Adaptive Concept Bottleneck for Foundation Models

Jihye Choi [1] **Jayaram Raghuram** [1] **Yixuan Li** [1] **Suman Banerjee** [1] **Somesh Jha** [1]

Abstract

Advancements in foundation models have led to a paradigm shift in deep learning pipelines. The rich, expressive feature representations from these pre-trained, large-scale backbones are leveraged for multiple downstream tasks, usually via light-weight fine-tuning of a shallow fully-connected network following the representation. However, the non-interpretable, black-box nature of this prediction pipeline can be a challenge, especially in critical domains such as healthcare. In this paper, we explore the potential of Concept Bottleneck Models (CBMs) for transforming complex, non-interpretable foundation models into interpretable decision-making pipelines using high-level concept vectors. Specifically, we focus on the test-time deployment of such an interpretable CBM pipeline "in the wild", where the distribution of inputs often shifts from the original training distribution. We propose a *light-weight adaptive CBM* that makes dynamic adjustments to the concept-vector bank and prediction layer(s) based solely on unlabeled data from the target domain, without access to the source dataset. We evaluate this test-time CBM adaptation framework empirically on various distribution shifts and produce concept-based interpretations better aligned with the test inputs, while also providing a strong average test-accuracy improvement of 15.15%, making its performance on par with that of non-interpretable classification with foundation models.

1. Introduction

Foundation models, trained on vast data corpora, are powerful feature extractors applicable across diverse distributions and tasks (Bommasani et al., 2021; Rombach et al., 2022). They can be applied to classification tasks via zero-shot predictions or linear probing with task-specific training data (Kumar et al., 2022; Radford et al., 2021). However, such models often operate as inscrutable black-boxes, presenting a barrier to user trust and understanding. Another challenge faced in the standard deployment of foundation model-based deep classifiers is their vulnerability to distribution shifts at test time caused *e.g.*, due to environmental changes, which can cause a drop in performance. This is particularly challenging in high-stakes domains such as healthcare (AlBadawy et al., 2018) and autonomous driving (Yu et al., 2020).

This work aims to tackle these challenges and develop an *interpretable classification* framework that leverages the rich, expressive feature representations of foundation models, while also preserving their robustness to (operational) distribution shifts at test time. To address interpretability, we leverage Concept Bottleneck Models (CBMs) (Koh et al., 2020) to transform a foundation model-based deep classifier into an interpretable, concept-based prediction pipeline. Unlike their early versions where a direct mapping from an input to its concept predictions is learned using concept labels, recent advances have shown the potential to transform any pre-trained neural network into a CBM (Yuksekgonul et al., 2023), and vision-language models can guide the construction of concept bottlenecks without explicit concept labels (Oikarinen et al., 2023; Wu et al., 2023). Concept-based predictions not only provide interpretability but are also beneficial for robustness under varying input distributions. A central premise of CBMs is that as complex feature embeddings go through the concept bottleneck, the resulting prediction should, in theory, become more invariant to inconsequential changes in the input (Kim et al., 2018; Adebayo et al., 2020).

However, we identify that CBMs that are directly deployed under distribution shifts do not necessarily produce more robust predictions compared to that obtained directly based on the feature representations (*i.e.*, foundation models either with zero-shot prediction or with fine-tuned linear prediction). This observation underscores the need for a dynamic approach to adapt the concept- (or CBM-) based prediction framework for real-world deployment in the wild.

To our knowledge, we make the first attempt at *test-time adaptation of CBMs* with a foundation model as the back-

[1]Department of Computer Sciences, University of Wisconsin - Madison, WI, USA. Correspondence to: Jihye Choi <jihye@cs.wisc.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

bone. Given unlabeled test data and a frozen foundation model, we propose to: **1)** adapt the concept bottleneck to align the target (test) input's concept-score patterns with that of the source domain, and **2)** adapt the label predictor (following the concept bottleneck) to dynamically adjust the contribution of different concepts to the prediction. Our empirical results show that the proposed method significantly enhances the accuracy of CBMs under various distribution shifts (*e.g.*, improving the accuracy of Yuksekgonul et al. (2023) by 28%), while also providing meaningful concept-based interpretations.

