Understanding and Enhancing Robustness of Concept-Based Models

Sanchit Sinha[1], **Mengdi Huai**[1,2], **Jianhui Sun**[1], **Aidong Zhang**[1]

[1]University of Virginia
[2]Iowa State University
{sanchit, js9gu, aidong}@virginia.edu, mdhuai@iastate.edu

Abstract

Rising usage of deep neural networks to perform decision making in critical applications like medical diagnosis and financial analysis have raised concerns regarding their reliability and trustworthiness. As automated systems become more mainstream, it is important their decisions be transparent, reliable and understandable by humans for better trust and confidence. To this effect, concept-based models such as Concept Bottleneck Models (CBMs) and Self-Explaining Neural Networks (SENN) have been proposed which constrain the latent space of a model to represent high level concepts easily understood by domain experts in the field. Although concept-based models promise a good approach to both increasing explainability and reliability, it is yet to be shown if they demonstrate robustness and output consistent concepts under systematic perturbations to their inputs. To better understand performance of concept-based models on curated malicious samples, in this paper, we aim to study their robustness to adversarial perturbations, which are also known as the imperceptible changes to the input data that are crafted by an attacker to fool a well-learned concept-based model. Specifically, we first propose and analyze different malicious attacks to evaluate the security vulnerability of concept based models. Subsequently, we propose a potential general adversarial training-based defense mechanism to increase robustness of these systems to the proposed malicious attacks. Extensive experiments on one synthetic and two real-world datasets demonstrate the effectiveness of the proposed attacks and the defense approach. An appendix of the paper with more comprehensive results can also be viewed at https://arxiv.org/abs/2211.16080.

Introduction

With growth of highly specialized architectures for a variety of use-cases and their superior performance, Deep Neural Networks (DNNs) are increasingly being used in sensitive and critical applications such as medical diagnosis, employment/recruiting, financial credit analysis, etc. However, widespread adoption of such models faces several challenges - primarily the black-box nature of their predictions. Many recent research works have proposed explanation methods which provide deep understanding of model predictions by providing "explanations". Explanations range from being local in nature where they assign importance scores to the features present in an input sample to being global in nature where the model identifies certain "concepts" present in the input sample. A concept can be thought of as an abstraction of features which are usually shared across multiple similar sample points. For example, in Figure 1, a concept can be entirely clinical "osteophyptes- femur", "sclerosis-tibia", etc. Usually DNNs are trained end-to-end, which makes it difficult to isolate concepts and even harder to make them human understandable. To alleviate this, concept-based approaches have been proposed (Koh et al. 2020; Alvarez Melis and Jaakkola 2018) which map a sample from input space to a concept space and subsequently map the concept space to the prediction space. The concept space usually consists of high-level human understandable concepts. A model trained incorporating either manually curated or automatically learned concepts increases both interpretability and reliability of its predictions. One such example, as proposed in (Koh et al. 2020), Concept Bottleneck Models (CBMs), can help domain experts quickly identify any discrepancy and intervene when and where needed. CBMs also offer generalizability in the sense that any DNN can be easily converted into a CBM by resizing an intermediate layer to correspond to the size of any closed concept set pre-selected by domain experts. The training of such models uses standard training procedure with a loss function augmented with an extra term from the bottleneck layer.

Although incredibly simple in formulation and training, many recent works have demonstrated certain flaws in CBMs which warrant an increased caution in their widespread applications. For exmaple, Margeloiu et al. (Margeloiu et al. 2021) demonstrated that computing the saliency maps with respect to a single concept does not capture the position of that concept in the image itself. Similarly, Mahinpei et al. (Mahinpei et al. 2021) demonstrate that CBMs suffer from "information leakage" where more than necessary information is encoded in a concept - making them adulterated with non-relevant noise resulting in unreliable downward predictions.

In this paper, we aim to study the security vulnerability and robustness of concept-based models to carefully crafted malicious attacks, where an adversary with a malevolent intent aims to introduce perturbations to clean sam-

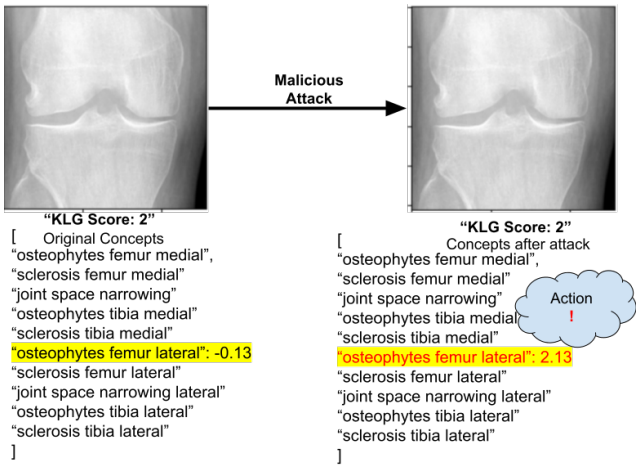

Figure 1: Example of how a value of a concept indicating osteophytes (bone spurs) can be maliciously changed although the actual severity of disease quantified by Kellgren-Lawrence grade (KLG Score) remains the same.

