LABEL-FREE CONCEPT BOTTLENECK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept bottleneck model (CBM) are a popular way of creating more interpretable neural network by having hidden layer neurons correspond to human-understandable concepts. However, existing CBMs and their variants have two crucial limitations: first, the need to collect labeled data for each of the predefined concepts, which is time consuming and labor intensive; second, the accuracy of a CBM is often significantly lower than that of a standard neural network, especially on more complex datasets. This poor performance creates a barrier for adoption in practical real world applications. Motivated by these challenges, we propose *Label-free* CBM which is a framework to transform any neural network into an interpretable CBM without labeled concept data, while retaining a high accuracy. Our Label-free CBM has many advantages, it is: *scalable* - we present the first CBM scaled to ImageNet, *efficient* - creating a CBM takes only a few hours even for very large datasets, and *automated* - training it for a new dataset requires minimal human effort.

1 INTRODUCTION

Deep neural networks (DNNs) have demonstrated unprecedented success in a wide range of machine learning tasks such as computer vision, natural language processing, and speech recognition. However, due to their complex and deep structures, they are often regarded as *black-box* models that are difficult to understand and interpret their behaviors. Interpretable models are important for many reasons such as creating calibrated trust in our models, i.e. understanding when we should trust the models, and making deep learning models interpretable and explainable is an active yet challenging research topic.

One approach to make deep learning more interpretable is Concept Bottleneck Models (CBMs) (Koh et al., 2020). CBMs typically have a concept bottleneck layer before the (last) fully connected layer of the neural network. The concept bottleneck layer is trained to have each neuron correspond to a single human understandable *concept*. This makes the final decision a linear function of interpretable concepts, greatly increasing our understanding of the decision making. Importantly, CBMs

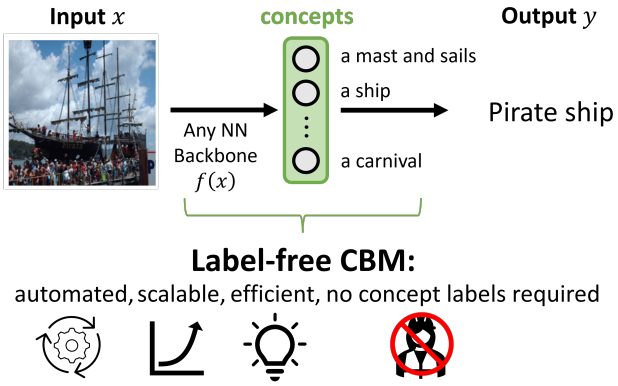


Figure 1: Our proposed **Label-free CBM** has many desired features of where existing CBMs lack and can transform any neural network backbone into an interpretable concept bottleneck model.

have been shown to be useful in a variety of applications, including model debugging and timely intervention on the DNNs. However, there are two crucial limitations of existing CBMs and their variants (Koh et al., 2020; Yuksekgonul et al., 2022; Zhou et al., 2018): (i) the labeled data is required for each of the predefined concept, which is time consuming and expensive to collect; (ii) the accuracy of a CBM is often significantly lower than the accuracy of a standard neural network, especially on more complex datasets.

To address the above two main challenges, we propose a new framework named **Label-free** CBM, which is capable of transforming any neural network into an interpretable CBM *without* labeled concept data while *preserving* comparable accuracy of the original neural network model by leveraging foundation models (Bommasani et al., 2021). Our Label-free CBM has many advantages:

- it is *scalable* - to our best knowledge, it is the first CBM that scales to ImageNet
- it is *efficient* - creating a CBM takes only a few hours even for very large datasets
- it is *automated* - training it for a new dataset requires minimal human effort

2 Related work

Post-hoc explanations (Samek et al., 2021). The approach of post-hoc explanations includes some classic methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), which try to explain individual model decisions by identifying which parts of the input data (e.g. pixels) are the most important for a given decision. However, these methods are based on local approximations of the DNN model and as such are not always accurate. Further, the explanations at the granularity of input pixels may not always be helpful and could require substantial subjective analysis from human. In contrast, our explanations in Section 4.4 are correct(not approximated) and explain predictions in terms of human-understandable concepts.

More interpretable final layer: (Wong et al., 2021) proposes making the FC layer sparse, and develop an efficient algorithm for doing so. They show that sparse models are more interpretable in many ways, but it still suffers from the fact the previous layer features are not interpretable. NBDT Wan et al. (2020) propose replacing the final layer with a neural backed decision tree for another form of more interpretable decisions.

