IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2024, accepted 5 September 2024, date of publication 11 September 2024,
date of current version 23 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3457784

RESEARCH ARTICLE

Exploring the Impact of Conceptual Bottlenecks on Adversarial Robustness of Deep Neural Networks

BADER RASHEED[1], MOHAMED ABDELHAMID[2], ADIL KHAN[3], (Member, IEEE), IGOR MENEZES[4], AND ASAD MASOOD KHATAK[1], (Senior Member, IEEE)

[1]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates
[2]Institute of Data Science and Artificial Intelligence, Innopolis University, 420500 Innopolis, Russia
[3]School of Computer Science, University of Hull, HU6 7RX Kingston upon Hull, U.K.
[4]Faculty of Business, Law and Politics, University of Hull, HU6 7RX Kingston upon Hull, U.K.

Corresponding author: Bader Rasheed (b.rasheed@innopolis.university)

ABSTRACT Deep neural networks (DNNs), while powerful, often suffer from a lack of interpretability and vulnerability to adversarial attacks. Concept bottleneck models (CBMs), which incorporate intermediate high-level concepts into the model architecture, promise enhanced interpretability. This study delves into the robustness of Concept Bottleneck Models (CBMs) against adversarial attacks, comparing their original and adversarial performance with standard Convolutional Neural Networks (CNNs). The premise is that CBMs prioritize conceptual integrity and data compression, enabling them to maintain high performance under adversarial conditions by filtering out non-essential variations in input data. Our extensive evaluations across different datasets and adversarial attacks confirm that CBMs not only maintain higher accuracy but also show improved defense capabilities against a range of adversarial attacks compared to traditional models. Our findings indicate that CBMs, particularly those trained sequentially, inherently exhibit higher robustness against adversarial attacks than their standard CNN counterparts. Additionally, we explore the effects of increasing conceptual complexity and the application of adversarial training techniques. While adversarial training generally boosts robustness, the increment varies between CBMs and CNNs, highlighting the role of training strategies in achieving adversarial resilience.

INDEX TERMS Concept bottleneck models, adversarial attacks, robustness, interpretable models.

I. INTRODUCTION

Deep neural networks (DNNs) have seen significant growth in recent years, being widely applied in fields like computer vision, natural language processing, and speech recognition, as well as in sectors such as healthcare, agriculture, energy, and transportation. In healthcare, transparency is crucial for medical image analysis and disease diagnosis, ensuring clinical trust and regulatory compliance. Similarly, secure systems are necessary in agriculture for crop monitoring and yield prediction to protect sensitive data from cyber threats. Moreover, security measures are essential in energy and transportation for optimizing consumption, traffic management, and autonomous driving, preventing malicious attacks, and ensuring safety and efficiency.

As DNNs continue to permeate other critical domains, the demand for models that not only deliver high accuracy but also uphold transparency and security standards is crucial for fostering trust and reliability in their applications. Nonetheless, DNNs have faced criticism due to their "black-box" nature, posing significant challenges in terms of interpretability. The interpretability issue of DNNs is often addressed through techniques like saliency maps [28] and class activation mapping (CAM) [15], [40], or perturbation-based methods such as Local Interpretable Model-agnostic Explanations (LIME) [27]. These techniques facilitate an understanding of what the network is focusing on or how it processes inputs. Despite the recent advancements in these

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

methods, their explanations can be highly subjective and may not consistently provide insights across different dataset samples.

Another issue is the susceptibility to adversarial attacks, wherein deliberate modifications to input data result in erroneous outputs. Such vulnerabilities present severe risks, especially in applications where safety and reliability are paramount [1], [16], [37]. Different defensive strategies have been developed to make DNNs less vulnerable to adversarial attacks. Among these, adversarial training stands out as a particularly effective method [7], [17] [38], [39] that has emerged as a robust strategy to strengthen DNNs by exposing them to adversarial examples during training. However, its effectiveness varies as it usually does not cover attacks that differ from those encountered during training.

To address these challenges and strengthen DNNs against adversarial attacks, thereby enhancing their resilience and robustness, this study conducted a series of simulations using CBMs. CBMs, introduced by [2], integrate an intermediary layer of human-understandable concepts preceding final decision-making. This approach not only enhances interpretability by making model reasoning accessible and modifiable but also maintains competitive accuracy compared to traditional DNNs. In this study, we hypothesize that by imposing a structured conceptual framework on the model through CBMs, DNNs may not only preserve predictive performance but also demonstrate greater resistance to adversarial perturbations, which typically exploit model-specific vulnerabilities in less structured prediction environments. We explore a range of adversarial attacks, from simple to complex, to evaluate the CBMs' ability to maintain integrity under different conditions. We aim to enhance the models' resilience by incorporating multiple concepts, ensuring that final predictions rely on robust features.

Our research adopts a comprehensive approach that integrates theoretical analysis with empirical validation. Theoretically, we establish a framework to assess CBMs' vulnerability to input perturbations, including a detailed Information Loss Analysis to understand the impact of concept integration on model robustness. Experimentally, we compare the resilience of CBMs and traditional CNNs against adversarial attacks. Our investigation aims to determine whether integrating explicit, human-understandable concepts within neural networks' decision-making pipelines enhances their resistance to such manipulations.

Our study follows a two-part structure. First, we benchmark the adversarial robustness of CBMs against standard CNNs across various datasets, focusing on white-box attacks where the adversary knows the model's architecture and parameters. We employ different attack methodologies in these scenarios. Second, we investigate how increasing conceptual complexity and the application of adversarial training techniques influence the robustness of these models. We evaluate how adversarial training affects CBMs and CNNs differently, highlighting the role of training strategies in achieving adversarial resilience.

This innovative approach offers a promising pathway to enhance our understanding of DNNs robustness against adversarial perturbations. This study highlights the potential of CBMs in critical applications and lays the groundwork for future research focused on developing more secure and interpretable AI systems.

II. BACKGROUND

A. CONCEPT BOTTLENECK MODELS

One of the principal strengths of employing CBMs lies in their capacity to enhance explainability and performance. They achieve this by mapping inputs to a series of understandable concepts, known as *bottleneck*, which are then utilized for predictive tasks, thereby augmenting accuracy and explainability [35]. Compared to alternative methods, concept bottleneck models have been shown to elucidate a higher percentage of model predictions, thereby surpassing them in terms of explainability [36].

The theoretical foundation of CBMs [2], [19], [20] involves decomposing the model into two stages. A neural network g maps the input $x \in \mathbb{R}^d$ to a human-specified concept space, where $c \in \mathbb{R}^k$ represents a vector of k concepts. Subsequently, another neural network h maps the concepts k to the final prediction $y \in \mathbb{R}$. As a result, the prediction of a CBM can be represented as $f(X) = h(g(X))$. For training CBMs, we use $L_{C_j} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ to measure the difference between the predicted and true concepts, and $L_Y : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ to measure the difference between predicted and true targets. The two common ways for training CBMs are:

1) The Sequential bottleneck models, which learn the network $g = argmin_g \sum_{i,j} L_{C_j}(g_j(x^{(i)}); c_j{}^{(i)})$, then it uses the trained concept predictor g to learn $h = \arg\min_h \sum_i L_Y \left(h(g(\mathbf{x}^{(i)})); y^{(i)}\right)$.
2) The Joint bottleneck models, which learn both of the networks h, g at the same time by minimizing the weighted sum

$$h, g = \arg\min_{h,g} \sum_i \left[L_Y \left(h \left(g \left(\mathbf{x}^{(i)} \right) ; y^{(i)} \right) \right) + \lambda \sum_j L_{C_j} \left(g \left(\mathbf{x}^{(i)} \right) ; c_j^{(i)} \right) \right] \quad (1)$$

for some $\lambda > 0$. This λ hyperparameter controls the trade-off between task and concept loss. For the models trained on the datasets CUB and Concept MNIST, we set λ to 0.01 and 1 respectively.