2. Background

Let $\mathcal{X}$ denote the space of inputs $\mathbf{x}$ and $\mathcal{Y} := \{1, \cdots, L\}$ the set of class labels y. We assume that the labeled training data from a source domain is sampled from an unknown probability distribution $p_{\text{s}}(\mathbf{x}, y)$, and unlabeled test data from a target domain is sampled from an unknown probability distribution $p_{\text{t}}(\mathbf{x})$. The subscripts 's' and 't' refer to the source and target domain respectively.

2.1. Foundation Models with a Concept Bottleneck

Consider a foundation model $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathbb{R}^d$, which is any pre-trained backbone model or feature extractor (Eslami et al., 2023; Jia et al., 2021; Girdhar et al., 2023) that maps the input $\mathbf{x}$ to an intermediate feature embedding $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^d$. $\boldsymbol{\phi}(\mathbf{x})$ is pre-trained on a large-scale, broad mixture of data for general purposes, *i.e.*, not restricted to a specific domain. For a specific downstream classification task, the general practice is to either apply zero-shot prediction on $\boldsymbol{\phi}(\mathbf{x})$, or to train a shallow label predictor $\mathbf{g}_s : \mathbb{R}^d \mapsto \mathbb{R}^L$ using a supervised loss $\ell(\mathbf{x}, y)$ (*e.g.*, cross-entropy) that maps $\boldsymbol{\phi}(\mathbf{x})$ to the un-normalized class predictions (logits) $\mathbf{g}_s(\boldsymbol{\phi}(\mathbf{x}))$.

A CBM (Koh et al., 2020) first projects the high-dimensional feature embeddings to a lower m-dimensional ($m \ll d$) *concept-score space* (acting like a bottleneck), and follows it with a *label predictor*, which is a simple linear or fully-connected layer that maps the concept scores into class predictions. The concept bottleneck is represented by a matrix of m unit *concept vectors* $\mathbf{C}_s = [\mathbf{c}_{s1} \,/\, \|\mathbf{c}_{s1}\|_2 \;\cdots\; \mathbf{c}_{sm} \,/\, \|\mathbf{c}_{sm}\|_2]^\top \in \mathbb{R}^{m \times d}$, where each $\mathbf{c}_{si} \in \mathbb{R}^d$ represents a high-level concept (*e.g.*, "stripes", "dots"). The m concept scores are obtained via a linear projection $\mathbf{v}_{\mathbf{C}_s}(\mathbf{x}) = \mathbf{C}_s\, \boldsymbol{\phi}(\mathbf{x})$, which is followed by the label predictor to obtain the CBM model, defined as $\mathbf{f}_s^{(\text{cbm})}(\mathbf{x}) := \mathbf{W}_s \mathbf{v}_{\mathbf{C}_s}(\mathbf{x}) + \mathbf{b}_s = \mathbf{W}_s \mathbf{C}_s\, \boldsymbol{\phi}(\mathbf{x}) + \mathbf{b}_s$. The label predictor is defined by $\mathbf{W}_s \in \mathbb{R}^{L \times m}$ and $\mathbf{b}_s \in \mathbb{R}^L$, and it outputs the un-normalized class predictions.

2.2. Distribution Shifts in the Wild.

Our focus in this work is on the performance of CBMs deployed in the wild, *i.e.*, when test inputs can undergo a distribution shift relative to the training data. This is a very practical scenario in real-world deployments, where inputs may undergo covariate shifts *e.g.*, due to noise, blur, snow, fog, lighting changes, etc (known as common corruptions) (Hendrycks & Dietterich, 2019b; Hendrycks et al., 2020). The distribution shift could also take the form of *disparate correlation to semantics*, *e.g.*, waterbirds always on a water background in the source dataset, but on a land background in the target dataset (Sagawa et al., 2019a).