CBM: Most related to our approach are Concept Bottleneck Models (Koh et al., 2020; Losch et al., 2019) which means creating a layer before the last fully connected layer where each neuron corresponds to a human interpretable concept. CBMs have been shown to be beneficial by allowing for human test-time intervention for improved accuracy, as well as being easier to debug. To reduce the training cost of a CBM in (Koh et al., 2020), a recent work (Yuksekgonul et al., 2022) proposed a post-hoc CBM that only need to train the last FC layer along with an optional residual fitting layer without the need to train the backbone from scratch by leveraging Test Concept Activation Vectors (TCAV) (Kim et al., 2018) or the multi-modal CLIP model (Radford et al., 2021). However, the post-hoc CBM does not fully address the problems of the original CBM as using TCAV still requires collecting annotated/labelled concept data and their use of CLIP model can only be applied to the NN backbone having the same architecture of CLIP image encoder which limits the application of post-hoc CBM. Besides, the performance of post-hoc CBMs without residual fitting layers is often significantly lower than the standard DNNs; although adding the residual layer can help to improve the accuracy, the residual layer is not interpretable and is thus undesired. Similarly, an earlier work Interpretable Basis Decomposition Zhou et al. (2018) proposes learning a concept bottleneck layer based on labeled concept data for explanable decisions, even though they do not call themselves a CBM. Extensive comparison of our method and existing approaches is shown in Table 1.

Model editing/debugging: Our approach is related to a range of works proposing ways to edit networks, such as (Bau et al., 2020; Wang et al., 2022) for generative vision models, (Bau et al., 2020) for classifiers, or language models (Meng et al., 2022; Mitchell et al., 2021). In addition (Abid et al., 2021) propose a way to debug model mistakes using TCAV activation vectors.

CLIP-Dissect: (Oikarinen & Weng, 2022). CLIP-Dissect is a recent method for understanding the roles of hidden layer neurons by leraging the CLIP multimodal model (Radford et al., 2021). It can provide a score of how close any neuron is to representing any given concept, which makes it useful as an optimization target for learning our interpretable projection in Step 3.

	Flexibility		Interpretability		Performance	
Method:	Without labeled concept data	Any network architecture	Sparse final layer	All interpretable features	Preserves accuracy	Extends to ImageNet scale
CBM	No	**Yes**	No	**Yes**	No	No
IBD	No	**Yes**	No	No	**Yes**	No
P-CBM	No	**Yes**	Yes	**Yes**	No	No
P-CBM (CLIP)	**Yes**	No	**Yes**	**Yes**	No	No*
P-CBM-h	No	**Yes**	**Yes**	No	**Yes**	No
P-CBM-h (CLIP)	**Yes**	No	**Yes**	No	**Yes**	No*
Label-free CBM (This work)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 1: Comparison of our method against existing methods for creating Concept Bottleneck models, CBM (Koh et al., 2020), IBD (Zhou et al., 2018) and 4 versions of P-CBM (Yuksekgonul et al., 2022), where '-h' indicates the hybrid model that uses uninterpretable residual term, '(CLIP)' means models using CLIP concepts and '*' indicates that their method might be able to scale if using our proposed GPT-3 procedure for generating concept set.

3 Label-free CBM: A new framework to build CBM

In this section, we propose **Label-free CBM**, a novel framework that builds a concept bottleneck model (CBM) in an automated, scalable and efficient fashion, which addresses the core limitations of existing CBMs (Koh et al., 2020; Yuksekgonul et al., 2022). Given a neural network backbone, **Label-free CBM** transforms the backbone into an interpretable CBM without the need of concept labels in the following 4 key steps as illustrated in Fig 2: (Step 1) Create the initial concept set and filter undesired concepts; (Step 2) Compute embeddings from the backbone and the concept matrix on the training dataset; (Step 3) Learn W_c to create a Concept Bottleneck Layer; (Step 4) Learn the weights W_F of the sparse final FC layer to make predictions.

Note that the backbone model here can either be a model trained on our target task, or some other task. In the following, we provide the details of **Label-free CBM** for Step 1 in Sec 3.1, Step 2 and 3 in Sec 3.2, and finally Step 4 in Sec 3.3 respectively.

3.1 Step 1: Concept set creation and filtering

In this step, we will describe how to create a concept set to serve as the basis of human-interpretable concepts in the concept bottleneck layer that will be constructed in the Step 3. This step consists of two sub-steps: **A. Initial concept set creation** and **B. Concept set filtering** below.

A. Initial concept set creation: To be specific, a concept set refers to the set of concepts represented in the Concept Bottleneck layer. In the original CBM paper (Koh et al., 2020), this is decided by

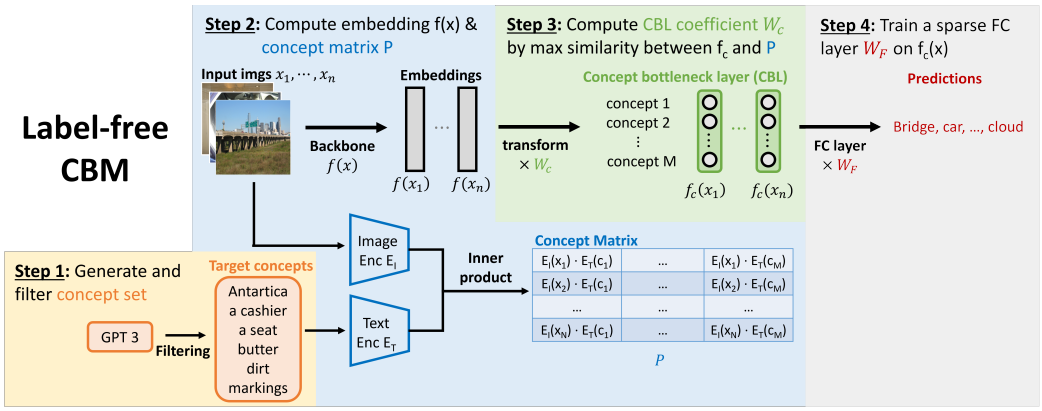


Figure 2: Overview of our pipeline for creating label-free CBM.