One of the main advantages of this new architecture of CBMs is that it allows turning an end-to-end neural network into a CBM model by resizing one of its layers to have k neurons that represent the concepts to predict, then attach the prediction layer to it.

B. STANDARD END-TO-END MODEL

For comparative analysis, a standard Convolutional Neural Network (CNN) is also examined. This model operates on

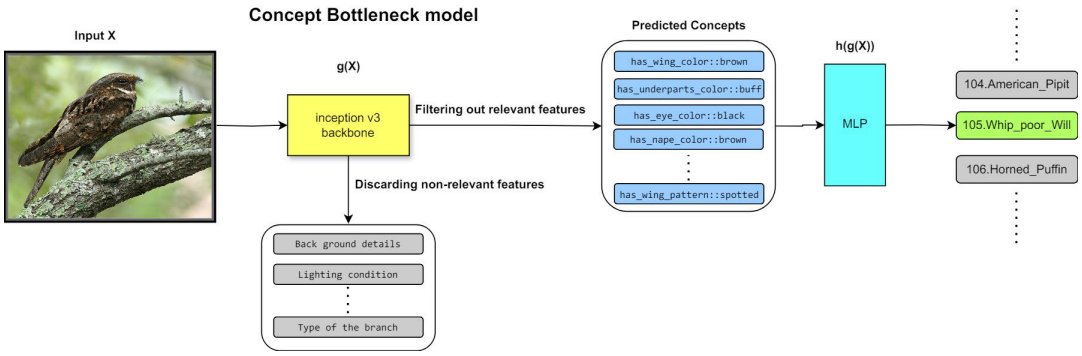


FIGURE 1. The architecture of a Concept Bottleneck Model (CBM). The input *X* is mapped to a series of human-understandable concepts *C* through the function *g*. The concepts are then used by the function *h* to predict the final output *Y*.

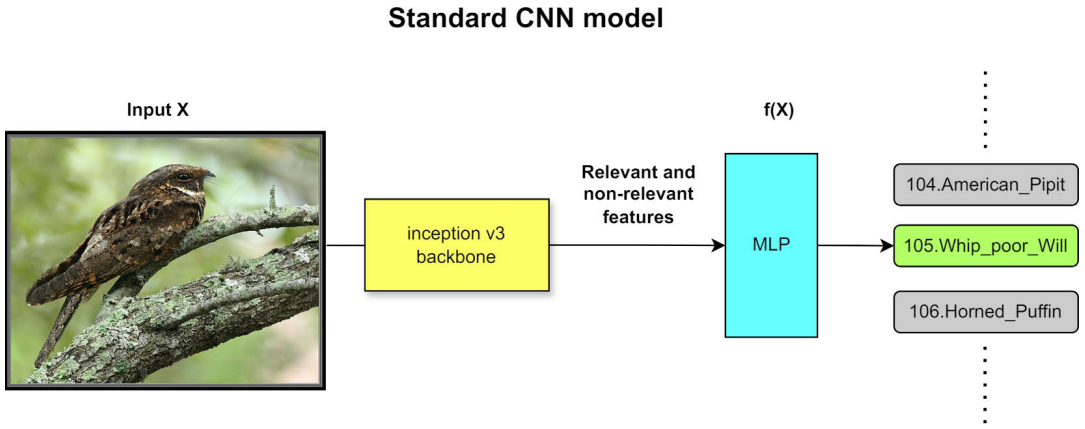


FIGURE 2. A standard CNN. The input *X* is directly used to predict the output *Y* through a series of convolutional layers without the intermediary step of mapping to human-understandable concepts.

an end-to-end basis, classifying inputs directly from raw data without the intermediate step of explicit concept prediction. Its primary mechanism involves automatic feature extraction, which does not segregate data into human-understandable concepts. Techniques like saliency maps and class activation mapping are required to interpret the results of CNNs. On the other hand, CBMs results can be interpreted effortlessly without using any technique.

Moreover, it's insightful to consider the architecture of standard CNN models through the lens of the CBM framework, conceptualized as $f(x) = h(g(x))$. In CNNs, the function g extracts features which are then processed by h to produce the final output. Unlike CBMs, the features extracted by g in CNNs do not represent distinct, interpretable concepts but rather are considered generic features that feed into subsequent network layers.

The training process for CNNs mirrors this architecture:

$$h, g = \arg\min_{h,g} \sum_i L_Y \left(h \left(g \left(\mathbf{x}^{(i)} \right) \right) ; y^{(i)} \right) \tag{2}$$

This framework ensures a balanced comparison between CBMs and CNNs, highlighting any differences in performance attributable to the interpretability and structured learning approach of CBMs.

C. ADVERSARIAL ATTACKS

Adversarial attacks [1] involve manipulating the input data to a neural network in a manner that causes the network to make a mistake. These perturbations are often imperceptible to humans but can drastically alter the network's predictions. In our study, we operate under the premise that the adversary has complete awareness of the model's architecture and parameters, characterizing the scenario as a white-box attack.

1) PROJECTED GRADIENT DESCENT (PGD)

Projected Gradient Descent [8] is a white-box attack that involves taking multiple steps of gradient ascent over the input data with respect to the loss function, with each step followed by a projection onto the set of allowable perturbations.

The PGD attack can be formalized as follows:

$$x^{(t+1)} = \Pi_{x+S} \left(x^{(t)} + \alpha \cdot \text{sign}(\nabla_x L(\theta, x^{(t)}, y)) \right) \tag{3}$$

where x is the input, y is the true label, θ represents the model parameters, L is the loss function, α is the step size, and Π denotes the projection operation on the set of possible perturbations S.

2) CARLINI & WAGNER (C&W)

The C&W attack [9] finds an adversarial example by solving an optimization problem that minimizes the distance to the original input while also misclassifying it. The C&W L_2 attack can be written as:

$$\min_{\delta} \|\delta\|_2 + c \cdot f(x+\delta) \tag{4}$$

where δ is the perturbation, x is the original input, c is a constant found via binary search, and f is the objective function that causes misclassification.

3) DeepFool

DeepFool [10] is an untargeted, iterative attack that aims to find the closest distance of the input data to the decision boundary of the classifier. For a binary classifier, it can be expressed as:

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{\left\|\nabla f(x^{(t)})\right\|_2} \nabla f(x^{(t)}) \tag{5}$$

where f represents the classifier's decision function.

For a multi-class classifier, DeepFool iteratively perturbs the input along the direction that is normal to the decision boundary of the current closest class. The specific algorithmic details are quite intricate; for an in-depth understanding, refer to the original paper.

D. ADVERSARIAL TRAINING

The role of adversarial training [7] is to enhance the robustness of neural networks by explicitly training them with adversarial examples. This approach incorporates adversarial examples into the training process, where the model learns to classify both clean and perturbed inputs correctly. The adversarial examples are typically generated by applying small but deliberate perturbations to training data, aiming to maximize the training loss. This technique is formalized as follows:

Let $\mathbf{x}$ be the clean input and y its corresponding label. The adversarial example $\mathbf{x}_{\text{adv}}$ is generated by solving the optimization problem:

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \arg\max_{\|\delta\| \leq \epsilon} L(\theta, \mathbf{x}+\delta, y) \tag{6}$$

where δ is the perturbation, ϵ is the magnitude of the allowed perturbation, L is the loss function, and θ are the model parameters.