There has been growing interest in the utility of concept-based explanations under distribution shifts. Since the first work (Kim et al., 2018) hinted at the potential of high-level concepts as "diagnosis units" against low-level perturbations (*e.g.*, adversarial examples), subsequent research has suggested the utility of concept-based explanations for the diagnosis and analysis of data drifts (Adebayo et al., 2020; Abid et al., 2022; Moayeri et al., 2023). However, prior works do *not* consider whether their *static concepts*, prepared before deployment, would be appropriate for the target data at test time, and do not provide a way of adapting the concepts according to the target data. Our work emphasizes the need for a dynamic approach to CBMs under distribution shift in order to maintain their accuracy and provide reliable explanations.

3. Adaptive Concept Bottleneck Models

We propose a dynamic approach to adapt CBMs at test time. Given a CBM $\mathbf{f}_s^{(\text{cbm})}(\mathbf{x})$ that is represented by a concept bank $\mathbf{C}_s$ and trained on a source dataset (that is not accessible), and an unlabeled test set $\mathcal{D}_t = \{\mathbf{x}_{tn}\}_{n=1}^{N_t}$ from a target distribution, our two-fold objective to adapt the CBM is [1]:

1. **Concept-Score Alignment (CSA):** Perform feature alignment of the concept scores of test inputs $\mathbf{v}_{\mathbf{C}}(\mathbf{x}_t) \in \mathbb{R}^m$ such that their class-conditional distributions are close to that of the concept scores from the source dataset. By adapting the concept vectors $\mathbf{C}$, this will ensure that the label predictor continues to "see" very similar class-conditional input distributions at test time, thereby maintaining accurate predictions.
2. **Linear Probing Adaptation (LPA):** Adapt the label predictor $(\mathbf{W}, \mathbf{b})$ of the CBM to account for any discrepancy or mismatch in the feature-alignment CSA step.

We next present our adaptation objectives for CSA and LPA. Following the convention in the TTA literature (Wang et al., 2021; Chen et al., 2022), we randomly split the test set into

[1] Here we drop the subscript 's' to denote that they are adaptation parameters, not specific to the source domain.

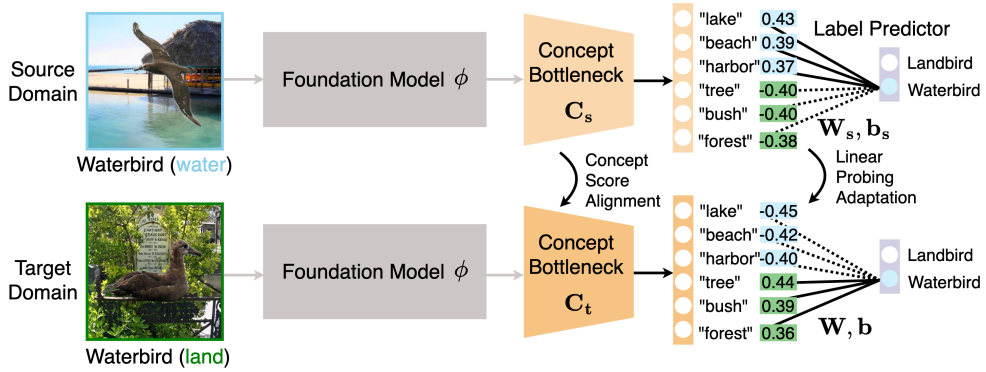

Figure 1: **Proposed test-time adaptation for a concept bottleneck model based on a (frozen) foundation model backbone.** As shown in the example inputs, there is a distribution shift in the target domain in the form of disparate correlation to semantics. In the target domain, the waterbird images have a land background, whereas the images have a water background in the source domain. Our proposed method consists of a *concept-score alignment* step which adapts the concept bottleneck layer, and a *linear probing adaptation* step which adapts the label predictor (classifier) layer.

fixed-size batches $\mathcal{D}_t = \bigcup_{b=1}^{B} \mathcal{D}_t^b$, and perform adaptation sequentially on each batch b, obtaining the adapted model's predictions on the subsequent (unseen) batch $b+1$. In the following, we refer to a specific test batch $\mathcal{D}_t^b$.