domain experts as the set of concepts that are important for the given task. However, since our objective is to automate the entire process of generating concept bottleneck models, it is not desirable to rely on having human experts. Instead, we propose generating the concept set via GPT-3 (Brown et al., 2020) using the OpenAI API. Somewhat surprisingly, GPT-3 has a good amount of domain knowledge of which concepts are important for detecting each class when prompted in the right way. Specifically, we ask the model the following:

- List the most important features for recognizing something as a {class}:
- List the things most commonly seen around a {class}
- Give superclasses for the word {class}:

Note that the {class} in here refers to the class name in the machine learning tasks. For GPT-3 to perform well on the above prompt, we provide two examples of the desired outputs for few-shot adaptation. Note that those two examples can be shared across all datasets, so no additional user input is needed for generating concept set for a new dataset. Full prompts and example outputs are illustrated in the Appendix Figures 7 and 8. To reduce variance we run each prompt twice and combine the results. This gives us a large, somewhat noisy set of initial concepts, which we improve by filtering in **B. Concept set filtering**. Note that Posthoc-CBM (Yuksekgonul et al., 2022) proposes using a knowledge graph known as ConceptNet (Speer & Havasi, 2013) to generate initial concept set, but it doesn't work for class names that consist of multiple words (i.e. almost all imagenet classes), and in general we found GPT-3 concepts to be of higher quality.

B. Concept set filtering: Several filters are employed to improve the quality and reduce the size of our concept set, as stated below:

- *Concept length*: We delete any concept longer than 30 characters in length, to keep concepts simple and avoid unnecessary complication.
- *Remove concepts too similar to classes*: We don't want our CBM to contain the output classes themselves as that would defeat the purpose of the explanation. To avoid this we remove all concepts that are too similar to the names of target classes. We measure this with cosine similarity in a text embedding space. In particular we use an ensemble of similarities of the CLIP ViT-B/16 text encoder (Radford et al., 2021) (denoted as CLIP cos) as well as the all-mpnet-base-v2 sentence encoder, so our measure is a combination of visual and textual similarity. In practice we deleted all concepts with similarity > 0.875 to any target class.
- *Remove concepts too similar to each other*: Similar to above, we don't want duplicate or too similar concepts in the bottleneck layer. We use the same embedding space as above, and remove any concept that has another concept with > 0.9 cosine similarity already in concept set.
- *Remove concepts not present in training data*: To make sure our concept layer accurately presents its target concepts, we remove any concept that doesn't activate CLIP highly. Note that this cut-off is a dataset specific metric and we delete all concepts with average top-5 activation less than the cut-off.
- *Remove concepts we can't project accurately*: This step is actually performed after step 3 and is described in section 3.2.

3.2 Step 2 and 3: Learning the Concept Bottleneck Layer (CBL)

Once the concept set is obtained, the next step is to learn a projection from the backbone models feature space into a space where the interpretable concepts lie in. Here, we present a way of learning the projection weights W_c without any labeled concept data using CLIP-Dissect (Oikarinen & Weng, 2022). To start with, let us define the set of target concepts that the bottleneck layer is expected to represent as $\mathcal{C} = \{t_1, ..., t_M\}$, as well as the training dataset (e.g. images) $\mathcal{D} = \{x_1, ..., x_N\}$ of the original task. Then we calculate and save the CLIP concept activation matrix P where $P_{i,j} = E_I(x_i) \cdot E_T(t_j)$, where E_I and E_T are the CLIP image and text encoders respectively. W_c is initialized as a random $d \times N$ matrix where d is the dimensionality of backbone features $f(x)$. Also let us define $f_c(x) = W_c f(x)$. We use k to denote a neuron of interest in the projection layer, and its activation pattern is denoted as q_k where $q_k^\top = [f_{c,k}(x_1), \ldots, f_{c,k}(x_N)]^\top$.