The model is then trained on a mixture of clean and adversarial examples, which can be expressed as minimizing the expected loss:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\alpha L(\theta, \mathbf{x}, y) + (1-\alpha) L(\theta, \mathbf{x}_{\text{adv}}, y)\right] \tag{7}$$

where $\mathcal{D}$ is the data distribution and α is a hyperparameter that balances the importance of clean and adversarial examples in the training process.

This process helps the model to not only perform well on clean data but also to resist potential adversarial attacks during deployment.

III. QUANTITATIVE ANALYSIS OF CBM ROBUSTNESS

In this section, we outline the theoretical underpinnings guiding our analysis of vulnerability and information loss, setting the stage for understanding how data compression and concept representation enhance model resilience. We begin by examining how CBMs improve resilience by reducing vulnerability to input perturbations through conceptual filtering. Next, we explore the impact of conceptual compression on information loss and its effects on the defensive capabilities of CBMs.

1) REDUCING VULNERABILITY WITH CONCEPTUAL FILTERING

Adversarial robustness consists of a model's ability to maintain its performance and resist crafted perturbations - adversarial attacks - designed to mislead the model's predictions. It is well-established that deep neural networks (DNNs) are vulnerable to such attacks, leading to erroneous predictions due to imperceptible perturbations in natural samples [32].

This study examines various levels of complexity in adversarial attacks by specifying intermediate features that serve as high-level concepts within the CBM framework. This approach paves the way for a detailed vulnerability analysis, revealing how CBMs mitigate the impact of adversarial perturbations on model predictions. It is widely acknowledged in the literature that a system's vulnerability, regardless of its scale, arises from its exposure and sensitivity to hazardous conditions, as well as its ability to manage, adapt to, or recover from such conditions [30].

Given that the interconnectedness of layers in DNNs functions as a complex system, this study similarly views DNNs' vulnerability (V) as the measure of susceptibility to attacks due to small variations in input data. This encompasses both the exposure and sensitivity of DNNs to adversarial conditions and their ability to maintain resilience and robustness in model output predictions. Reducing vulnerability enhances the model's resilience to minor variations, especially those introduced by adversaries to deceive the model. Such resilience is crucial in adversarial scenarios where attackers subtly manipulate input data to mislead the model's predictions without detection.

Reducing vulnerability enhances the model's robustness against adversarial attacks by making it less reactive to high-frequency noise. It also improves the model's ability to generalize across diverse datasets and various attack methods not encountered during training.

In CBMs, vulnerability is inherently reduced ($\Delta V > 0$) as the model's output relies more on distilled concepts C rather than direct input features, as we will demonstrate later. Mathematically, vulnerability represents the norm of the

gradient of the model's output (Y) with respect to its input (X), denoted as $V = \|\nabla_X f(X)\|$, where $f : \mathbb{R}^n \to \mathbb{R}^m$ is a neural network function mapping the input vector $X \in \mathbb{R}^n$ to the output vector $Y \in \mathbb{R}^m$, and $\|\cdot\|$ denotes a suitable norm (e.g., the Euclidean norm).

In CBMs, the prediction function is decomposed into two stages: $f(X) = h(g(X))$. The central hypothesis posits that the concept representation C encapsulates the relevant information in X for predicting Y, thereby filtering out irrelevant variations.

To illustrate the impact on vulnerability, we compare the vulnerability of the model with and without the concept bottleneck:

- Without CBM: The vulnerability is $V_{\text{non-CBM}} = \|\nabla_X f(X)\|$.
- With CBM: The vulnerability becomes $V_{\text{CBM}} = \|\nabla_X h(g(X))\|$.

The reduction in vulnerability, ΔV, is then calculated as $\Delta V = V_{\text{non-CBM}} - V_{\text{CBM}}$. To establish $\Delta V > 0$, we employ the chain rule for differentiation, yielding: $\nabla_X h(g(X)) = (\nabla_X g(X))(\nabla_C h(C))$ where $C = g(X)$.

Since $g(X)$ is designed to capture only the relevant concepts for predicting Y, it inherently acts as a filter, reducing the impact of small perturbations in X on Y. Consequently, this filtering effect means that $\|\nabla_X h(C)\|$ is expected to be smaller than $\|\nabla_X f(X)\|$, as $g(X)$ discards irrelevant variations in X.

Because $h(C)$ operates on a more compressed, concept-focused representation of X, the gradient $\nabla_C h(C)$ is less susceptible to small, irrelevant perturbations in C (and thus X) compared to the gradient of $f(X)$ with respect to X directly.

As a result, the product of these gradients, representing the vulnerability of the CBM, $\|\nabla_X h(g(X))\|$, is smaller than the vulnerability of the non-CBM model, $\|\nabla_X f(X)\|$. This is because both the reduction of irrelevant information by $g(X)$ and the focused prediction mechanism of $h(C)$ contribute to dampening the effect of input perturbations on the output. Thus, we have $\|\nabla_X h(g(X))\| < \|\nabla_X f(X)\|$, which implies that $\Delta V = V_{\text{non-CBM}} - V_{\text{CBM}} > 0$.

This mathematical exposition demonstrates that introducing a concept bottleneck reduces the model's vulnerability to input perturbations, thereby enhancing its adversarial robustness through a quantifiable reduction in vulnerability ($\Delta V > 0$). The theoretical underpinning provided by the Information Bottleneck principle [23] supports our hypothesis that CBMs, by focusing on relevant concepts, reduce the vulnerability of the model's output to input perturbations.

2) IMPACT OF CONCEPTUAL COMPRESSION ON INFORMATION LOSS

In CBMs, Conceptual Compression involves a strategic selection and retention of essential information while discarding non-essential details. This approach shares similarities with information dropout techniques, which are designed to mitigate information loss during processing stages [22]. To understand the impact of this strategy, we turn to Information Loss Analysis. This methodology explains how conceptual compression in CBMs enhances adversarial robustness by filtering out "irrelevant" features from inputs—those that do not substantially contribute to the target outputs. By focusing on relevant features, the model minimizes its vulnerability to being misled by adversarial perturbations that target these non-robust features [13], [26]. This filtration process effectively reduces the attack surface that adversaries can exploit, making the model's predictions more robust by anchoring them in the relevant conceptual features that are less susceptible to adversarial noise.

The mutual information [12] $I(X; Y)$ and $I(X; C; Y)$ can be used to quantify the information loss due to the concept bottleneck. The reduction in mutual information, $\Delta I = I(X; Y) - I(X; C; Y)$, reflects the amount of "irrelevant" information filtered out by the concepts, which is not necessary for predicting Y.

To formalize the relationship between mutual information changes and vulnerability reduction, we start by expressing the mutual information metrics in terms of entropy [11]:

Mutual Information $I(X; Y) = H(Y) - H(Y|X)$: The amount of information that the input variable X contains about the output variable Y.

Mutual Information $I(X; C; Y) = H(Y) - H(Y|X, C)$: The amount of information that X contains about Y mediated through the concepts C.

The reduction in mutual information due to concept compression can then be linked to a reduction in the model's vulnerability to input perturbations. This can be modeled as a function of the entropy reduction:

$$\begin{aligned}\Delta I &= I(X; Y) - I(X; C; Y) \\ &= H(Y) - H(Y|X)] - [H(Y) - H(Y|X, C)\end{aligned}$$

Simplifying, we obtain:

$$\Delta I = H(Y|X, C) - H(Y|X)$$

To demonstrate mathematically and theoretically how C serves as an efficient compression of X in terms of relevant information for predicting Y, we will leverage the concept of a Markov chain. The objective is to establish that C, as a distilled representation of X for predicting Y, does not exceed the information X provides about Y, leading to $I(X; Y) \geq I(C; Y)$.