Pseudo-labeling. Since the test samples are unlabeled, it becomes challenging to perform class-conditional alignment of the concept scores $\mathbf{v}_{\mathbf{C}}(\mathbf{x}_t)$ (*i.e.*, CSA). We utilize the idea of pseudo-labeling to address this, as commonly done in the TTA and semi-supervised learning literature (Chen et al., 2022; Lee et al., 2013; Sohn et al., 2020). A simple approach for pseudo-labeling the test set is to use the class predictions of the source-domain CBM (referred to as "self-labeling"). However, the CBM is often not robust to distribution shifts and could lead to poor-quality pseudo-labels. Instead, we leverage the powerful feature-extraction backbone $\phi(\mathbf{x})$, which is a foundation model pre-trained on diverse data distributions, and use its pseudo-label $\hat{y}_t$ via *zero-shot prediction* (as done *e.g.*, in Radford et al. (2021)).

Objective for CSA. The general form of the adaptation objective (to minimize) for concept-score alignment is:

$$L_{\text{CSA}}(\mathbf{C}) = \frac{1}{|\mathcal{D}_t^b|} \sum_{\mathbf{x}_t \in \mathcal{D}_t^b} \ell_{ada}(\mathbf{f}_t^{(\text{cbm})}(\mathbf{x}_t), \hat{y}_t) + \lambda_{\text{CSA}} \|\mathbf{C} - \mathbf{C}_s\|_F^2, \quad (1)$$

where ℓ_{ada} is an adaptation loss (to be defined) guided by the pseudo-label, and the second term is a regularization on how much the concept vectors can to deviate from their source domain values in terms of the Frobenius norm.

Motivated by class-aware feature alignment (Jung et al., 2023), we design $\ell_{ada}(\cdot)$ to adapt the concept bottleneck to achieve concept-score alignment on a per-class level. Jung et al. (2023) model the intermediate feature representation of a DNN classifier as a class-conditional multivariate Gaussian, whose parameters are estimated from the source-domain dataset. We also base our design of $\ell_{ada}(\cdot)$ on the well-explored idea that for discriminative feature alignment, the *intra-class distances (compactness)* should be small, and the *inter-class distances (separation)* should be large on the test samples (Ye et al., 2021; Ming et al., 2023).

Suppose the class-conditional distributions of the concept-score vector in the source domain are modeled as Gaussians: $\mathbb{P}(\mathbf{v}_{\mathbf{C}_s}(\mathbf{x}_s) \,|\, y_s = y) = \mathcal{N}(\mathbf{v}_{\mathbf{C}_s}(\mathbf{x}_s); \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y),\ \forall y \in \mathcal{Y}$. Given a labeled source-domain dataset, it is straight-forward to estimate $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_y$ using the sample mean and sample covariance of $\mathbf{v}_{\mathbf{C}_s}(\mathbf{x}_s)$ on the samples from class y. Although we cannot access the source dataset during adaptation, we assume (as in Jung et al. (2023)) that we have access to these distribution statistics $\{(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)\}_{y \in \mathcal{Y}}$.