To make the neurons in the CBL actually interpretable, we need to enforce the projected neurons to possess the target concepts, which can be formulated as optimizing W_c to maximize the CLIP-Dissect similarity between activation pattern and the target concept. To optimize this similarity, we have designed a new and fully differentiable similarity function $\text{sim}(q_i, t_i)$ that can be applied to CLIP-Dissect, called *cosine cubed*, that still achieves very good performance in explaining the neuron functionality. Our goal is to minimize the following objective L over W_c in Equation (1):

$$L(W_c) = \sum_{i=1}^{N} -\text{sim}(q_i, t_i) := \sum_{i=1}^{N} -\frac{\bar{q_i}^3 \cdot \bar{P_{:,i}}^3}{||\bar{q_i}^3||_2 ||\bar{P_{:,i}}^3||_2}. \tag{1}$$

Here $\bar{q}$ indicates vector q normalized to have mean 0 and standard deviation 1, and as can be seen the *cosine cubed* similarity $\text{sim}(q_i, t_i)$ is simply the cosine similarity between two activation vectors after both have been normalized and raised to third power element-wise. The third power is necessary to make the similarity more sensitive to highly activating inputs. As this is still a cosine similarity, it takes values between $[-1, 1]$. We optimize L using the Adam optimizer on training data, with early stopping when similarity on validation data starts to decrease. Finally to make sure our concepts are truthful, we drop all concepts j with $\text{sim}(q_j, t_j) < 0.45$ on validation data at the end of training W_c. This cutoff was selected manually as an indicator of a neuron being interpretable.

3.3 Step 4: Learning the sparse final layer

Now that the Concept Bottleneck Layer is learned, the next task is to learn the final predictor with the fully connected layer W_F. The goal is to keep W_F sparse, since sparse layers have been demonstrated to be much more interpretable, as shown in (Wong et al., 2021). Given that both the backbone $f(x)$ and concept projection W_c are fixed, this is a problem of learning a sparse linear model. This can be solved efficiently with the elastic net penalty:

$$\min_{W_F, W_{F0}} ||f_c(X)^\top W_F + W_{F0} - y||_2^2 + \lambda R_\alpha(W_F), \tag{2}$$

where $R_\alpha = (1-\alpha)\frac{1}{2}||W_F||_2^2 + \alpha||W_F||_1$ and W_{F0} denotes the bias of the FC layer. We optimize Equation (2) using the GLM-SAGA solver created by (Wong et al., 2021). For the sparse models, $\alpha = 0.99$ and λ are such that each model has 25 to 35 nonzero weights per output class. This level was found to still able to result in interpretable decisions while retaining good accuracy. Depending on the number of features/concepts in the previous layer this corresponds to $0.7-15\%$ of the weights of the model being nonzero.

4 Results

4.1 Evaluation setup

To evaluate our approach, we train Label-free CBMs on 5 datasets. These are CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), CUB (Wah et al., 2011), Places365 (Zhou et al., 2017) and ImageNet (Deng et al., 2009). This is a diverse set of tasks, where CIFAR-10/100 and ImageNet are general image classification datasets, CUB is a fine-grained bird-species classification dataset and Places365 is focused on scene recognition. Their sizes vary greatly, with CUB having 5900 training samples, CIFAR datasets 50,000 each and ImageNet and Places365 have 1-2 million training images. CUB contains annotations with 312 concepts for each datapoint, such as *has wing color:blue* or *has head pattern:spotted*. Neither the concept names nor concept annotations were used for the training to showcase automation, still our method was successful in discovering similar concepts and being competitive with methods that utilize the available concept information.

For the model backbone, for CIFAR and CUB we use the same backbones as (Yuksekgonul et al., 2022) for fair comaprison, so we use CLIP(RN50) image encoder as the backbone for CIFAR, and ResNet-18 trained on CUB from https://github.com/osmr/imgclsmob for CUB. For both ImageNet and Places365 we use ResNet-50 trained on ImageNet as the backbone network. For the number of concepts, it mostly depends on the number of output classes on the task. The number of concepts for our models are as follows: 211 for CUB, 128 for CIFAR-10, 824 for CIFAR-100, 2202 for Places-365 and 4505 for ImageNet. CUB has a smaller number of concepts because we only used the *List the most important features for recognizing something as a "class"* prompt due to the specificity of

		Dataset				
Model	Sparse final layer	CIFAR10	CIFAR100	CUB200	Places365	ImageNet
Standard	No	88.80%	70.10%	76.70%	48.56%	76.13%
Standard (sparse)	Yes	82.96%	58.34%	**75.96%**	38.46%	**74.35%**
P-CBM*	Yes	70.50%	43.20%	59.60%	N/A	N/A
P-CBM (CLIP)*	Yes	84.50%	56.00%	N/A	N/A	N/A
Label-free CBM (Ours)	Yes	**86.40%**	**65.26%**	74.59%	**43.71%**	71.98%

Table 2: Accuracy comparison, best performing sparse model bolded. We can see our method outperforms Posthoc-CBM and matches or outperforms a sparse standard model. *Indicates reported accuracy.

the dataset. All models are trained on a single Nvidia Tesla P100 GPU, and the full training run takes anywhere from few minutes to 20 hours depending on the dataset size. The majority of runtime is taken by step 2, where we save activations for both the backbone and CLIP over the entire training dataset. Fortunately, these results only need to calculated once and can be reused for training new CBMs on the same dataset. Once the activations have been saved, learning the model takes less than 4 hours on all datasets.