ple image and modify it in an adversarial manner to manipulate the concepts predicted by the model. Specifically, we first demonstrate how concepts learned by a concept-based model can be manipulated by introducing adversarially generated perturbations in input samples. The goal of attacker is to effectively manipulate concepts **without** changing the final model predictions. We propose and study different concept attacks - concept erasure, concept introduction and concept confounding - all of which disrupt the concept set predicted by a well trained concept-based model. Note that proposed attacks can be generalized for any concept-based model. We utilize Concept Bottleneck Models (Koh et al. 2020) in this paper as an example to demonstrate the efficacy of our attacks on one of the most popular concept-based modelling paradigm. To improve trust and reliability of concept-based models, it is important that both concepts and predictions are robust to malicious attacks. Instilling trust in predictions is a well researched problem in adversarial literature. However, the robustness of concepts is an open question. In our paper, we focus on analyzing robustness of concepts without changing model predictions. This critical difference creates important distinction between our proposed attacks and standard adversarial attacks where the goal of attacker is to change the prediction label.

As shown in Figure 1, an attacker can easily disrupt concepts without changing the prediction label. These disrupted concepts can cause misinterpretations as shown in Figure 1 - the concept "osteophytes femur lateral"- which quantifies amount of growth of bone spurs in the upper bone has been maliciously changed to a very high value. This disruption, especially in high security settings, can prompt remedial actions - like expensive oral medicines or even surgery to fix a "supposed" problem even if it does not exist. The fact that such concepts can be manipulated without any perceptible change in the appearance of the input sample and its final prediction essentially defeats the utility of concept-based models in critical applications as these attacks can be very hard to detect. To alleviate this, in addition to studying attacks, we also propose a general adversarial training-based defense mechanism to improve the robustness of the learning models against the proposed attacks on concepts. We conduct comprehensive experiments on different datasets of varying risk levels - ConceptMNIST, CUB and OAI, and the derived experimental results demonstrate the efficacy of both our attacks and defense.

Related Work

Related work on concept-level explanations. To incorporate a broader perspective on model decision making in sensitive applications such as medical diagnosis or financial forecasting (Suo et al. 2020; Xun et al. 2020), concept attribution methods have been proposed. These methods provide a high level abstract notion of explanations by aligning model explanations with human-understandable concepts improving overall reliability. Several popular methods which automatically learn concepts are detailed (Kim et al. 2018; Ghorbani et al. 2019; Yeh et al. 2020; Wu et al. 2020; Goyal et al. 2019). On the other hand providing concept priors have been utilized to align model concepts with human understandable concepts (Zhou et al. 2018; Murty, Koh, and Liang 2020; Chen et al. 2019).

Related work on concept bottleneck models (CBMs). Concept bottleneck models were initially limited to specific use-cases. More recently, the applications of such bottleneck models was generalized in a recent work (Koh et al. 2020) which postulated that any prediction model architecture can be transformed into a CBM by simply resizing any intermediate layer to represent a human-understandable concept representation. Similar work on utilizing and improving CBMs for various downstream tasks include (Sawada and Nakamura 2022; Jeyakumar et al. 2021; Pittino, Dimitrievska, and Heer 2021; Bahadori and Heckerman 2020).

Related work on robustness of interpretations. Although explanations have enabled deep understanding of DNNs, there are concerns regarding their robustness. (Ghorbani, Abid, and Zou 2019) showed that explanations can easily be misled by introducing imperceptible noise in the input image. Several other works have highlighted similar problems on vision, natural language and reinforcement learning such as (Adebayo et al. 2018; Dombrowski et al. 2019; Slack et al. 2020; Kindermans et al. 2019; Sinha et al. 2021; Huai et al. 2020). Similarly, concept explanation methods are also fragile to small perturbations to input samples (Brown and Kvinge 2021). Such concerns regarding fragility of model explanations have prompted related research in improving robustness of explanation methods. For example, (Levine, Singla, and Feizi 2019; Lakkaraju, Arsov, and Bastani 2020; Mangla, Singh, and Balasubramanian 2020) proposed learning more robust feature attributions while (Alvarez Melis and Jaakkola 2018; Soni et al. 2020; Huai et al. 2022) try to learn more robust concepts. However, existing defense methods (Levine, Singla, and Feizi 2019; Lakkaraju, Arsov, and Bastani 2020; Mangla, Singh, and Balasubramanian 2020) on improving the robustness of the feature-level model explanations cannot be directly adopted here. The

reason is that they focus on post-hoc interpretations, while we work on intrinsic concept-based interpretable network.

Methodology

This section investigates vulnerability of CBMs to malicious attacks. We first introduce details of the proposed attack strategies. Subsequently, we propose a defense mechanism **Robust Concept Learning (RCL)**, based on which we train robust models to prevent malicious attacks. Even though our experiments are conducted on CBMs, attacks are general enough and can be used to attack any concept-based model.

