Semi-supervised Concept Bottleneck Models

Jeeon Bae† Sungbin Shin‡ and Namhoon Lee‡

†: Kyung Hee University, ‡: Pohang University of Science and Technology

Summaries

- Concept bottleneck models (CBMs) enhance the interpretability of deep neural networks by adding a concept layer between the input and output layers.
- However, this improvement comes at the cost of labeling concepts, which can be prohibitively expensive.
- To tackle this issue, we develop a semi-supervised learning approach to CBMs that can make accurate predictions given only a handful of concept annotations.

Motivation: Concept Labeling Costs for CBMs

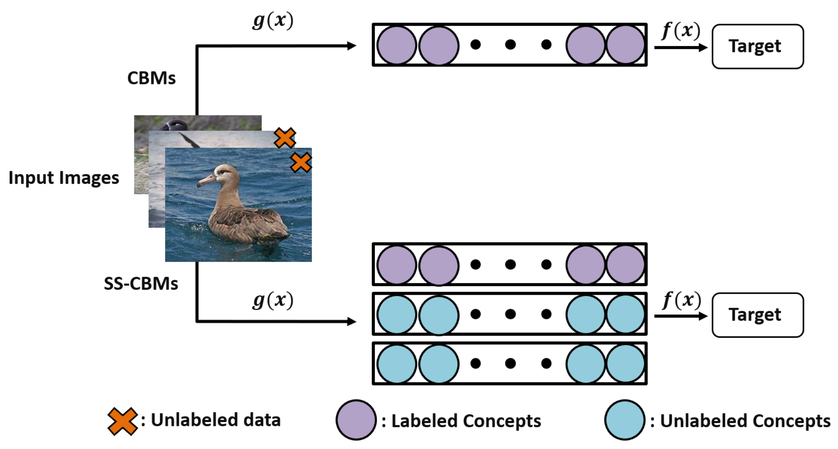

Fig. 1: Comparison between the standard CBMs and SS - CBMs

- CBMs [1] secure interpretability by having a layer of concepts in the network, by which the model first makes predictions for the high-level concepts and then predicts the target class based on the predicted concept values.
- However, CBMs[1] require ground-truth annotations for concepts to be trained. Compared to the standard end-to-end learning models, this labeling cost can be quite high, undermining their utility in practice.
- we suggest a new way to tackle this issue via semi-supervised learning (SSL), learning with only a small number of labeled examples while having access to potentially a large amount of unlabeled data.

SS-CBMs

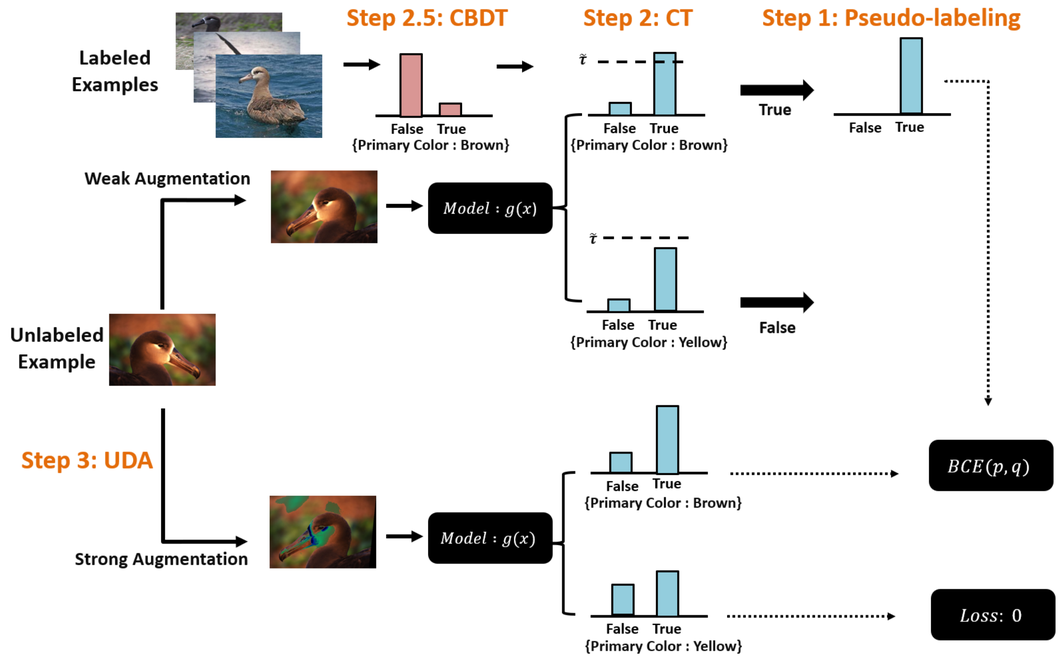

Fig. 2: Overview of SS - CBMs

- We propose a semi-supervised concept bottleneck model (SS-CBM), which leverages a limited set of labeled concepts while simultaneously utilizing a large amount of unlabeled concepts.
- We integrate a new method called Class Based Dynamic Thresholding (CBDT) with FixMatch[2], an existing state-of-the-art SSL approach.
- We consider the two labeling scenarios for SS-CBM: "easy case" and "hard case" depending on the availability of the class labels for the concept-unlabeled dataset.