4.2 Accuracy

Table 2 shows the performance of **Label-free CBM** on all 5 datasets. We can see that our method can create a CBM with a sparse final layer on all datasets with little loss in accuracy, including a model with 72% top-1 accuracy on the challenging ImageNet dataset. Our **Label-free CBM** clearly outperform Post-hoc CBM on all the datasets. We did not include methods with non-interpretable components (i.e. P-CBM-h or IBD) for comparison as we want to focus on fully interpretable CBM models. Some rows for P-CBM are N/A as they do not provide results for those datasets, and it is unclear if their method would scale. Note that P-CBM uses the expert provided concept set for CUB200, yet we can outperform it using our fully GPT-3 derived set of features. The standard sparse models were finetuned by us to also have a sparse decision layer with 25-35 nonzero weights per class and have very similar accuracy as our CBM.

4.3 Explainable decision rules

Perhaps the biggest benefit of Concept bottleneck models is that their decisions can be explained as a simple linear combination of understandable features. To showcase this, in Figure 3 we visualize the prediction rules for the classes "Orange" and "Lemon" on ImageNet, and classes "Mountain" and "Mountain Snowy" on Places365. The learned decision rules largely align with our intuitions. For ImageNet, the concept "citrus" fruit is highly connected to both, while orange colors activate the Orange class and yellow colors and limes activate the lemon class. On Places, concepts like "a high elevation" activate both classes, while "Mountain Snowy" is also activated by snow and ice related concepts and "Mountain" is activated by concepts related to volcanoes like "lava" and "crater". Visualizations like this allows us to gain a *global* understanding of how our model behaves.

4.4 Explainable individual decisions

In addition to global understanding, Concept Bottleneck models allow us to understand the reasoning behind individual decisions. Since our decisions are linear functions of interpretable features they can be explained in a simple and accurate way. After learning the Concept Bottleneck Layer, we normalize the activations of each concept to have mean 0 and standard deviation 1 on the training data. Given normalized features, the contribution of feature j to output i on input x_k can be naturally calculated as $\textit{Contrib}(x_k, i, j) = W_{F[ij]} * f_c(x_k)_j$. Since our W_F is sparse, most contributions will be 0 and the important contributions can be easily visualized with a bar plot. Example

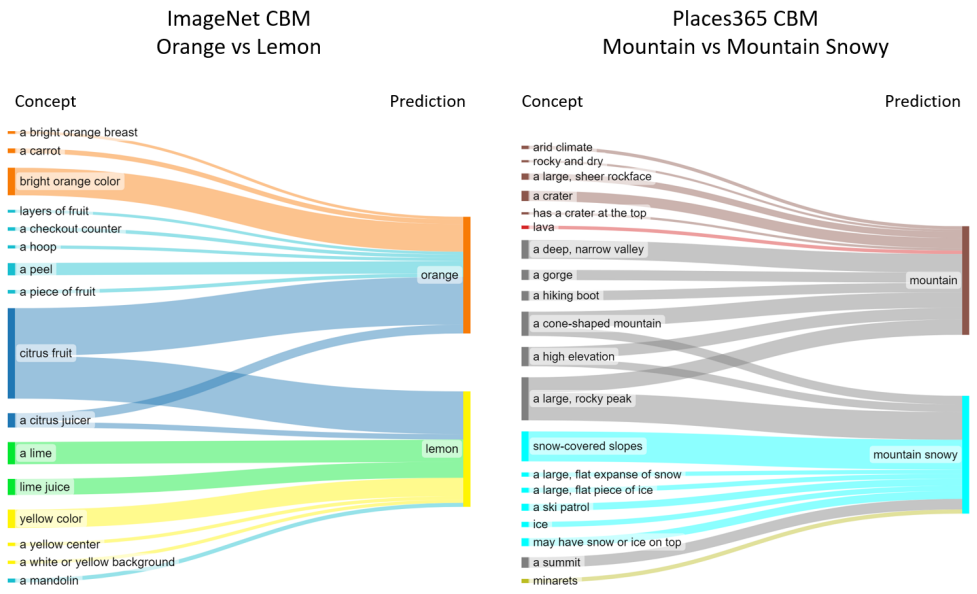
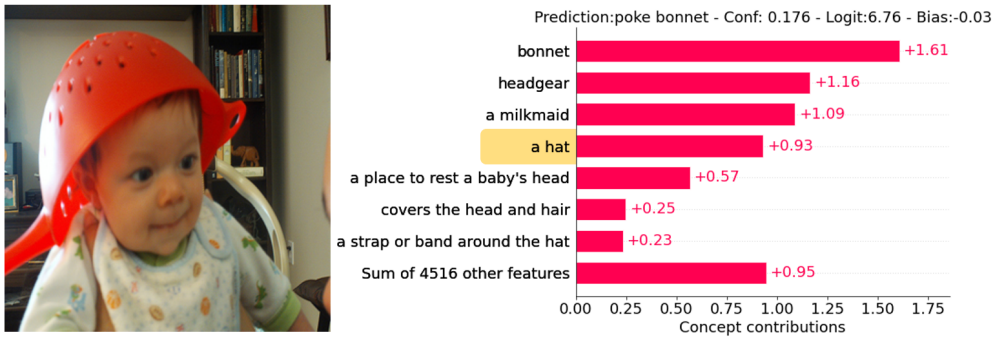


Figure 3: Visualization of the final layer weights of our **Label-free CBM**. Showcasing how our models differentiate between two similar classes. The width of the line is proportional to connecting weight.