Malicious Attacks against CBMs

In this section, we propose a general optimization framework for designing malicious attacks. Here, we use K and T to denote the number of class labels and the number of concepts, respectively. In a CBM model, we are given a set of training samples $\{(x_i, y_i, c_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^D$ denotes the i-th training sample, $y_i \in \{1, \cdots, K\}$ is the target classification label for sample x_i, and $c_i \in \mathbb{R}^T$ is a vector of T concepts. CBMs usually consider two components, i.e., the concept component $g(\cdot)$ and the prediction component $f(\cdot)$. Specifically, CBMs consider the form $f(g(x))$, where $g : \mathbb{R}^D \rightarrow \mathbb{R}^T$ maps an input x into the concept space, and $f : \mathbb{R}^T \rightarrow \mathbb{R}^K$ maps concepts into the final prediction. CBMs define task accuracy as how accurately $f(\cdot)$ predicts label y, and concept accuracy as how accurately $g(\cdot)$ predicts concept c. For sample x, we use $\mathcal{U}(x; f, g)$ to denote its concept-based explanations generated by g to explain the predicted classification label (i.e., $argmax f(g(x))$). Let $G(\mathcal{U}(x; f, g), \mathcal{U}(x + \delta; f, g))$ denote the attacker's goal of maximizing the difference between the generated concept-based explanations before and after the attacks. In order to achieve the attacker's attacking goal, we propose the following framework:

$$\max_{||\delta||_\infty \leq \epsilon_{thresh}} \quad G(\mathcal{U}(x; f, g), \mathcal{U}(x + \delta; f, g))$$
$$\text{s.t.}\, argmax f(g(x + \delta)) = argmax f(g(x)), x + \delta \in [0, 1]^D$$

where δ denotes the adversarial perturbation, ϵ_{thresh} controls the magnitude of the whole adversarial perturbations, and $\mathcal{U}(x + \delta; f, g)$ is the generated concept-based explanations to interpret the predicted class label for the crafted adversarial sample $x + \delta$. Objective function is used to maximize the difference of generated concepts before and after the attacks. The first constraint is enforced to make sure that predictions of sample x is identical before and after the attack. The second constraint guarantees that generated perturbation is imperceptible so it cannot be easily detected. l_∞ norm is most commonly used when considering imperceptible perturbations and measures the feature with the largest amount perturbation, regardless of number of other features that have been maliciously modified. By solving the above optimization problem, the attacker can find an optimal perturbation that can maximize the attacker's goals. Depending on how to define the attacker's goal, we categorize three different types of attacking, which are given as following:

- **Erasure:** Concept erasure attack seeks to subtly delete a particular concept without changing the class prediction result. The gap in perception and absence of concepts would be puzzling to an analyzer and very difficult to detect, especially in datasets where every image of the same class does not have the same concepts - while still seemingly giving the same final prediction. Note that in CBMs, the importance score of the j-th concept for sample x is calculated as $g_j(x)$. In practice, for CBMs, we usually have a pre-defined threshold γ that is used to determine whether a concept is a relevant concept. Specifically, for sample x, the j-th concept is a relevant concept if $g_j(x) - \gamma \geq 0$. Let $S_{x,Rev}$ denote the set of the targeted initially relevant concepts. In order to remove the presence of an initially relevant concept, the attacker's goal is defined as follows,

$$\sum_{j \in S_{x,Rev}} (\mathbb{I}[\gamma - g_j(x + \delta)] - \mathbb{I}[\gamma - g_j(x)]), \quad (1)$$

  where $\mathbb{I}[\cdot]$ is the indicator function, δ denotes the crafted adversarial perturbation, and γ is the given threshold. Note that for the j-th initially relevant concept, we have $g_j(x) - \gamma > 0$, which means that $\mathbb{I}[\gamma - g_j(x)] = 0$. The attacker aims to craft the adversarial perturbation δ such that this j-th concept becomes the non-relevant concept, i.e., $\gamma > g_j(x + \delta)$. In other words, the attack is successful if and only if $\mathbb{I}[\gamma - g_j(x + \delta)] = (\mathbb{I}[\gamma - g_j(x + \delta)] - \mathbb{I}[\gamma - g_j(x)]) = 1$, where $\mathbb{I}[\gamma - g_j(x)] = 0$. The above objective is used to maximize the attacker's goal by reducing the importance score of these initially relevant concepts such that their importance scores are less than the threshold γ. By solving the above objective, the attacker can find an optimal perturbation that can remove the presence of initially relevant concepts for sample x.
- **Introduction:** Concept introduction attack aims to manipulate the presence of non-relevant concepts without modifying the classification result. This hinders accurate analysis of model's interpretations by providing mixed interpretations. The attacker tries to introduce new non-relevant concepts which were not previously present in the concept set of the original sample. For sample x, let $S_{x,Non}$ denote the set of targeted concepts that do not originally present in sample x. The attacker's goal of attacking the presence of these targeted initially non-relevant concepts can be formulated as follows,

$$\sum_{j \in S_{x,Non}} (\mathbb{I}[g_j(x + \delta) - \gamma] - \mathbb{I}[g_j(x) - \gamma]), \quad (2)$$

  where δ denotes the perturbation to be optimized. Note that for the j-th initially non-relevant concept, if $g_j(x + \delta) - \gamma \geq 0$, we can say this initially non-relevant concept becomes the relevant concept after perturbation. The above loss defines attacker's goal - maximizing presence of targeted non-relevant concepts. Specifically, above loss function aims to maximize the attacker's goal by increasing the importance scores of the targeted initially non-relevant concepts such that these targeted concepts' importance scores are larger than the threshold γ. To

achieve his goal of maximizing the presence of the initially non-relevant concepts for sample x, attacker can solve above objective to find an optimal perturbation.