3) CONCEPTUAL FRAMEWORK AND MARKOV CHAIN ANALYSIS

The Markov chain model [12] used in the conceptual setup illustrates how breaking direct dependencies between raw input features and outputs can inherently limit the pathways through which adversarial perturbations influence the model's predictions. This setup implies that the output Y depends solely on the concept C, and not directly on the raw input X, establishing a form of conditional independence that enhances security measures [33].

Considering the relationship between X, C, and Y, we model it as a Markov chain: $X \rightarrow C \rightarrow Y$. This implies that C is a function of X and that Y, the output, is generated based on C without directly accessing X. In this setup, C captures the relevant aspects of X needed for predicting Y, acting as a bottleneck. This is crucial for the model's focus on relevant features, potentially reducing the model's complexity and enhancing interpretability.

In this study, the Markov chain is defined by the property that the future state (Y) depends only on the current state (C), and not on the sequence of events (or states) that preceded it (X). Mathematically, this is represented as:

$$P(Y|X, C) = P(Y|C)$$

This equation encapsulates the principle of conditional independence in CBMs, indicating that once we know the concepts C, the input X provides no additional information about the output Y.

4) REDUCTION IN UNCERTAINTY ABOUT Y FROM C

The reduction in uncertainty about Y from knowing C can be quantified by $\Delta(I) = H(Y|X, C) - H(Y|X)$. A negative $\Delta(I)$ indicates a more reliable model in the face of data variability and potential perturbations as the uncertainty will be reduced.

Given the identity $I(Y; C \mid X) = H(Y \mid X) - H(Y \mid X, C)$, it follows that $H(Y \mid X, C) = H(Y \mid X) - I(Y; C \mid X)$. This rearrangement emphasizes that the mutual information $I(Y; C \mid X)$ quantifies the reduction in uncertainty about Y due to the knowledge of C, given X.

Since mutual information is inherently non-negative

$$I(Y; C \mid X) \geq 0,$$

this non-negativity implies that:

$$H(Y \mid X, C) \leq H(Y \mid X)$$

Therefore, the entropy of Y conditioned on both X and C can never exceed the entropy of Y conditioned on X alone. This inequality also implies that equality holds if and only if C provides no additional information about Y beyond what is already contained in X. In such a case, C perfectly encapsulates the relevant information from X necessary for predicting Y.

IV. EXPERIMENTAL DESIGN

Our experimental setup involves a comparative analysis of CBMs and traditional CNNs across multiple datasets. We meticulously outline the specifics of the network architectures, training procedures, and adversarial attack scenarios to ensure reproducibility and clarity in evaluating our hypotheses. The goals of our methodology are threefold: 1) To demonstrate the inherent robustness features of CBMs, 2) To explore the effects of adversarial training, and 3) To investigate the impact of conceptual complexity on model performance. Through rigorous testing and analysis, we aim to provide a comprehensive assessment of how well CBMs maintain their integrity and accuracy when faced with sophisticated adversarial challenges.

A. DATASETS DESCRIPTION AND CONCEPTUAL ANNOTATIONS

CBMs heavily rely on datasets that are not only rich in images or data points but also annotated with human-understandable concepts directly related to the output predictions. In our study, we explore two such datasets: Concept MNIST and CUB.

1) CONCEPT MNIST DATASET

A variant of the MNIST dataset, augmented with additional concept labels for each digit, was introduced in [6]. The creation of datasets like MNIST demonstrates the evolving landscape of concept-based AI systems and highlights the need to effectively address uncertainties and incorporate human interventions [34]. The Concept MNIST dataset includes initial and additional concepts:

The initial concepts in the dataset include:

- Non-overlapping concept: This concept is simply the one-hot encoding of the digit in the image, considered as a single concept.
- Overlapping concepts: These include the presence of curved lines and straight lines in the digit.

The new set of overlapping concepts added later to the Concept MNIST dataset includes:

- Intersection Points: Identifies whether a digit has points where lines intersect. For example, digits like '4', '8', '9', and '0' have intersection points, while '1', '2', '3', '5', and '7' do not.
- Closed Loops: Determines whether a digit contains closed loops. Digits such as '6', '8', '9', and '0' feature closed loops, whereas others do not.
- Presence of Horizontal/Vertical Lines: This concept specifically looks for the presence of horizontal or vertical lines in addition to identifying straight and curved lines.
- Top/Bottom Heavy: Examines whether a digit has more visual weight at the top (like '9') or at the bottom (like '6').

2) CUB DATASET

The second dataset used was the Caltech-UCSD Birds-200-2011 (CUB) dataset. This dataset contains 11,788 images of 200 different classes of birds, as detailed in [14] The original dataset included 312 binary concepts representing the features of each bird class. However, after processing the dataset as described in [2], the number of concepts was reduced to 112 for each class.

B. CBM AND CNN ARCHITECTURES FOR EXPERIMENTATION

We derived our model structures from [2] and [6]. As mentioned in section II-B, we consider the architecture of CNNs through the lens of the CBM framework. This approach

allows us to describe the architectures by detailing the g and h networks. Both models employ identical g and h networks as demonstrated in Figs. 1 and 2, but utilize different training mechanisms. The training mechanisms for the Sequential, Joint, and CNN models are explained in detail in sections II-A and II-B. This structure ensures a fair comparison between the models. An in-depth comparison of the architectures of CBMs and CNNs is presented in the subsequent subsections:

1) CONCEPT MNIST MODELS

- **Architecture of Network g:** The network g, acting as the initial feature extractor and concept mapper, comprises the following components:
  - **Convolutional Layers:** Two convolutional layers each with 32 filters of size 3×3, designed to extract spatial hierarchies from input images. These layers employ a stride of 1 and are followed by ReLU activation functions.
  - **Max Pooling:** Each convolutional layer is followed by a max pooling layer with a window of 2×2 and a stride of 2, which reduces the spatial dimensions of the feature maps, thus condensing the information and enhancing feature robustness against small translations.
  - **Fully Connected Layers:** After flattening the output from the convolutional stacks, the data is passed through two fully connected layers. The first has 128 units and the second is split into two segments: 10 units for non-overlapping concepts and an additional set designed to capture overlapping concepts, calculated as $10 + n\_concepts \times 2$ where $n\_concepts$ is the number of overlapping concepts being modeled.
- **Architecture of Network h:** Following the concept layer g, the network h maps these high-level concepts to final class predictions:
  - **Input Processing:** The network accepts input from g, which includes a vector for non-overlapping concepts and another for overlapping concepts.
  - **Fully Connected Layers:** Comprises an initial layer with 32 units followed by a final output layer with 10 units corresponding to the class predictions. Both layers use ReLU activation.

2) CUB MODELS

- **Architecture of Network g**: the network g is an Inception V3 model pretrained on ImageNet and fine-tuned following instructions in [25]. The size of the output layer is changed to match the number of concepts.
- **Architecture of Network h:** Network h operates as a straightforward Multi-Layer Perceptron (MLP), directly mapping the processed concepts from network g to the final class predictions:
  - **Input Layer:** Receives an input of dimension 112 from network g, which encapsulates high-level concepts.
  - **Output Layer:** A fully connected layer that maps the input dimensions directly to the number of classes. This layer is designed to output the final class predictions based on the input concept vector, utilizing a linear transformation followed by a softmax activation.

V. EXPERIMENTS

A. TESTING THE ROBUSTNESS OF CBMS VERSUS DNNS

In this experiment, we evaluate the robustness of Concept Bottleneck Models (CBMs) against standard Convolutional Neural Network (CNN) models by deploying L_∞-norm Projected Gradient Descent (PGD) attacks with varying intensities of ϵ.