Figure 4: Visualization of two correct decisions made by our **Label-free CBM**.

visualizations for explaining the models predicted class are shown in Figure 4. These visualizations show the concepts with highest absolute value contribution. Concepts with negative activation can still be important contributors, and they are shown as "NOT concept" in our visualizations. We can see for example the CUB model correctly identifies "a red head" as the most important feature in recognizing this image as "Red headed Woodpecker", while the fact that there is no "a rosy breast" present also increases models confidence.



Intervene: activation of “a hat” 2.70 -> 0. New prediction: strainer ✓

Figure 5: Here we fix an incorrect model prediction on ImageNet by simply zeroing out the incorrect activation for concept ”a hat”. Model originally predicted ”poke bonnet” which is a type of hat.

5 Case study: Manually improving an ImageNet model

In this section we take a deep dive into our ImageNet **Label-free CBM** – we analyze the errors it makes and design ways to debug/improve our network based on these findings. In particular, we show how we can inspect a handful of individual incorrect decisions and manually change our model weights to not only fix those predictions but also multiple other predictions, improving the networks overall accuracy. As far as we know, this is the first example of manually editing a large neural network with no obvious flaws in a way that improves test accuracy on the same distribution.

5.1 Types of model errors

During the course of our experiment, we were able to identify 4 different types of mistakes our model makes, all with different strategies for fixing them:

Type 1: Incorrect/ambiguous label: Despite looking like errors in terms of accuracy, these are simply a consequence of having noisy labels and are unavoidable. We will not attempt to fix this type of mistakes.

Type 2: No sufficient concept in CBL: Despite having a large set of concepts, our CBL does not include all concepts important for detecting certain classes. For example, detecting the difference between two species of snakes may require recognizing very fine-grained patterns that can't be easily explained in words. While these could often be fixed by adding new concepts to the concept set, we do not focus on it as it requires retraining the projection W_c and final layer.

Type 3: Incorrect concept activations: Some of the time, the activations in our CBL are incorrect and do not match the image, causing the prediction to become incorrect. Since the network up to the Concept Bottleneck layer is basically a black box, we can't easily improve the predictions, but predictions can be improved with test time interventions as seen in Figure 5.

Type 4: Incorrect final layer weight W_F: Sometimes even if all the concept activations are correct, the models final layer weights cause it to still make an erroneous prediction. As the final layer is fully interpretable, fixing these errors is the main focus of our study.

See Figure 10 in the Appendix for examples of Type 1 and Type 2 errors.

5.2 Editing final layer weights

In this section we describe our procedure for manually editing final layer weights of our model. This is done with the following procedure:

1. **Find an incorrect prediction belonging to the fourth class of errors**: This is done by visualizing incorrect model predictions and their explanations, and identifying cases which

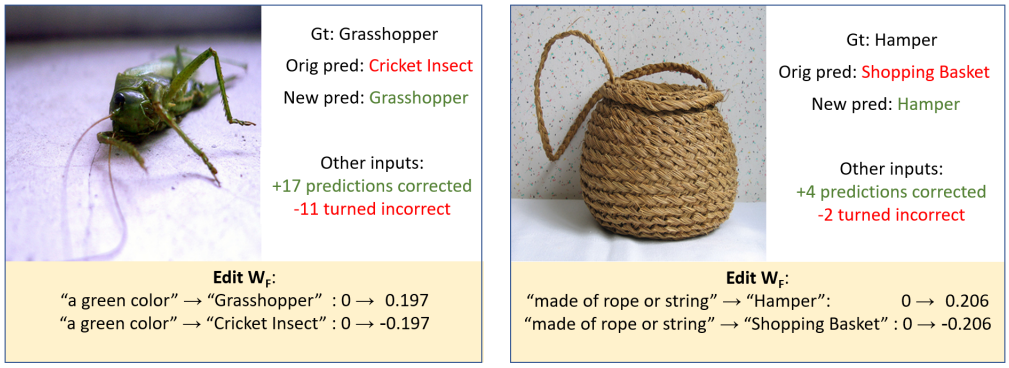


Figure 6: Examples of two edits we performed on our model predictions. Both edits fix the error we found, and fix more other predictions than they cause mistakes on other unseen inputs.

belong to the fourth error type, i.e. highly activating concepts looks correct and the label is correct and unambiguous.