- **Confounding:** Concept confounding attack attempts to build on top of both erasure and introduction by simultaneously removing relevant concepts and introducing non-relevant concepts. The concept confounding attack is a much more powerful attack than just the concept introduction attack as it also removes concepts while maintaining the same model prediction. This can be especially troublesome as it would defeat any purpose of training models with concept bottlenecks. Let $S_{x,Rev}$ and $S_{x,Non}$ denote index set of the targeted initially relevant concepts and the set of the targeted initially non-relevant concepts, respectively. In this case, attacker's goal can be mathematically represented as follows,

$$\sum_{j \in S_{x,Rev}} (\mathbb{I}[\gamma - g_j(x+\delta)] - \mathbb{I}[\gamma - g_j(x)]) \quad (3)$$
$$+ \sum_{j \in S_{x,Non}} (\mathbb{I}[g_j(x+\delta) - \gamma] - \mathbb{I}[g_j(x) - \gamma]),$$

  where δ denotes the adversarial perturbation to be optimized. The above objective is used to maximize the attacker's goal by decreasing the importance scores of these targeted initially relevant concepts to reduce their presence and increasing these non-relevant concepts' importance scores to introduce their presence.

The above schemes define attacker's goals from different aspects. The detailed implementation of the above schemes is given in the extended version. Based on above proposed adversarial attacks, we can perform the security vulnerability analysis to understand how motivated attackers can craft malicious examples to mislead CBMs to generate wrong concepts. The magnitude of perturbation reflect the robustness of CBMs to attacks. The smaller the magnitude of the crafted adversarial perturbations is, the less robust the generated concepts are to the adversarial attacks.

Improving Concept Robustness

Our goal here is to design a defense mechanism which can effectively generate concept-based explanations robust to malicious attacks. Note that in CBMs, we consider bottleneck models of form $f(g(x))$, where g maps an input into the concept space and f maps concepts into a final class prediction. Let $\mathcal{L}_Y = l(f(g(x_i); y_i))$ and $\mathcal{L}_C = \Sigma_j^T l(g_j(x_i), c_i^j)$ denote the classification training loss and the concept training loss over the i-th training data, respectively, where T is the total number of concepts and l represents Binary Cross Entropy or Root Mean Square Error loss.

Hybrid training paradigm: In order to learn the concept component g and the class prediction component f, traditional works (Koh et al. 2020; Margeloiu et al. 2021) usually adopt two common ways of learning CBMs - sequential and joint. We discuss both paradigms in brief below:

- **Sequential Training:** Learns the concept model g by minimizing the concept training loss and subsequently learns the class prediction model f by minimizing the classification loss independently. Mathematically it can be thought of minimizing training objective detailed in Equation 4 first with $\gamma = 0$ and then subsequently minimizing with $\lambda = 0$. As concepts once learned are never updated again during prediction model optimization, the concepts learned are completely independent of the prediction task.
- **Joint Training:** Learns both concept and prediction models (f and g) by minimizing both concept and classification loss jointly in an end-to-end manner. Mathematically it can be thought of minimizing the entire training objective Equation 4 with appropriate values of γ and λ. As concepts and prediction task are learned jointly, concepts learned are not independent of the prediction task as there is some guidance of gradient directions from the prediction part of the model f in the concept model g.

As demonstrated in the extended version, sequential training has lower concept error but worse task performance as compared to joint training (consistently shown by Figure 2 in (Koh et al. 2020)). Hence, there exists a tradeoff between concept and task loss while using joint or sequential training paradigm. However, as we will demonstrate in Tables 1, 2 and 3, joint training shows higher vulnerability of concepts to malicious attacks - implying that concepts learned during joint training are less robust as compared to those learned in sequential. This behavior is expected - as concepts learned during joint training have higher chances of being spuriously correlated to predictions, making them easier to be maliciously attacked.

To overcome this and achieve a better trade-off between concept robustness and prediction performance, we propose a new hybrid training paradigm by combining the sequential and joint training methods. Specifically, in our proposed hybrid training method, we first freeze the prediction model and only let the concept model learn for the first half of total epochs. Subsequently, we unfreeze the complete model and let training continue for the remainder of epochs with a lower learning rate. Based on this, we formulate training loss as follows, where (x_i, c_i, y_i) is a data point sampled from image set (X), concept set (C), and label set (Y):

$$\mathcal{L}_{f,g} = \Sigma_i[\gamma * l(f(g(x_i); y_i)) + \lambda * \Sigma_j^T l(g_j(x_i); c_i^j)], \quad (4)$$

where the first and second terms represent the task and concept losses for i-th training sample with T total number of concepts, respectively, and the values of $\gamma \in \{0, 1\}$ and $\lambda \in \mathbb{R}$. Using the above loss formulation, the complete model parameters $\theta_{f,g}$ are updated as follows:

$$\theta_{f,g} = \begin{cases} \theta_g - \omega * \nabla_{\theta_g} \mathcal{L}_{f,g} \ \ (\gamma = 0) & \text{if epoch} \leq \text{N/2}, \\ \theta_{f,g} - \omega^{'} * \nabla_{\theta_{f,g}} \mathcal{L}_{f,g} & \text{epoch} > \text{N/2} \end{cases}$$

where ω and $\omega^{'}$ represent learning rates. The above proposed hybrid training method is a two-stage training paradigm. Specifically, during the optimization procedure, for the first half of epochs, we set $\gamma = 0$ and $\lambda = 1$ and learning rate ω such that we can first freeze the class prediction model and only train the concept model. Subsequently, in the remaining epochs, we focus on full model training by setting γ as