CBMs with FixMatch

- We incorporated the Pseudo-labeling[3] and the Unsupervised Data Augmentation (UDA)[4] which were originally employed in FixMatch[2].
- Pseudo-labeling generates artificial labels by selecting the label with the highest probability from the model's predicted distribution for unlabeled data.
- Unsupervised data augmentation, a consistency regularization method in SSL, employs two types of augmentation (weak and strong) on unlabeled data. It calculates the consistency loss by comparing the predictions obtained from these two augmented inputs.

Class-based Dynamic thresholding

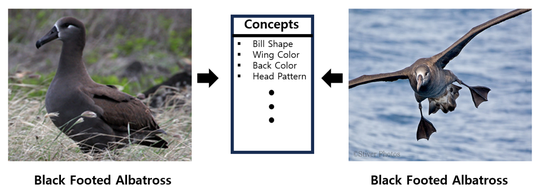

Fig. 3: Samples with the same class have similar conceptual values.

Confidence thresholding is a technique that uses a predefined threshold to determine which unlabeled data points to use to train a model. If the confidence thresholding of FixMatch is directly applied to CBM, there are two main issues:

- When using fixed values, it fails to regulate erroneous pseudo-labels as training progresses.
- It only considers the relationship between the input and the concept, thus it does not take into account the relationship between the concept and the target.

Thus, we applied the concept that samples from the same class should exhibit more similarity than those from different classes. Using this prior knowledge, we dynamically adjusted the threshold by analyzing the probability distribution of concepts for each class in the labeled samples.

In the hard case, however, the class labels for unlabeled datasets are not available. To apply our approach in this situation, we generate pseudo class labels from the auxiliary end-to-end neural network which predicts the class label from the inputs.

Evaluation of SS-CBM

Method	Target accuracy (%) $n=1$	$n=2$	$n=3$	$n=6$	$n=30$	Concept accuracy (%) $n=1$	$n=2$	$n=3$	$n=6$	$n=30$
CBM (No SSL)	17.47	28.27	36.50	51.07	76.70	87.09	89.07	90.35	92.83	96.60
CBM w/ FixMatch	15.48	26.48	39.25	56.83	–	87.60	89.18	91.05	93.80	–
SS-CBM (easy)	**53.33**	**65.46**	**68.32**	**71.40**	–	**94.03**	**95.50**	**95.80**	**96.30**	–
SS-CBM (hard)	**19.34**	**38.91**	**49.74**	**64.16**	–	**88.95**	**92.05**	**93.48**	**95.27**	–

Fig. 4: Target and concept accuracy of different models for the CUB

- We evaluate SS-CBM for varying labeling cost and present the results in Fig 4.
- SS-CBM consistently outperforms the other baselines across all settings.

Analysis of SS-CBM

Model	$n=1$	$n=2$	$n=3$	$n=6$
EfficientNet V2	30.00	44.41	51.81	63.76

Fig. 5: Target accuracy of the auxiliary network for the CUB

The knowledge about the relationship between concept and target can still be helpful even when the target accuracy of the auxiliary network is low. This improvement can be attributed to two key factors:

- **Indirect use:** SS-CBMs do not directly use class knowledge. Instead, they indirectly incorporate it into the threshold. This means that SS-CBMs can prevent the training of accurate pseudo-label in the worst-case scenario. It becomes evident that prioritizing the avoidance of inaccurate pseudo-labels is more crucial.
- **Similar class has similar features:** We speculate that even when predicting an incorrect pseudo class label, the model would tend to predict classes with features similar to the correct class label.

Conclusion

- We address the problem of high annotation cost in learning CBMs via a semi-supervised learning approach.
- To this end, we propose a novel CBMs called SS-CBMs, which combines existing semi-supervised learning methods with techniques to regulate inaccurate pseudo-labels by utilizing the essence of CBMs.
- Our approach not only encompasses various label scenarios but also offers valuable insights on effectively utilizing unlabeled data for CBMs.
- As a result, SS-CBMs are able to attain a target accuracy of 93% and a concept accuracy of 99.6% with a mere 20% subset of the full dataset.

References

[1]. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. ICML, 2020.
[2]. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H.,Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS, 2020.
[3]. Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. Workshop on challenges in representation learning, ICML, 2013.
[4]. Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. NeurIPS, 2020.

Contact: jeeon98b@khu.ac.kr, {ssbin4, namhoonlee}@postech.ac.kr

Please refer to the main paper for more details.