Following the training of the models on the Concept MNIST dataset, we subjected them to adversarial attacks as shown in Table 1. Without any adversarial perturbations, the models exhibit comparable accuracy, with the Sequential model slightly surpassing the others. As the attack's ϵ value increases beyond 0.1, the Sequential model maintains a performance edge, outstripping the Joint model by 2.15% and the standard CNN model by 3.67%.

Regarding the CUB dataset results in Table 2, the standard CNN model initially has higher clean accuracy without perturbations. Moreover, within the ϵ range of 0.0005 to 0.001, it continues to outperform the CBMs. Nonetheless, when ϵ surpasses 0.001, the Sequential model begins to demonstrate markedly enhanced robustness. Specifically, when $\epsilon = 0.005$, the Sequential model exceeds the performance of both the Joint and standard CNN models by a margin of 9.03% and 10.7%, respectively. These results illustrate that while standard CNNs may achieve higher accuracy in the absence of perturbations, CBMs, especially the Sequential model, offer superior robustness against adversarial attacks. This robustness is crucial for applications where models need to perform reliably under potentially hostile conditions, making CBMs a valuable approach for developing resilient AI systems.

B. IMPACT OF ADVERSARIAL TRAINING

In this experiment, we are going to adversarially train the CBMs and CNN models using the *PGD* L_∞ attack with different values of ϵ. After training, we test the accuracy of each model before and after the adversarial training against *PGD* L_∞ attacks with various ϵ values. Additionally, we aim to understand how well each model generalizes to different ϵ values other than the ones it was trained on.

- On the Concept MNIST dataset, as represented in Fig 3 (a), the performance of the Sequential model before adversarial training is slightly better than the Joint and CNN models, as observed in the previous section. After adversarial training, the standard CNN model shows a slight edge in robustness compared to the other models across different ϵ values. An interesting observation from Fig 4 is that when the models are adversarially trained with $\epsilon = 0.05$ and 0.1, the Joint model

TABLE 1. The accuracy of the models —sequential, joint, and CNN— on the concept MNIST dataset after being subjected to projected gradient descent (PGD) attacks at various epsilon (eps) intensities.

Epsilon (eps)	Sequential Model Accuracy (%)	Joint Model Accuracy (%)	CNN Model Accuracy (%)
0 (No attack)	**98.77**	98.72	98.67
0.05	93.76	93.14	**93.87**
0.1	**71.39**	69.86	71.36
0.2	**10.21**	8.06	6.54
0.3	**10.21**	8.07	6.54
0.4	**10.21**	8.06	6.54

TABLE 2. The accuracy of the models —Sequential, Joint, and CNN— on the CUB dataset after being subjected to projected gradient descent (PGD) attacks at various epsilon (eps) intensities.

Epsilon (eps)	Sequential Model Accuracy (%)	Joint Model Accuracy (%)	CNN Model Accuracy (%)
0 (No attack)	72.01	75.87	**77.63**
0.0005	34.92	28.75	**43.13**
0.001	21.82	13.65	**23.58**
0.0025	**15.33**	5.30	6.89
0.005	**13.57**	4.54	2.87

generalizes better to higher ϵ values than the other models. As the value of ϵ increases, the performance of the Joint and standard CNN models becomes almost identical across the different ϵ values.

- On the CUB dataset, as represented in Fig 3 (b), the results are more definitive. The Sequential model performs much better before adversarial training when $\epsilon > 0.001$. However, after adversarial training, the standard CNN model outperforms the CBMs by approximately 9%. The standard CNN model also generalizes significantly better than the other models to ϵ values it was not trained on. These findings highlight the importance of adversarial training in enhancing model robustness. While the Sequential CBM shows promising performance on clean data and low-intensity adversarial attacks, the standard CNN model demonstrates superior robustness and generalization capabilities when subjected to higher-intensity attacks and different ϵ values.

C. GENERALIZATION TO OTHER ATTACKS

In this experiment, we evaluate the ability of adversarially trained models to generalize to unseen attacks by subjecting them to *Deepfool* and $C\&WL_2$ attacks.

- On Concept MNIST:
  - $C\&WL_2$ attack: as shown in Table 3 (a), the Joint model outperforms the other models across all ϵ values, except when $\epsilon = 0.1$, where the standard CNN model slightly surpasses the Joint model.
  - *Deepfool* attack: the results of the *Deepfool* attack, displayed in Table 4 (a), indicate that initially, the standard CNN model performs better than the CBMs. However, when the value of $\epsilon > 0.1$, the Sequential model outperforms the other models.
- On CUB dataset: For the CUB dataset, the performance of the standard CNN model is superior to that of the other models when subjected to both $C\&W$ and *Deepfool* attacks. These findings suggest that, while the Joint model shows strong performance against the $C\&W$ attack on the Concept MNIST dataset, and the Sequential model demonstrates resilience under the *Deepfool* attack for higher ϵ values, the standard CNN model consistently shows superior performance on the CUB dataset across both types of attacks.

D. IMPACT OF ADDING MORE CONCEPTS TO THE DATASET

This experiment was conducted exclusively on the Concept MNIST dataset. The original Concept MNIST dataset includes two concepts, as described in section IV-A1. Initially, we start with one concept, which is the non-overlapping one-hot encoding of the numbers in the image. Gradually, we add more concepts to the dataset to test how the number of concepts affects the robustness of the models.

Fig. 5 presents a comparative analysis of three different models' performance on normal and adversarial images across a range of added concepts. The models maintain nearly uniform high accuracy on normal examples, with negligible variance between them. Conversely, the performance on adversarial images reveals distinct patterns. The CNN model's performance on adversarial images is consistently the lowest across all concept counts. This outcome is expected because the CNN model does not rely on the concepts, i.e., changing the number of concepts does not affect its performance on adversarial examples. The Joint model exhibits a variable pattern: its accuracy on adversarial images starts comparably low to the CNN model with zero to three new concepts but then increases sharply, surpassing the Sequential model at four added concepts. As more concepts are added, the Joint model's performance fluctuates, displaying a series of peaks and troughs. This variability

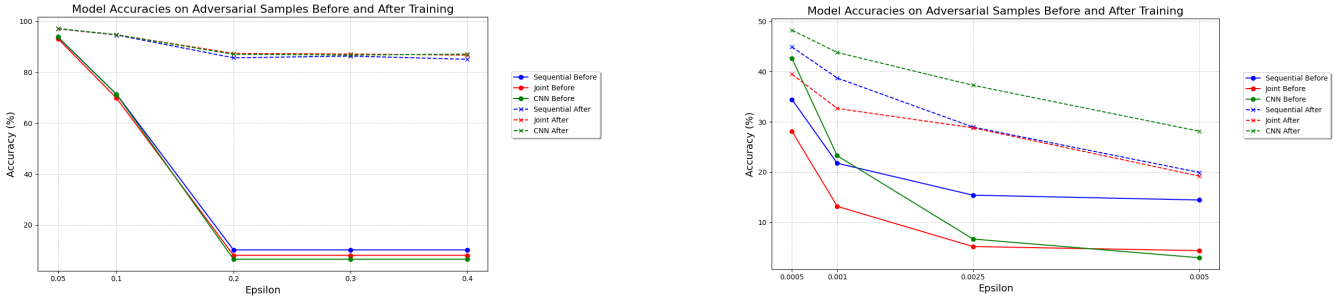


FIGURE 3. Comparing the performance of the models before and after adversarial training on (a) Concept MNIST and (b) CUB datasets.