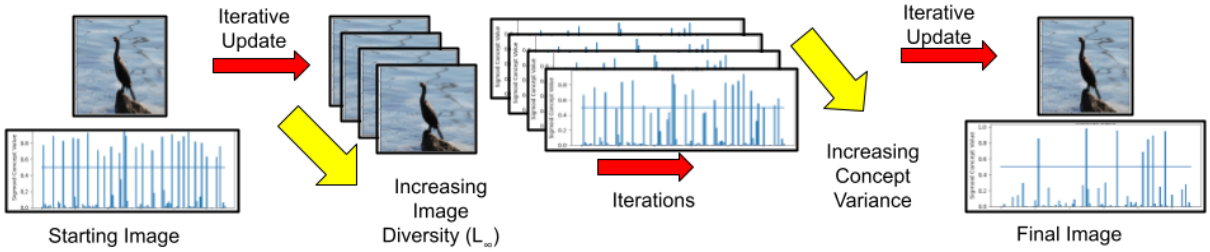

Figure 2: Iterative perturbations to generate diverse training images. Corresponding histograms represent concepts across the spectrum. Proposed augmentation generates images that should belong to same concept class but contain wider variance

1 and assigning a pre-defined appropriate weight value to λ and a different (smaller) learning rate ω'. The specifics of the training procedures are further detailed in the extended version.

Generate diverse training data using adversarial augmentation. The essential reason why an attacker can easily introduce malicious perturbations in a sample is the lack of sample diversity in each concept class. The data distribution of each concept can be discrete and highly dispersed. For example, in the CUB (birds) dataset (Reed et al. 2016) - a sample set containing numerous different types of birds with e.g. 'WingColor==Black' concept class, which would still not be enough to cover all possible combinations of birds of different sizes, shapes, etc. Hence, the distribution of 'WingColor==Black' concept has huge vacancies in its domain that CBM fails to explore while a malicious attacker can easily manipulate. One way to make it difficult for the attacker to exploit such 'vacancies' in data distribution (previously unexplored by CBMs) is to augment the training set by injecting diverse training samples which smoothen the concept distribution space. Intuitively, it simulates a weak attacker and generates images that look perceptually similar - but with potentially different concept classes which in turn, significantly enriches the spectrum of concepts existing in the training data.

Robust concept learning (RCL). We introduce our proposed approach to effectively generate robust concept-based explanations here. Our framework alternates between an inner maximization, where images are iteratively updated with perturbations that increase diversity in concept distribution; and an outer minimization, where model parameters are optimized to find a sweet spot between class prediction, concept accuracy, and concept robustness. Specifically, in the inner loop, we aim to find a perturbed input $\tilde{x}_i$, such that, its difference from true input x_i is smaller than a budget ϵ_{thresh} (i.e., $\|x_i - \tilde{x}_i\|_\infty \leq \epsilon_{thresh}$), while it maximizes the concept divergence loss $l(g(\tilde{x}_i), c_i))$ (i.e. the concept misclassification error) at the same time. The motivation is to generate images that appear identical but are widely diversified in terms of the concept distribution. Formally, we iteratively update $\tilde{x}_i$ in the inner loop as follows,

$$\tilde{x}_i \leftarrow \tilde{x}_i + \epsilon * sign(\nabla_{\tilde{x}_i} l(g(\tilde{x}_i), c_i)) \quad (5)$$

Figure 2 provides an illustration of how the inner loop is effective in generating images with high diversity. We plot the original image as well as intermediate images at each updating step and the ultimate generated perturbed image along with their associated concepts. As can be seen from the concept histogram of each image - concept distributions vary without much perceptual changes in image.

Once the perturbed sample is iteratively generated, in the outer loop, we aim to optimize the model weight such that it achieves a good balance between task classification, concept prediction as well as concept robustness. The updated total loss $\mathcal{L}_{f,g}$ we optimize is,

$$\mathcal{L}_{f,g} = \Sigma_i[\gamma * \mathcal{L}_Y + \lambda * \mathcal{L}_C + \alpha * \mathcal{L}_{adv}] \quad (6)$$

where $\mathcal{L}_Y$ and $\mathcal{L}_C$ denote task classification loss and concept prediction loss, respectively as described above. The adversarial loss $\mathcal{L}_{adv}$ is calculated by $\mathcal{L}_{adv} = l(g(\tilde{x}_i), c_i)$ (l is the same as defined before). γ, λ, and α are tunable weights.

To combine the advantages from both joint and sequential models, we adopt a hybrid training paradigm as described previously, in which we disable the training of the prediction model and only allow the concept model to be trained in the first half, before unfreezing the prediction model and training the whole model with a lower learning rate in the second half. Our empirical investigation shows this hybrid paradigm outperforms both sequential and joint training alone by a nontrivial margin. Pseudocode of RCL is detailed in the extended version.