The Mahalanobis distance measures the distance of a test input's concept-score vector to a class-conditional Gaussian as $D_{\text{mah}}(\mathbf{x}_t; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) = (\mathbf{v}_{\mathbf{C}}(\mathbf{x}_t) - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{v}_{\mathbf{C}}(\mathbf{x}_t)) - \boldsymbol{\mu}_y)$. For a test input $\mathbf{x}_t$ with pseudo-label $\hat{y}_t$, the intra-class and inter-class distances are defined as

$$D_{\text{intra}}(\mathbf{x}_t, \hat{y}_t) = D_{\text{mah}}(\mathbf{x}_t; \boldsymbol{\mu}_{\hat{y}_t}, \boldsymbol{\Sigma}_{\hat{y}_t}) \quad \text{and} \quad (2)$$

$$D_{\text{inter}}(\mathbf{x}_t, \hat{y}_t) = \sum_{\ell=1:\ell \neq \hat{y}_t}^{L} D_{\text{mah}}(\mathbf{x}_t; \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell). \quad (3)$$

Finally, the adaptation loss in Eqn (1) based on the Mahalanobis distance, which aims to minimize the intra-class distances and maximize the inter-class distances is

$$\ell_{ada}(\mathbf{f}_t^{(\text{cbm})}(\mathbf{x}_t), \hat{y}_t) = \log \frac{D_{\text{intra}}(\mathbf{x}_t, \hat{y}_t)}{D_{\text{inter}}(\mathbf{x}_t, \hat{y}_t)}. \quad (4)$$

Objective for LPA. We achieve linear probing adaptation by adjusting the label predictor of the CBM using cross-entropy loss between the predictions of the target-domain

Dataset			ZS	LP	Yuksekgonul et al. (2023)				Yeh et al. (2020)			
					w/o adaptation	+ CSA	+ LPA	+ CSA + LPA	w/o adaptation	+ CSA	+ LPA	+ CSA + LPA
Waterbirds	Source	AVG	0.821	0.973	0.978 ± 0.001	-	-	-	0.988 ± 0.001	-	-	-
		WG	0.662	0.949	0.964 ± 0.003	-	-	-	0.983 ± 0.001	-	-	-
	Target	AVG	0.613	0.538	0.333 ± 0.004	0.388 ± 0.014	0.589 ± 0.003	0.613 ± 0.003	0.440 ± 0.003	0.444 ± 0.004	0.595 ± 0.004	0.634 ± 0.006
		WG	0.419	0.447	0.299 ± 0.011	0.153 ± 0.030	0.337 ± 0.005	0.419 ± 0.001	0.370 ± 0.009	0.381 ± 0.010	0.336 ± 0.007	0.389 ± 0.012
Metashift	Source	AVG	0.957	0.972	0.979 ± 0.001	-	-	-	0.972 ± 0.001	-	-	-
		WG	0.934	0.960	0.969 ± 0.003	-	-	-	0.960 ± 0.001	-	-	-
	Target	AVG	0.947	0.783	0.844 ± 0.014	0.844 ± 0.015	0.944 ± 0.001	0.948 ± 0.001	0.901 ± 0.001	0.900 ± 0.002	0.917 ± 0.002	0.929 ± 0.003
		WG	0.928	0.605	0.739 ± 0.032	0.738 ± 0.033	0.932 ± 0.001	0.932 ± 0.002	0.846 ± 0.002	0.842 ± 0.005	0.878 ± 0.005	0.914 ± 0.002

Table 1: **Our test-time adaptation significantly improves the test accuracy of CBMs.** For each setting, we compare the performance of CBM based on three variants of our adaptation method: adaptation with CSA loss in Equ (1), with LPA loss in Equ (5), and with both the CSA and LPA losses included. We also report the baseline performance of the foundation model based on zero-shot prediction (ZS) and linear probing (LP).