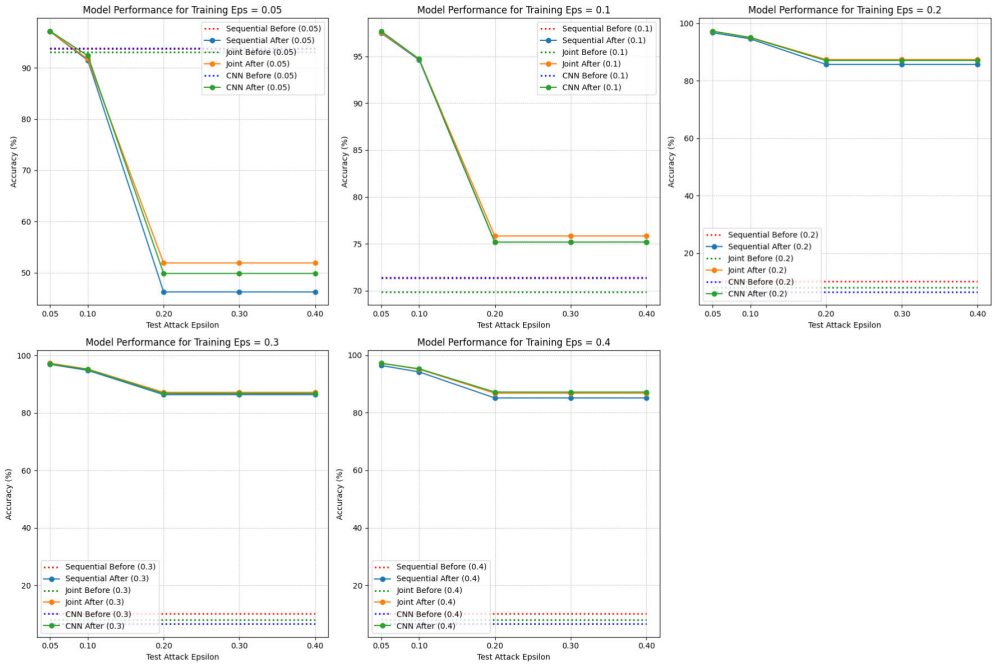


FIGURE 4. The results of adversarial training the models on Concept MNIST dataset. The figures represent training the models with one epsilon and testing their generalization to the other epsilon.

suggests that the Joint model benefits from the addition of concepts but in an inconsistent manner. The Sequential model's accuracy on adversarial images, while higher than the CNN's, also demonstrates some variability but remains between the performance of the Joint and CNN models throughout the range of added concepts. This indicates that while the Sequential model benefits from additional concepts, it does not achieve the same level of robustness as the Joint model at certain points.

VI. DISCUSSION

A. ROBUSTNESS OF CONCEPT-BOTTLENECK MODELS (CBMS) COMPARED TO STANDARD CNNS

The experiments reveal that Concept-Bottleneck Models (CBMs), including Sequential and Joint models, exhibit higher robustness to adversarial attacks than standard Convolutional Neural Networks (CNNs). This is primarily due to the structured prediction strategy that CBMs utilize, effectively maintaining conceptual integrity even under adversarial conditions. Sequential models, which sequentially model intermediate concepts before addressing the final classification task, display the highest robustness, suggesting a protective effect against adversarial manipulation. On the other hand, Joint models, which concurrently train on concept and class predictions, may inadvertently transfer non-robust features to the prediction layer, potentially compromising their robustness [4], [29]. Additionally, CBMs' ability to compress data into concept-based representations reduces the available attack surface, making it challenging for adversaries to exploit detailed data features effectively. This synergy between maintaining conceptual integrity and minimizing data complexity through compression enhances CBMs' defense capabilities, marking them as suitable for security-sensitive applications and highlighting the

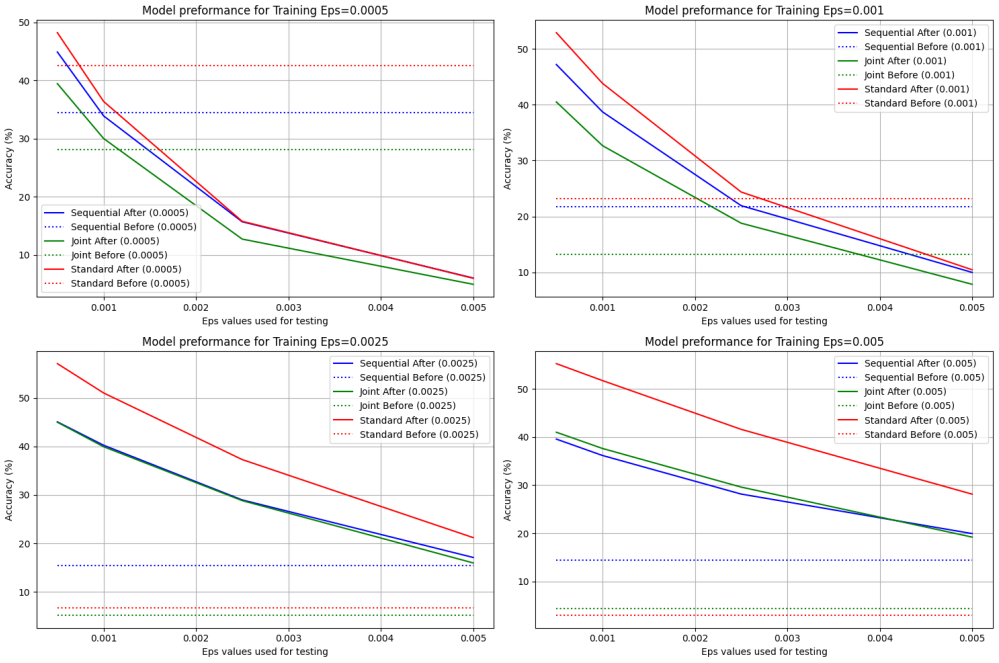


FIGURE 5. The results of adversarial training the models on CUB dataset. The figures represent training the models with one epsilon and testing their generalization to the other epsilon.

TABLE 3. Attacking the adversarially trained models on Concept MNIST (a) and CUB (b) datasets with $C\&WL_2$ attack.

(a)

Epsilon	Sequential	Joint	Standard
0.05	10.16	**15.04**	9.71
0.1	20.66	20.26	**20.87**
0.2	48.94	**51.15**	48.01
0.3	47.14	**49.44**	46.98
0.4	49.95	**52.16**	47.69

(b)

Epsilon	Sequential	Joint	Standard
0.0005	58.51	51.05	**61.74**
0.001	55.13	48.69	**62.15**
0.0025	49.59	48.74	**61.65**
0.005	43.04	44.63	**58.73**

TABLE 4. Attacking the adversarially trained models on concept MNIST (a) and CUB (b) datasets with Deepfool attack.

(a)

Epsilon	Sequential	Joint	Standard
0.05	16.73	15.36	**21.67**
0.1	14.99	13.66	**16.7**
0.2	**22.53**	20.34	20.47
0.3	**21.8**	18.57	20.25
0.4	**22.48**	18.95	21.31

(b)

Epsilon	Sequential	Joint	Standard
0.0005	43.61	38.45	**48.27**
0.001	37.09	32.26	**43.46**
0.0025	24.59	25.18	**35.28**
0.005	13.77	15.59	**24.66**

importance of concept-bottleneck strategies in developing resilient AI systems.