Experimental Study

Dataset Description

We test the proposed approaches on the following 3 datasets of varying domains and levels of security and trust required. For a standard classification task such as digit or bird identification, a wrong concept set is not a very concerning outcome - however for a medical diagnosis - a wrong concept set can be catastrophic. For a more comprehensive description of datasets, please refer to the extended version.

- **ConceptMNIST (C-MNIST):** We augment the original MNIST dataset by constructing concepts of each image by including 2 physical characteristics of numbers in the image along with 10 standard non-overlapping concepts representing one hot encodings of the numbers, resulting in a size 12 concept vector for each image. **[Low Risk]**
- **CUB:** The Caltech-UCSD Birds-200-2011 dataset (Reed et al. 2016) consists of photos of 200 classes of birds. Pre-processing of the dataset is performed exactly as

(Koh et al. 2020). The final dataset consists of 112 concepts for each class with concepts representing physical traits of the birds like wing color, beak size, etc. **[Low Risk]**

- **OAI:** The Osteoarthritis Initiative (OAI) dataset (Nevitt, Felson, and Lester 2006) consists of X-ray images and clinical data for about 36,000 patients over 4 years of study who pose a risk of knee osteoarthritis. The task is X-ray grading into 4 different risk categories (KLG Score). Each image has 10 medical concepts from X-ray images such as bone spacing. For more comprehensive description, refer to (Pierson et al. 2019). **[High Risk]**

Benchmarking and Ablation Study

We train CBMs on all 3 datasets using different training strategies - sequential and joint proposed by (Koh et al. 2020) and hybrid as previously discussed with hyperparameters mentioned in the extended version. We train the respective models for CUB and OAI datasets based on the hyperparameters mentioned in (Koh et al. 2020) as well as train hybrid models on both datasets to compare with standard models. In addition, we also train robust models using RCL (Algorithm detailed in the extended version) utilizing both joint and hybrid training paradigms. The task errors for C-MNIST and CUB are classification error while for OAI, task error is Root Mean Square Error (RMSE) as the prediction label is a continuous variable. Concept error for C-MNIST and CUB is 0-1 error (binary concepts), while for OAI, concept error is RMSE (concepts are continuous variables). The benchmark results are reported in the extended version. As expected, performance of hybrid models lie between joint and standard models in task and concept performance. Usage of all 3 training paradigms presents a trade-off between task and concept performance depending on use-case. For example, in high-risk settings, where concept accuracy is paramount (e.g. medical diagnosis), sequential can be utilized. Whereas tasks where small errors in concepts can be tolerated but prediction performance is important, joint can be utilized. Hybrid paradigm provides a good trade-off between both sequential and joint paradigms.

Attack Results and Discussion

We report results on a set of 500 randomly chosen samples from the test set for all 3 datasets. We skip all samples which - a) have wrong task prediction label and b) have concept accuracy $\leq 60\%$ for binary valued concepts (C-MNIST, CUB) or concept Root Mean Square Error (RMSE) ≥ 0.6 for continuous valued concepts (OAI). In all our experiments, we begin by reporting attack success results using standard adversarial attack setting on the joint model (Adv. Attack (Joint)), followed by results for proposed attacks on standard Joint, Sequential and Hybrid models, and finally on joint and hybrid models trained using RCL. As concept scores are not explicitly used during optimization in standard adversarial setting, we expect attack success metrics to be relatively low. Mathematically, standard adversarial setting can be formulated with β setting to 0.

Hyperparameter Selection All attacks are performed with 2 distinct sets of hyperparameters. The first set of hyperparameters controls properties of attacks - budget (ϵ_{thresh}), number of steps (N) and learning rate (ϵ). We refer popular benchmarks[1] for hyperparameter selection decisions. The second set of hyperparameters controls the influence of concepts (α) and influence of predictions (β) to the loss optimized during attacks.

	C-MNIST	CUB	OAI
Adv. Attack(Joint)	$4 \pm 0\%$	$1 \pm 0\%$	$0 \pm 0\%$
Joint	$67 \pm 5\%$	$66 \pm 7\%$	$62 \pm 3\%$
Sequential	$44 \pm 4\%$	$56 \pm 4\%$	$54 \pm 5\%$
Hybrid	$51 \pm 4\%$	$59 \pm 6\%$	$54 \pm 4\%$
RCL-Joint	$22 \pm 2\%$	$32 \pm 2\%$	$5 \pm 2\%$
RCL-Hybrid	$18 \pm 2\%$	$23 \pm 2\%$	$1 \pm 0\%$

Table 1: Attack results on erasure attacks for datasets - C-MNIST, CUB and OAI averaged over 3 different seeds.