CBM and the pseudo-labels:

$$L_{\text{LPA}}(\mathbf{W}, \mathbf{b}) \;=\; -\frac{1}{|\mathcal{D}_t^b|} \sum_{\mathbf{x}_t \in \mathcal{D}_t^b} \log \boldsymbol{\sigma}_{\hat{y}_t}(\mathbf{f}_t^{(\text{cbm})}(\mathbf{x}_t)), \quad (5)$$

where $\boldsymbol{\sigma}_k(\mathbf{r})$ takes the logits $\mathbf{r}$ and outputs the softmax probability for class k. Using this objective, the (linear) label predictor is adapted such that the CBM's predictions on test samples are consistent with their pseudo-labels.

4. Experiments

4.1. Setup

We evaluate our proposed test-time adaptation with two different approaches for preparing the concept bottleneck (CBM): 1) using a general-purpose concept bank where natural language concept descriptions and modern vision-language models (*e.g.*, Stable Diffusion (Rombach et al., 2022)) are being leveraged to automatically generate concept examples for finding the Concept Activation Vectors (CAVs) (Kim et al., 2018) (each CAV corresponds to a $\mathbf{c}_{si} \in \mathcal{C}_s$) (Yuksekgonul et al., 2023; Wu et al., 2023); and 2) using learned concept vectors that optimize the concept-based prediction accuracy in an unsupervised manner (Yeh et al., 2020).

We simulate distribution shifts by adopting the settings from Wu et al. (2023) and evaluate on two datasets: Waterbirds (Sagawa et al., 2019b) and Metashift (Liang & Zou, 2021). Waterbirds dataset is for a two-class classification task ("landbird" vs. "waterbird"). In the source domain, landbird (waterbird) images are always associated with the land (water) background, while in the target domain, the correlation with the background is flipped, *i.e.*, landbird (waterbird) images are always on the water (land) background. Metashift has two classes of "cat" and "dog", and it simulates the disparate correlation to the backgrounds in a similar way. Source cat images are always correlated with a sofa or bed in the background, while dog images are always correlated with a bench or bike in the background. For evaluation, we randomly split 90:10 equally across the correlation types, *i.e.*, 10% of dog images with sofa, 10% of dog images with bed, 10% of cat images with bench, and 10% of cat images with bike. In the target domain, both classes are always correlated with a shelf in the background. We use CLIP:ViT-L-14 (Radford et al., 2021) as the backbone foundation model.

We report the performance in terms of two metrics: averaged group accuracy (AVG) and worst-group accuracy (WG). AVG is the average (per-class) accuracy across the classes, and WG is the minimum accuracy across the classes. A model well-generalized to distribution shifts should have high AVG and WG, with a small gap between the source and target domains. We repeated each experiment for 50 trials and report the mean and standard error.

4.2. Results and Discussion

The results are given in Table 1. In the source domain, we observe that CBMs always outperform the classification accuracy obtained based on feature embeddings of the foundation model (using either zero-shot (ZS) or linear probing (LP)). However, in the target domain, without adaptation, the accuracy of CBMs is not even close to the accuracy of ZS or LP. This observation confirms the need for test-time adaptation in order to achieve reliable performance of CBMs post-deployment (see Tables 2 and 3 in the appendix for additional results).

With our proposed adaptive concept bottleneck, the test-time accuracy (both AVG and WG) is significantly increased in all cases, even outperforming that of feature-based predictions. For the explored type of distribution shifts (disparate reliance on covariates), we find that LPA is more crucial than CSA to boost performance. But having both of them is more beneficial, and it achieves the best results.

We also highlight that the performance of our method is dependent on the quality of pseudo-labels (see Table 3). Future work would include exploring approaches to im-

prove the quality of pseudo-labels via ensembling, augmentations (Zhang et al., 2022), or nearest-neighbors voting (Chen et al., 2022).

5. Conclusion

We have explored the robustness of concept bottleneck for foundation models under distribution shifts at test time. We proposed a dynamic concept bottleneck approach, leveraging concept-score alignment (CSA) and linear probing adaptation (LPA) to enhance the model's interpretability and adaptability. Our preliminary findings indicate that this classification pipeline, using foundation models as a backbone followed by an adaptive concept bottleneck, offers not only strong test-time performance but also valuable post-deployment insights. In future work, we plan to extend our analysis to a broader array of foundation models and distribution shifts, explore better pseudo-labeling, and provide an in-depth analysis of the resulting interpretations to further validate our approach.