B. IMPACT OF ADVERSARIAL TRAINING

Adversarial training has shown to influence the resilience of both CBMs and CNNs, albeit in distinct manners across model architectures. CNNs, with their focus on local feature extraction, demonstrate enhanced adversarial accuracy primarily against perturbations similar to those seen during training. The localized feature adaptability of CNNs makes them robust against specific types of adversarial attacks where perturbations are aligned with their receptive fields. However, this adaptability may not necessarily extend to novel, sophisticated adversarial strategies that bypass or exploit these local features. CBMs, on the other hand, operate on a higher level of abstraction, focusing on the relationships and structures between concepts within the data. This approach makes them less sensitive to the specific

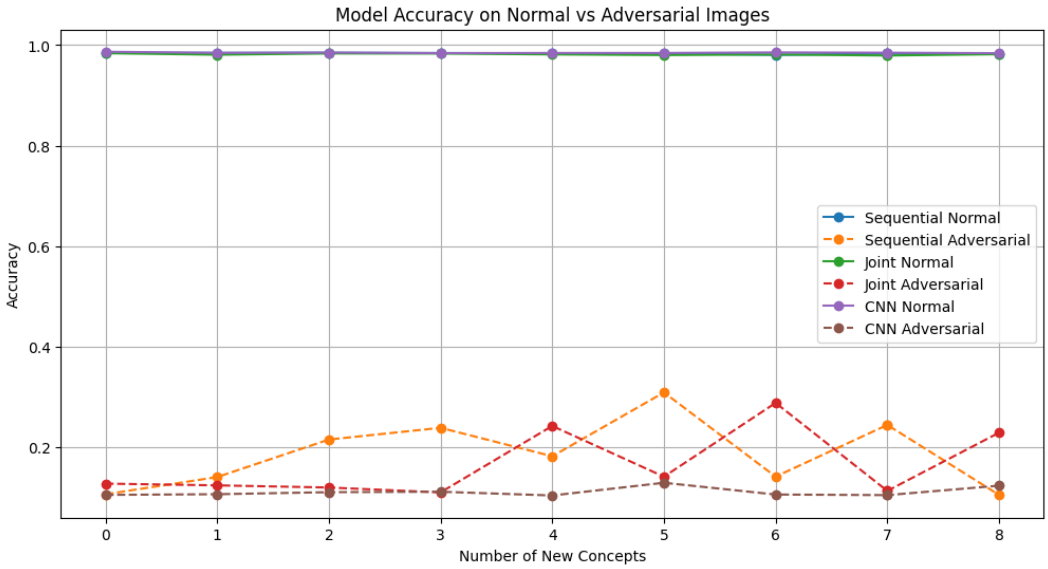


FIGURE 6. Comparing the accuracy of the models on clean and adversarial examples as the number of concepts increase in the dataset.

pixel-level changes that adversarial attacks often employ. However, it also means that their ability to adapt through adversarial training might be more nuanced, as improvements in robustness require adjustments at the conceptual level rather than the feature or pixel level. If adversarial examples do not directly perturb the concepts learned during training but instead manipulate other aspects of the input space, CBMs may fail to adapt. CNNs are able to better adapt to adversarial perturbations seen during training, as they are not confined to a predetermined set of concepts. This partially aligns with theoretical insights, which suggest that CBMs' robustness is inherently tied to their conceptual structure. Despite these insights, all models exhibited vulnerabilities to sophisticated or highly variable adversarial inputs, highlighting the limitations of adversarial training. Future advancements in hybrid training approaches or architectural innovations may lead to the development of neural networks robust not only to known threats but also to evolving adversarial strategies.

C. ROBUSTNESS AND CONCEPTS NUMBER

The results show that the impact of adding concepts on model robustness does not follow a straightforward pattern; instead, it fluctuates, with certain concepts affecting the models differently. This variability indicates that not all concepts contribute equally to robustness, and their impact can depend on the specific architecture (Sequential, Joint, or CNN) and the nature of the concepts themselves. Certain concepts have a differential impact on model accuracy and robustness, indicating that the nature of the concept (e.g., presence of closed loops, intersection points) and its relevance to the task at hand can influence how a model generalizes from training to adversarial contexts.

VII. FUTURE RESEARCH DIRECTIONS

Our findings have laid a foundational understanding of the robustness of Concept-Bottleneck Models (CBMs) and Convolutional Neural Networks (CNNs) against adversarial attacks. To build on this foundation, key areas for future research include:

- **Diverse Datasets:** It is crucial to expand testing to a broader array of datasets. This will help ensure that findings on the robustness of CBMs are not skewed by class and concept imbalances, such as those observed in the CUB dataset. Diverse datasets can provide a more comprehensive assessment of model robustness across different scenarios and data distributions.
- **Adversarial Training Refinement:** There is a need to refine adversarial training algorithms to specifically target the concepts modeled by CBMs. By focusing on concept-specific perturbations, we can enhance the robustness of DNNs, making them more resilient to attacks that exploit specific conceptual vulnerabilities.
- **Minimizing Information Compression:** Developing training mechanisms that minimize the information compressed within the concepts used to predict Y can also reduce vulnerabilities. By focusing on the essential aspects necessary for robust prediction, we can ensure that the models retain critical information while discarding extraneous details that may be exploited by adversarial attacks.
- **Semantic-based Attacks:** Exploring the impact of semantic-based adversarial attacks is another important area of research. These attacks could reveal additional vulnerabilities or strengths in CBMs, particularly when compared to CNNs that do not explicitly model concepts [18]. Understanding how CBMs handle semantic

perturbations can provide insights into their robustness and guide improvements.

- **Concept Investigation:** A systematic investigation into which concepts significantly influence model robustness and why is essential. By focusing on these concepts' geometric properties or semantic depth, we can design CBMs that prioritize robust and essential concepts for classification.

These directions promise to deepen our understanding of adversarial robustness and guide the development of more secure AI systems capable of withstanding sophisticated threats. Addressing these research areas will advance the field of adversarial robustness, leading to the creation of models that are not only accurate but also resilient to adversarial manipulation.

VIII. CONCLUSION

Our comprehensive study demonstrates the enhanced robustness of Concept Bottleneck Models (CBMs) against adversarial attacks compared to traditional Convolutional Neural Networks (CNNs). This robustness stems from the conceptual integrity and data compression capabilities inherent in CBMs, which effectively filter out non-essential input variations. However, when adversarially trained, standard CNNs adapt better to the perturbations seen during training. While standard CNNs can achieve comparable robustness through targeted adversarial training, CBMs inherently offer a more robust framework due to their structured conceptual integration.

Additionally, we investigated the impact of increasing conceptual complexity within CBMs. Our findings indicate that the effect of adding more concepts depends on the robustness of the concepts themselves. Robust concepts enhance the model's resilience to adversarial attacks, while non-robust concepts may not significantly improve performance.

This research underscores the potential of CBMs in sensitive and critical applications where robustness and interpretability are paramount. It also lays the groundwork for future explorations into creating more secure and interpretable AI systems. Future research can enhance the robustness and reliability of CBMs by focusing on diverse datasets, refining adversarial training methods, minimizing information compression, exploring semantic-based attacks, and systematically investigating critical concepts, paving the way for advancements in secure AI technologies.

ACKNOWLEDGMENT

(Bader Rasheed and Mohamed Abdelhamid contributed equally to this work.)

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[2] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," 2020, *arXiv:2007.04612*.

[3] I. Fischer and A. A. Alemi, "CEB improves model robustness," *Entropy*, vol. 22, no. 10, pp. 10–81, Sep. 2020, doi: 10.3390/e22101081.

[4] A. Margeloiu, M. Ashman, U. Bhatt, Y. Chen, M. Jamnik, and A. Weller, "Do concept bottleneck models learn as intended?" 2021, *arXiv:2105.04289*.

[5] I. Fischer, "The conditional entropy bottleneck," *Entropy*, vol. 22, no. 9, p. 999, Sep. 2020, doi: 10.3390/e22090999.