Results on Erasure Attack. As erasure attacks attempt to remove or "flip" relevant concepts in a particular sample, we run our attack by targeting all possible concepts for each selected sample. For C-MNIST and CUB, we classify a sample as being "flipped" if it is no longer classified as being 'present' based on sigmoid classification ($\geq$0.5) after the attack. For OAI, we consider a concept as being "flipped" if its absolute value changes with more than a pre-defined threshold after attack. In the experiments, we set this threshold as 2 (hyperparameter settings - extended version) which we believe can result in a significant shift in medical diagnosis of knee-pain. Table 1 shows the percentage of successful flips for standard adversarial attack, followed by joint, sequential and hybrid models across all 3 datasets. A higher percentage of flipped concepts implies a higher success rate for the attack. We observe about 60% of concepts are successfully flipped across 3 datasets for joint, sequential and hybrid models with the highest and lowest success rates being on joint and sequential respectively as discussed before. Joint and hybrid models trained using RCL show significantly lower attack success rates of 18%, 23% and 1% on C-MNIST, CUB and OAI - demonstrating RCL's success as a defense. As targeted concept scores are not used during optimization in standard adversarial attack setting, we observe successful flip percentages to be low (4%, 1% and 0% for C-MNIST, CUB and OAI). Figure 3 demonstrates attack results on a sample from CUB dataset.

Results on Introduction Attack. As opposed to erasure, introduction attacks attempt to introduce non-relevant concepts to concept prediction set of a perturbed sample image. As an introduction attack specifically targets non-relevant concepts, this attack is not suitable for data with continuous concept values (e.g. on OAI, all concepts are deemed to be relevant for prediction). We report percentage of new concepts introduced (%Introduced) in perturbed image before and after attack. Goal of attacks here is to introduce previously non-relevant concepts, hence higher value of %Introduced implies higher success of the attack. In addition, we also report percentage of concepts retained (%Re-

[1] github.com/MadryLab/mnist_challenge,cifar10_challenge

Figure 3: Top-most: Original image and associated concepts. Following 3 images show final concept set after attack on selected concept (red). Concepts in red previously classified as "present" selectively attacked and removed.

tained) from the original concept set to ensure no significant change in originally relevant concepts (ideally close to 100%). Table 2 shows the percentage of non-relevant concepts successfully introduced on standard adversarial setting followed by Joint, Sequential and Hybrid models across CUB and C-MNIST datasets. For CUB, around 33%, 30% and 31% while for C-MNIST, 114%, 71% and 102% concepts are successfully introduced for joint, sequential and hybrid models respectively. Average percentage of relevant concepts retained are relatively high ($\geq$90%) for all 3 models. Models trained with RCL are less susceptible to attack, with introduction percentages around 20% for both datasets. Non-relevant concept scores are not explicitly used during optimization in standard adv. attack, we observe low values of both percentage introduced and retained.

Results on Confounding Attack. Confounding is a combination of both erasure and introduction attacks. As confounding essentially maximizes the difference between original and perturbed concept sets, we report the Jaccard Similarity index (JSI) for binary concepts (CUB, C-MNIST) and average (Avg-Δ) and minimum (Min-Δ) absolute change in concept values for continuous concepts (OAI). Lower JSI values indicate a greater difference in concept sets before and after attack, implying higher success of confounding attack. Similarly, higher values of Avg-Δ implies confounding attack disrupts values for all concepts by a significant amount whereas, high Min-Δ implies that even minimum concept disruption caused is still relatively large - reducing trust in all concept predictions. Table 3 reports JSI on CUB and C-MNIST for joint, sequential and hybrid models across all 3 datasets. We observe relatively low values of JSI (around 0.2 for CUB and 0.4 for C-MNIST respectively)

	C-MNIST		CUB	
	%Intro.	%Ret.	%Intro.	%Ret.
Adv. Attack	$53 \pm 2\%$	86%	8 ± 2 %	77%
Joint	$114 \pm 4\%$	96%	$33 \pm 5\%$	92%
Sequential	$71 \pm 2\%$	93%	$30 \pm 4\%$	97%
Hybrid	$102 \pm 2\%$	94%	$31 \pm 3\%$	95%
RCL-Joint	$18 \pm 3\%$	96%	$13 \pm 2\%$	93%
RCL-Hybrid	$13 \pm 2\%$	**96%**	23 ± 2**%**	**97%**

Table 2: Attack results on introduction attacks for C-MNIST and CUB averaged over 3 different seeds. %Intro denotes the percentage of new concepts introduced wrt. original concept set, while %Ret denotes the percentage of concepts retained from the original concept set. If more than original number of concepts are introduced, introduction percentage $\geq$100%.

	C-MNIST	CUB	OAI	
	Jacc Sim	Jacc Sim	Avg-Δ	Min-Δ
Adv. Attack	0.61	0.51	0.21	0.03
Joint	0.38	0.20	0.57	0.13
Sequential	0.44	0.23	1.06	0.35
Hybrid	0.41	0.25	0.7	0.21
RCL-Joint	0.52	0.49	5.8e-3	3.1e-4
RCL-Hybrid	0.55	0.54	1.9e-3	1.8e-5

Table 3: Attack results on confounding attacks for datasets - C-MNIST, CUB and OAI avg. over 3 seeds. Jaccard Sim. represents Jaccard Similarity indices (JSI). Lower JSI value implies concept set before and after are more dissimilar.

showcasing the success of proposed attack. We also observe models trained using RCL demonstrate relatively higher JSI values of around 0.5 for both datasets, thus making them less susceptible attack. Similarly, for OAI, RCL demonstrates much better robustness against confounding attacks with average absolute change 2 orders of magnitude less than standard (0.0019 vs 0.35) - which further validates success of RCL. Adversarial attack's JSI is relatively high as none of the concept scores are utilized during optimization.

