"Explain me how you do it"
Concept Bottleneck Model Distillation by Ensemble Teacher

Pedro Fernandes Salgado

Dr. Michael Wollowski Dr. Reinhold von Schwerin

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

THU Technische Hochschule Ulm University of Applied Sciences

Concept Bottleneck Models

Concept bottleneck models differ from end-to-end models by the inclusion of an intermediary interpretable bottleneck layer. This limits the information that can be used by the model to interpretable information. The following are its main benefits:

- inherent interpretability
- predictability
- local and global intervenability

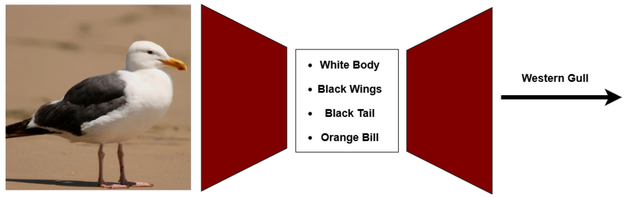

Figure 1. Concept Bottleneck Model Architecture. A concept encoder outputs a set of concepts, which are then used by the label decoder to predict the labels. The bottleneck layer is limited to a finite set of well-defined concepts in which each layer unit predicts if the corresponding concept appears in the sample.

However, concept labels are needed to train such models, and manual labeling is expensive. Even when concept labels are present they are often insufficient, leading to poor model accuracy (e.g. see Figure 2). Previous automatic concept extraction methods also produce an insufficient set of concepts.

Figure 2. Example of an indistinguishable cluster from the Caltech-UCSD Birds-200-2011 (CUB) dataset. From left to right the classes are groove billed ani, brewer blackbird, and shiny cowbird. Some shared concepts are black wing color and black bill color. On the other hand, there are no distinguishing concepts in the CUB dataset. This leads to them being confused by robust concept bottleneck models. However, a finetuned ResNet-50 model achieves an accuracy of 0.80. Therefore, there must exist distinguishing characteristics (e.g. the distinct bill of the ani). The goal is to find new interpretable concepts that are not shared by birds of a cluster.

Research Statement

In this work, an iterative unsupervised concept bottleneck model distillation method is proposed. I hypothesize that this method is able to find previously undiscovered concepts, improving the accuracy of distilled concept bottleneck models in comparison to previous methods.

Solution Architecture

Label Clustering

The goal is to find clusters, such that samples of classes in a cluster cannot be distinguished robustly by their concepts. If any concept appears frequently in one class but is not plausible for another, the classes must not be in the same cluster and can easily be distinguished. See Figure 2 for an example cluster.

1. Ensemble Training

Each model of the ensemble is trained on a subset of the clusters, so that they may find distinguishing details. In addition, random samples from classes outside the clusters are included in the subtasks to optimize for finding generalizable and widely-applicable concepts.

2. Segment Extraction

To extract informative segments, GradCAM is used. GradCAM scores subsections of the feature space based on gradient descent (i.e. the importance of the segment to the model output). A binary threshold is applied to this scoring to gain the most important sections (see Figure 3).

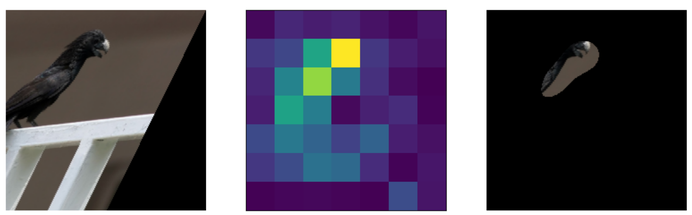

Figure 3. Example Segmentation. As seen, the model focuses strongly on the bill of the ani, which is one of its characteristic features.

3. Concept Discovery

Similarly to previous methods, the extracted segments are clustered into concepts by similarity. The similarity is measured by the L2 distance of segments in the embedding space of a model. Concepts can be scored by different metrics, e.g. number of appearances, robustness against noise, number of classes is distinguished (i.e how are the clusters changed by it). High-ranking concepts can be added to the set of known concepts.

Concept Bottleneck Model Distillation

The previous steps are executed iteratively until sufficient concepts have been found. Significantly, attention-guided training is possible since the location of each concept is known.

Further

Attention Model

Using GradCAM for segmentation is suboptimal for two main reasons:

- Only one segment can be extracted per model per sample
- Post-hoc explainability methods are known to be unfaithful

For this reason, an interpretable attention model is being developed. A segmentation model dynamically applies a fixed amount of masks to the input data. This resulting masked data is then fed into a classifier (see Figure 4). This resolves both problems faced by GradCAM. While similar to concept bottleneck models, this architecture does not guarantee that each filter represents a distinct concept, nor is it intervenable.

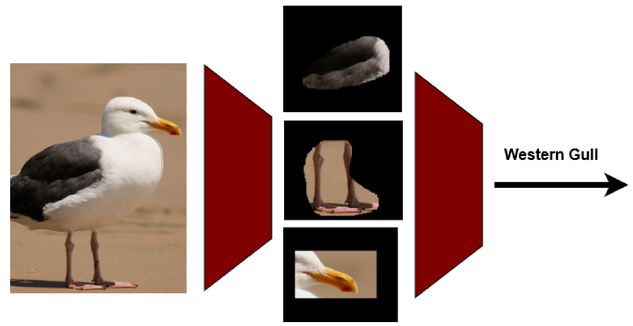

Figure 4. Architecture of Proposed Attention Model. The goal is to produce a faithful method for extracting task-relevant segments. In addition to minimizing classification loss, the masks can be trained to follow certain desiderata (e.g. as few masks as possible should be used, masks should be as small as possible and should not overlap, and they should be dissimilar to existing concept segments).

Variational Auto-Encoder

For segment clustering, a model with a well-structured latent space is required, otherwise distances in the latent space do not reflect conceptual similarity, impairing concept finding. The most promising solution appears to be variational auto-encoders. Variational auto-encoders infer latent variable distributions from observed variable. Crucially, they have a generative latent space as opposed to auto-encoders and can be used to encode samples unlike GAN models.