References

Abid, A., Yuksekgonul, M., and Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pp. 66–88. PMLR, 2022.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 700–712, 2020.

AlBadawy, E. A., Saha, A., and Mazurowski, M. A. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Chen, D., Wang, D., Darrell, T., and Ebrahimi, S. Contrastive test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 295–305. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00039. URL https://doi.org/10.1109/CVPR52688.2022.00039.

Choi, J., Raghuram, J., Feng, R., Chen, J., Jha, S., and Prakash, A. Concept-based explanations for out-of-distribution detectors. In *International Conference on Machine Learning*, pp. 5817–5837. PMLR, 2023.

Eslami, S., Meinel, C., and De Melo, G. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1151–1163, 2023.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019a.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jung, S., Lee, J., Kim, N., Shaban, A., Boots, B., and Choo, J. CAFA: Class-aware feature alignment for test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19060–19071, 2023.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2021.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=aEFaE0W5pAd.

Moayeri, M., Rezaei, K., Sanjabi, M., and Feizi, S. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, pp. 25037–25060. PMLR, 2023.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019a. URL http://arxiv.org/abs/1911.08731.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019b.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E. D., Kurakin, A., and Li, C. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023.

Ye, H., Xie, C., Cai, T., Li, R., Li, Z., and Wang, L. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Zhang, M., Levine, S., and Finn, C. MEMO: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

A. Background

Concept Bottleneck Models.

A Concept Bottleneck Model (CBM) (Koh et al., 2020) projects the feature embeddings from a backbone model onto a *concept space* (linear subspace of $\mathbb{R}^d$ spanned by concept vectors), and make class predictions based on the concept-projected embeddings. Let $\mathcal{C}_s = \{\mathbf{c}_{s1}, \ldots, \mathbf{c}_{sm}\}$ define a concept bank (basis) consisting of m concept vectors, where each concept vector $\mathbf{c}_{si} \in \mathbb{R}^d$ lies in the feature embedding space, and represents a high-level concept (*e.g.*, "stripes", "dots"). Let $\mathbf{C}_s = [\mathbf{c}_{s1} \,/\, \|\mathbf{c}_{s1}\|_2 \;\cdots\; \mathbf{c}_{sm} \,/\, \|\mathbf{c}_{sm}\|_2]^\top \in \mathbb{R}^{m \times d}$ define the corresponding concept-projection matrix, whose rows are the unit-normalized concept vectors. This matrix projects a feature representation $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^d$ to a vector of m *concept scores* as $\mathbf{v}_{\mathbf{C}_s}(\mathbf{x}) = \mathbf{C}_s\, \boldsymbol{\phi}(\mathbf{x})$, where the i-th concept score is given by $\frac{\langle \boldsymbol{\phi}(\mathbf{x}),\, \mathbf{c}_{si} \rangle}{||\mathbf{c}_{si}||_2} \in \mathbb{R}$. The CBM first maps a high-dimensional feature representation to a lower-dimensional (here $m \ll d$) concept-score space (acting like a bottleneck), and follows it with a *label predictor*, which is a simple linear or fully-connected layer that maps the concept scores into class predictions (Yuksekgonul et al., 2023; Oikarinen et al., 2023). Formally, the CBM can be defined as $\mathbf{f}_s^{(\text{cbm})}(\mathbf{x}) := \mathbf{W}_s \mathbf{v}_{\mathbf{C}_s}(\mathbf{x}) + \mathbf{b}_s = \mathbf{W}_s \mathbf{C}_s\, \boldsymbol{\phi}(\mathbf{x}) + \mathbf{b}_s$, where $\mathbf{W}_s \in \mathbb{R}^{L \times m}$, $\mathbf{b}_s \in \mathbb{R}^L$ defines the linear *label predictor*. A key advantage of the CBM is that its predictions are a linear combination of the high-level concept scores, which allows for better interpretability of the model. Since the label predictor of a CBM is chosen to be simple (*e.g.*, a linear layer), its performance is strongly dependent on the construction (richness) of the concept bank.