Effect of varying attack budget (ϵ_{thresh}): We also report additional results with varying attack budgets in the extended version. As expected, attack success rates increase with increasing value of attack budgets (ϵ_{thresh}). However, with higher ϵ_{thresh}, images start losing visual imperceptibility implying a trade-off between attack success and budget.

Conclusion

In this paper, we conducted the first systematic study on malicious attacks against concept bottleneck models (CBMs). Specifically, we first proposed 3 different novel attack methods to show that current CBMs are vulnerable to adversarial perturbations. To defend such adversarial attacks and enhance the robustness of CBMs against adversarial attacks, we proposed a generic adversarial training-based defense mechanism. Extensive experimental results on real-world datasets not only show that current CBMs are vulnerable to malicious perturbations, but also demonstrate the effectiveness of the proposed defense mechanism.

Acknowledgments

This work is supported in part by the US National Science Foundation under grants 2213700, 2217071, 2008208, 1955151. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Alvarez Melis, D.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Bahadori, M. T.; and Heckerman, D. E. 2020. Debiasing concept bottleneck models with instrumental variables. *arXiv preprint arXiv:2007.11500*.

Brown, D.; and Kvinge, H. 2021. Brittle interpretations: The Vulnerability of TCAV and Other Concept-based Explainability Tools to Adversarial Attack. *arXiv preprint arXiv:2110.07120*.

Chen, R.; Chen, H.; Ren, J.; Huang, G.; and Zhang, Q. 2019. Explaining neural networks semantically and quantitatively. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9187–9196.

Dombrowski, A.-K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.-R.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3681–3688.

Ghorbani, A.; Wexler, J.; Zou, J.; and Kim, B. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*.

Goyal, Y.; Feder, A.; Shalit, U.; and Kim, B. 2019. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Huai, M.; Liu, J.; Miao, C.; Yao, L.; and Zhang, A. 2022. Towards Automating Model Explanations with Certified Robustness Guarantees. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Huai, M.; Sun, J.; Cai, R.; Yao, L.; and Zhang, A. 2020. Malicious Attacks against Deep Reinforcement Learning Interpretations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 472–482.

Jeyakumar, J. V.; Dickens, L.; Cheng, Y.-H.; Noor, J.; Garcia, L. A.; Echavarria, D. R.; Russo, A.; Kaplan, L. M.; and Srivastava, M. 2021. Automatic Concept Extraction for Concept Bottleneck-based Video Classification. *arXiv preprint arXiv:2206.10129*.

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

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 267–280. Springer.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348. PMLR.

Lakkaraju, H.; Arsov, N.; and Bastani, O. 2020. Robust and stable black box explanations. In *International Conference on Machine Learning*. PMLR.

Levine, A.; Singla, S.; and Feizi, S. 2019. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*.

Mahinpei, A.; Clark, J.; Lage, I.; Doshi-Velez, F.; and Pan, W. 2021. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*.

Mangla, P.; Singh, V.; and Balasubramanian, V. N. 2020. On Saliency Maps and Adversarial Robustness. *arXiv preprint arXiv:2006.07828*.

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*.

Murty, S.; Koh, P. W.; and Liang, P. 2020. ExpBERT: Representation Engineering with Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2106–2113.

Nevitt, M.; Felson, D.; and Lester, G. 2006. The osteoarthritis initiative. *Protocol for the cohort study*, 1.

Pierson, E.; Cutler, D.; Leskovec, J.; Mullainathan, S.; and Obermeyer, Z. 2019. Using machine learning to understand racial and socioeconomic differences in knee pain. *NBER Machine Learning and Healthcare Conference*.

Pittino, F.; Dimitrievska, V.; and Heer, R. 2021. Hierarchical Concept Bottleneck Models for Explainable Images Segmentation, Objects Fine Classification and Tracking. *Objects Fine Classification and Tracking*.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 49–58.

Sawada, Y.; and Nakamura, K. 2022. Concept Bottleneck Model with Additional Unsupervised Concepts. *IEEE Access*.

Sinha, S.; Chen, H.; Sekhon, A.; Ji, Y.; and Qi, Y. 2021. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 420–434.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Soni, R.; Shah, N.; Seng, C. T.; and Moore, J. D. 2020. Adversarial TCAV–Robust and Effective Interpretation of Intermediate Layers in Neural Networks. *arXiv preprint arXiv:2002.03549*.

Suo, Q.; Zhong, W.; Xun, G.; Sun, J.; Chen, C.; and Zhang, A. 2020. GLIMA: Global and Local Time Series Imputation with Multi-directional Attention Learning. In *IEEE International Conference on Big Data, Big Data 2020, Atlanta, GA, USA, December 10-13, 2020*, 798–807. IEEE.

Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M. R.; and Tai, Y.-W. 2020. Towards Global Explanations of Convolutional Neural Networks With Concept Attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8652–8661.

Xun, G.; Jha, K.; Sun, J.; and Zhang, A. 2020. Correlation Networks for Extreme Multi-label Text Classification. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

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

Zhou, B.; Sun, Y.; Bau, D.; and Torralba, A. 2018. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 119–134.