2. **Identify a concept to edit.**: Perhaps the most tricky step, to flip a prediction we need to select a concept that is highly activated in this image, important for the ground truth class and not important for the incorrect class using our domain knowledge and/or information online and will have a small impact on other predictions. For example to flip a prediction from "cricket insect" to "grasshopper" we chose the concept "a green color".

3. **Change weights by sufficient magnitude**: Next we choose magnitude of weight change Δw, and edit the weights in the following way: $W_{F[gt,concept]} \leftarrow W_{F[gt,concept]} + \Delta w$ and $W_{F[pred,concept]} \leftarrow W_{F[pred,concept]} - \Delta w$. Typically we want Δw to be slightly more than required to flip the prediction on our example, but it understanding how it affect other predictions is hard and should be checked with a validation dataset.

Figure 6 shows two examples of model edits we performed, and their effects on model predictions on the validation set. In total we identified 5 such beneficial edits during our quick exploration, the other 3 are shown in 9. After applying all of them, the total validation accuracy of our model goes from 71.98% to 72.02%, correcting 38 predictions while turning 17 incorrect. While this might seem like a small difference, it is worth noting we only change 10 weights in the final layer out of 4.5 million total weights, of which 34,550 are non-zero. Our edits only affect 10 classes which have a total of 500 validation examples, so in effect we increase accuracy on this subset by 4.2% which is a significant boost. While labor-intensive, we believe this editing approach is very promising for future use, where practitioners could for example notice an incorrect prediction in a production system, quickly devise a fix for that prediction, and check how it affects other predictions in validation data.

6 Conclusion

We have presented **Label-free CBM**, a fully automated and scalable method for generating concept bottleneck models and used it to create the first high performing CBM on ImageNet. In addition, we have demonstrated how our models are easily understandable in terms of both global decision rules and individual predictions, and how to use this understanding to manually improve model weights in a fully trained ImageNet model in a way that improves validation accuracy.

7 Reproducibility

We have described our algorithm in detail, including hyperameters and other details in sections 3 and 4. In addition we will release our code publicly prior to publication.

REFERENCES

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully explaining model mistakes using conceptual counterfactuals. *arXiv preprint arXiv:2106.12723*, 2021.

David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *European conference on computer vision*, pp. 351–369. Springer, 2020.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks, 2019. URL https://arxiv.org/abs/1907.10882.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*, 2022.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2021. URL https://arxiv.org/abs/2110.11309.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Robyn Speer and Catherine Havasi. Conceptnet 5: A large semantic network for relational knowledge. *The people's web meets NLP, theory and applications of natural language processing*, pp. 161–176, 02 2013. doi: 10.1007/978-3-642-35085-6_6.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Henry Jin, Suzanne Petryk, Sarah Adel Bargal, and Joseph E. Gonzalez. Nbdt: Neural-backed decision trees, 2020.

Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Rewriting geometric rules of a gan. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, pp. 11205–11216. PMLR, 2021.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.

List the things most commonly seen around a "tench":

-a pond
-fish
-a net
-a rod
-a reel
-a hook
-bait

List the things most commonly seen around a "beer glass":

-beer
-a bar
-a coaster
-a napkin
-a straw
-a lime
-a person

List the things most commonly seen around a "mountain":

-a peak
-a valley
-a forest
-rocks
-dirt
-a trail
-an animal

Give superclasses for the word "tench":

-fish
-vertebrate
-animal

Give superclasses for the word "beer glass":

-glass
-container
-object

Give superclasses for the word "grasshopper":

-insect
-arthropod
-animal

Figure 7: Full prompts used for GPT-3 concept set creation. Text in green generated by GPT, rest is our prompt.

APPENDIX: ADDITIONAL FIGURES

Figures 7, 8 provide examples of the full prompts we used for GPT-3, as well as GPT outputs.

Figure 9 shows the additional model edits performed in our ImageNet CBM experiment, and Figure 10 showcases Type 1 and Type 2 errors made by our ImageNet CBM.

List the most important features for recognizing something as a "goldfish":

-bright orange color
-a small, round body
-a long, flowing tail
-a small mouth
-orange fins

List the most important features for recognizing something as a "beerglass":

-a tall, cylindrical shape
-clear or translucent color
-opening at the top
-a sturdy base
-a handle

List the most important features for recognizing something as a "scuba diver":

-a person wearing a scuba mask and breathing apparatus
-a wet suit
-flippers
-a diving tank

Figure 8: Full prompts used for GPT-3 concept set creation. Text in green generated by GPT, rest is our prompt.