Preparing the concept bottleneck.

There are various ways of defining the concept vectors $\mathbf{c}_{si}$ in the concept prediction layer $\mathbf{v}_{\mathbf{C}_s}(\mathbf{x})$. Early works on CBM required the training (source) dataset to have concept annotations from domain experts in addition to the class labels, and the concept predictor is trained on this (Koh et al., 2020). Subsequent works have also explored learning the concept vectors in an unsupervised manner (*i.e.*, without any concept annotations) (Yeh et al., 2020; Choi et al., 2023). More recently, natural language concept descriptions and modern vision-language models (*e.g.*, Stable Diffusion (Rombach et al., 2022)) are being leveraged to automatically generate concept examples (Yuksekgonul et al., 2023; Wu et al., 2023) for finding the Concept Activation Vectors (CAVs) (Kim et al., 2018) (each CAV corresponds to a $\mathbf{c}_{si} \in \mathcal{C}_s$), or to directly guide the construction of concept bank $\mathcal{C}_s$ (Oikarinen et al., 2023). We highlight that in all prior works (to our knowledge) the *concept bank remains static*, *i.e.*, once the set of concept vectors is defined and the CBM is deployed, its predictions are made based on the predefined concepts, regardless of any distribution shift at test time.

B. Additional Experiments

Method		CIFAR10	
		AVG (↑)	WG (↑)
Backbone	Source	0.723	0.381
+ ZS	Target	0.374 ± 0.038	0.132 ± 0.032
Backbone	Source	0.872	0.764
+ LP	Target	0.496 ± 0.013	0.215 ± 0.015
PCBM (Yuksekgonul et al., 2023)	Source	0.796	0.665 ± 0.001
	Target	0.426 ± 0.012	0.160 ± 0.013
Yeh *et al.* (Yeh et al., 2020)	Source	0.885 ± 0.002	0.760 ± 0.032
	Target	0.496 ± 0.170	0.206 ± 0.197

Table 2: **Predictions based on high-level semantic concepts are not necessarily more robust against distribution shifts.** For CIFAR10, we use CLIP:ResNet50 as the backbone. We report the average accuracy (AVG) and worst-group accuracy (WG) across the classes.

Dataset			ZS	LP	Yuksekgonul et al. (2023)				Yeh et al. (2020)			
					w/o adaptation	+ CSA	+ LPA	+ CSA + LPA	w/o adaptation	+ CSA	+ LPA	+ CSA + LPA
Metashift	Source	AVG	0.957	0.972	0.979 ± 0.001	-	-	-	0.972 ± 0.001	-	-	-
		WG	0.934	0.960	0.969 ± 0.003	-	-	-	0.960 ± 0.001	-	-	-
	Target	AVG	0.705	0.835	0.890 ± 0.006	0.620 ± 0.049	0.713 ± 0.005	0.676 ± 0.009	0.840 ± 0.009	0.834 ± 0.009	0.749 ± 0.008	0.690 ± 0.005
		WG	0.460	0.720	0.850 ± 0.013	0.279 ± 0.110	0.476 ± 0.017	0.398 ± 0.018	0.712 ± 0.018	0.700 ± 0.020	0.512 ± 0.016	0.400 ± 0.010

Table 3: **Negative results of our test-time adaptation.** In the target domain, the model faces Metashift images with random Gaussian noise (Hendrycks & Dietterich, 2019a). When the performance of zero-shot inference is poor in the target domain, the pseudo-label cannot serve as a reliable reference for the test-time adaptation.