Image Recognition by Constructing Transparent Embedding Space. In *ICCV 2021*, 875–884.

Xue, M.; Huang, Q.; Zhang, H.; Cheng, L.; Song, J.; Wu, M.; and Song, M. 2022. ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition. *arXiv preprint arXiv:2208.10431*.

Yüksekgönül, M.; Wang, M.; and Zou, J. 2023. Post-hoc Concept Bottleneck Models. In *ICLR 2023*. OpenReview.net.