[6] S. Sinha, M. Huai, J. Sun, and A. Zhang, "Understanding and enhancing robustness of concept-based models," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 12, pp. 15127–15135, doi: 10.1609/aaai.v37i12.26765.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57, doi: 10.1109/SP.2017.49.

[10] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582, doi: 10.1109/CVPR.2016.282.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2005, doi: 10.1002/047174882x.

[13] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 125–136. [Online]. Available: https://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Tech. Rep. CNS-TR-2011-001, 2011.

[15] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018, doi: 10.1631/fitee.1700808.

[16] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019, doi: 10.1109/TNNLS.2018.2886017.

[17] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*.

[18] A. Joshi, A. Mukherjee, S. Sarkar, and C. Hegde, "Semantic adversarial attacks: Parametric transformations that fool deep classifiers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4772–4782, doi: 10.1109/ICCV.2019.00487.

[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep. 2009, pp. 365–372, doi: 10.1109/ICCV.2009.5459250.

[20] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 951–958, doi: 10.1109/CVPR.2009.5206594.

[21] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*.

[22] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2897–2905, Dec. 2018, doi: 10.1109/TPAMI.2017.2784440.

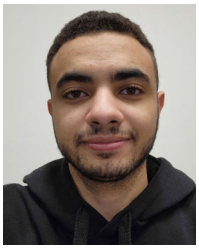
[23] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," 2015, *arXiv:1503.02406*.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[25] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4109–4118, doi: 10.1109/CVPR.2018.00432.

[26] L. Engstrom, *A Discussion of 'Adversarial Examples are not Bugs, They are Features*, vol. 4, no. 8. San Francisco, CA, USA: Distill, Aug. 2019, doi: 10.23915/distill.00019.
[27] H. Uzunova, J. Ehrhardt, T. Kepp, and H. Handels, "Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Mar. 2019, p. 36, doi: 10.1117/12.2511964.
[28] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. NIPS*, 2018, pp. 9505–9515. [Online]. Available: https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf
[29] M. T. Bahadori and D. E. Heckerman, "Debiasing concept-based explanations with causal analysis," 2020, *arXiv:2007.11500*.
[30] B. Smit and J. Wandel, "Adaptation, adaptive capacity and vulnerability," *Global Environ. Change*, vol. 16, no. 3, pp. 282–292, Aug. 2006, doi: 10.1016/j.gloenvcha.2006.03.008.
[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[32] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–39, Aug. 2023, doi: 10.1145/3547330.
[33] J. Selimkhanov, B. Taylor, J. Yao, A. Pilko, J. Albeck, A. Hoffmann, L. Tsimring, and R. Wollman, "Accurate information transmission through dynamic biochemical signaling networks," *Science*, vol. 346, no. 6215, pp. 1370–1373, Dec. 2014, doi: 10.1126/science.1254933.
[34] K. M. Collins, M. Barker, M. E. Zarlenga, N. Raman, U. Bhatt, M. Jamnik, I. Sucholutsky, A. Weller, and K. Dvijotham, "Human uncertainty in concept-based AI systems," 2023, *arXiv:2303.12872*.
[35] D. Steinmann, W. Stammer, F. Friedrich, and K. Kersting, "Learning to intervene on concept bottlenecks," 2023, *arXiv:2308.13453*.
[36] V. V. Ramaswamy, S. S. Y. Kim, N. Meister, R. Fong, and O. Russakovsky, "ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features," 2022, *arXiv:2206.07690*.
[37] B. Rasheed, A. Khan, S. M. A. Kazmi, R. Hussain, M. J. Piran, and D. Y. Suh, "Adversarial attacks on featureless deep learning malicious URLs detection," *Comput., Mater. Continua*, vol. 68, no. 1, pp. 921–939, 2021, doi: 10.32604/cmc.2021.015452.
[38] B. Rasheed, A. Khan, M. Ahmad, M. Mazzara, and S. M. A. Kazmi, "Multiple adversarial domains adaptation approach for mitigating adversarial attacks effects," *Int. Trans. Electr. Energy Syst.*, vol. 2022, pp. 1–11, Oct. 2022, doi: 10.1155/2022/2890761.
[39] B. Rasheed, A. M. Khattak, A. Khan, S. Protasov, and M. Ahmad, "Boosting adversarial training using robust selective data augmentation," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 89, May 2023, doi: 10.1007/s44196-023-00266-x.
[40] T. Hussain and H. Shouno, "Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping," *Information*, vol. 14, no. 12, p. 642, Nov. 2023, doi: 10.3390/info14120642.

BADER RASHEED received the master's degree from Bauman Moscow State Technical University and the Ph.D. degree in computer science from Innopolis University. He is currently the Head of the Recognition Systems Department, Innopolis University. He is also a Research Assistance at the College of Technological Innovation, Zayed University.



MOHAMED ABDELHAMID is currently pursuing the degree in computer science with Innopolis University, specializing in applied artificial intelligence. His academic journey started with the STEM Assuit High School under a scholarship from Egyptian Government and USAID, with a focus on STEM subjects and academic research. He received another scholarship to study computer science with Innopolis University, where he deepened his expertise in AI. His current research interests include the robustness of concept bottleneck models against adversarial attacks, marking his first contribution to AI safety research. His work aims to enhance the interpretability and security of AI systems, a crucial step toward reliable AI applications.

ADIL KHAN (Member, IEEE) is currently a seasoned Professor and a Prolific Researcher in machine learning. With a robust background in machine learning, deep learning, and representation learning, he is passionately dedicated to both pedagogy and innovative research in the realm of artificial intelligence. His research journey started in South Korea, in 2006, where he concentrated on human activity recognition through wearable sensors. His groundbreaking discoveries were published in reputable journals and employed by leading technology firms for their healthcare applications. Over his career, he has undertaken more than ten research projects, obtaining substantial funding, and has published in excess of 90 research articles. He has supervised more than two dozen M.S. and Ph.D. students to completion. His expertise and experience are not limited to a single geographic location. His academic career has spanned across various prestigious universities in South Korea, Denmark, Russia, United Arab Emirates, Switzerland, and U.K. These diverse collaborations and experiences have enriched his knowledge and cultural understanding, augmenting his holistic approach toward research, and education.

IGOR MENEZES received the Ph.D. degree from the University of Cambridge. He was a Research Associate with the Judge Business School. He is currently an Associate Professor of people analytics with Hull University Business School. With over fifteen years of experience, he has coordinated teams and laboratories as a Researcher and the Principal Investigator for various research grants and funded projects. Throughout his career, he has published more than 50 research articles and presented at numerous conferences. His current research interests include integrating AI and computer vision algorithms with psychometric techniques to develop enhanced solutions for individuals and businesses. He is an AI Psychometrician, a RSS Graduate Statistician (GradStat), a Chartered Psychologist (CPsychol), and an Associate Fellow (AFBPsS) of the British Psychological Society.

ASAD MASOOD KHATAK (Senior Member, IEEE) received the M.S. degree in information technology from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2008, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2012. He was a Postdoctoral Fellow and an Assistant Professor with the Department of Computer Engineering, Kyung Hee University. In August 2014, he joined the College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates. He is currently an Associate Professor with the College of Technological Innovation, Zayed University. He is leading three research projects, collaborating in four research projects, and has successfully completed five research projects in the fields of data curation, context-aware computing, the IoT, and secure computing. He has authored/co-authored more than 120 journals and conference papers in highly reputed venues. He has delivered keynote speeches, invited talks, guest lectures, and short courses in many universities. He served as a reviewer, a program committee member, and the guest editor for many conferences and journals. He and his team have secured several national and international awards in different competitions.