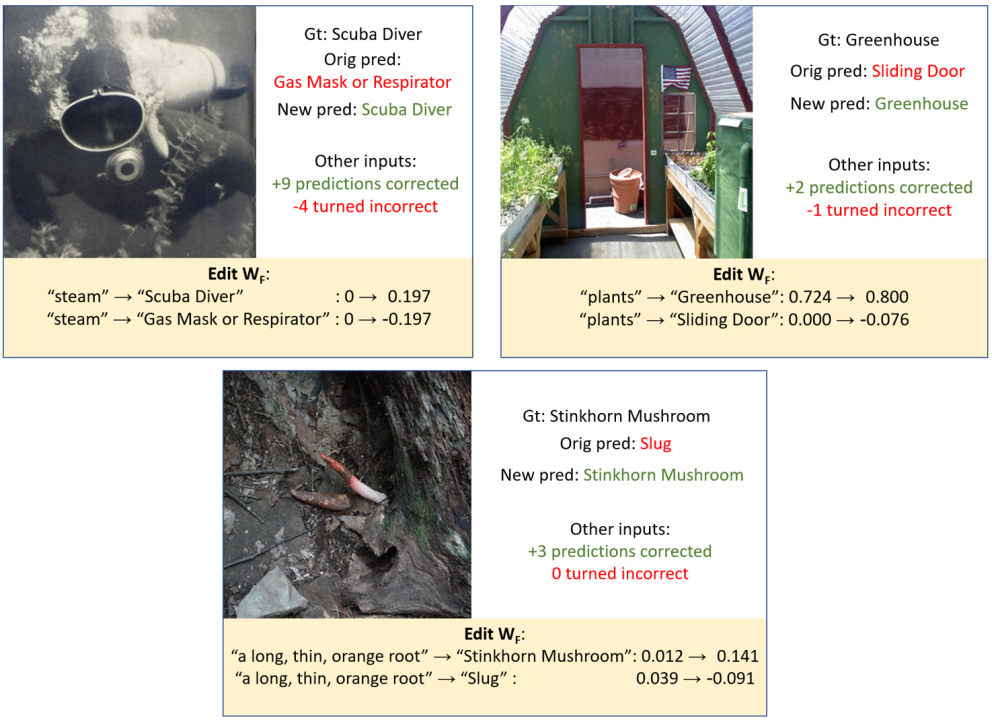


Figure 9: The other 3 model edits we performed in our experiment

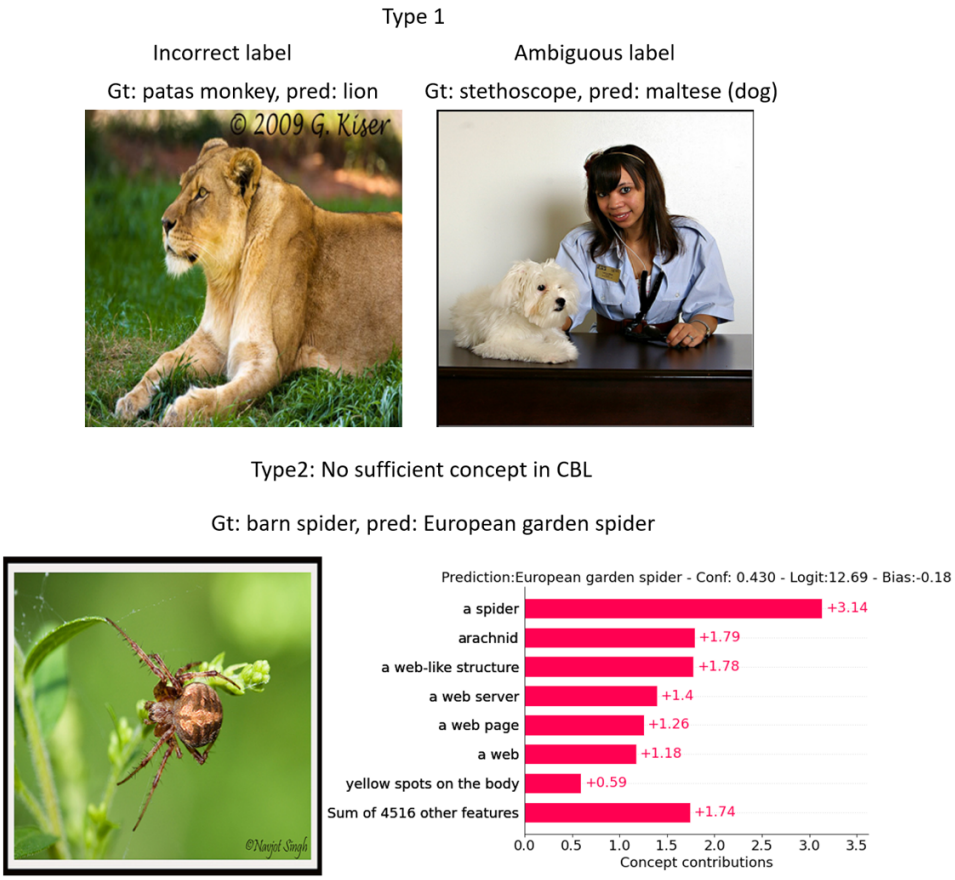


Figure 10: Examples of Type 1 and type 2 model errors.