Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057866

ARTICLE

Exploratory Research on Defense against Natural Adversarial Examples in Image Classification

Yaoxuan Zhu, Hua Yang and Bin Zhu*

The State Key Laboratory of Pulsed Power Laser Technology, National University of Defense Technology, Hefei, 230037, China

*Corresponding Author: Bin Zhu. Email: zhubin@nudt.edu.cn

Received: 29 August 2024 Accepted: 13 November 2024 Published: 17 February 2025

ABSTRACT

The emergence of adversarial examples has revealed the inadequacies in the robustness of image classification models based on Convolutional Neural Networks (CNNs). Particularly in recent years, the discovery of natural adversarial examples has posed significant challenges, as traditional defense methods against adversarial attacks have proven to be largely ineffective against these natural adversarial examples. This paper explores defenses against these natural adversarial examples from three perspectives: adversarial examples, model architecture, and dataset. First, it employs Class Activation Mapping (CAM) to visualize how models classify natural adversarial examples, identifying several typical attack patterns. Next, various common CNN models are analyzed to evaluate their susceptibility to these attacks, revealing that different architectures exhibit varying defensive capabilities. The study finds that as the depth of a network increases, its defenses against natural adversarial examples strengthen. Lastly, Finally, the impact of dataset class distribution on the defense capability of models is examined, focusing on two aspects: the number of classes in the training set and the number of predicted classes. This study investigates how these factors influence the model's ability to defend against natural adversarial examples. Results indicate that reducing the number of training classes enhances the model's defense against natural adversarial examples. Additionally, under a fixed number of training classes, some CNN models show an optimal range of predicted classes for achieving the best defense performance against these adversarial examples.

KEYWORDS

Image classification; convolutional neural network; natural adversarial example; data set; defense against adversarial examples

1 Introduction

As a classic and effective network architecture in deep learning, Convolutional Neural Networks (CNNs) have become one of the fundamental building blocks of various novel deep learning networks. In 2013, Krizhevsky et al. [1] introduced the AlexNet model based on CNNs, achieving first place in the ImageNet competition, which drew attention to the unparalleled feature extraction capabilities of CNNs. Due to their outstanding feature extraction functionality, CNNs have found widespread application in the field of images, leading to the emergence of numerous models such as Visual

Geometry Group (VGG) [2], Inception [3], ResNet [4], DenseNet [5], and ConvNeXt [6]. Initially, CNNs were applied solely in the visible light image domain; however, in recent years, with the proliferation of deep learning hardware and software, CNNs have also played a significant role in infrared images [7] and Synthetic Aperture Radar (SAR) images [8].

However, as research has advanced, particularly with the advent of adversarial examples, an growing number of scholars have recognized the inherent shortcomings in the robustness of these end-to-end learning "black box" models. Szegedy et al. [9] introduced the concept of adversarial examples at 2014 International Conference on Learning Representations (ICLR), defining them as input samples formed by deliberately adding subtle perturbations to the dataset, which cause the model to produce erroneous outputs with high confidence after perturbation. The emergence of adversarial examples has sparked a research frenzy in the field of deep learning, focusing on both adversarial attacks and defenses. With the development of theoretical frameworks around deep learning and adversarial attacks, merely understanding the existence of adversarial examples in the digital realm is no longer sufficient. Many studies have begun to explore adversarial examples generated by physical objects in the real world. For instance, Lee et al. [10] demonstrated a physical adversarial patch attack targeting object detectors, particularly the YOLOv3 (You Only Look once, YOLO) detector, indicating that a well-designed patch can suppress nearly all detected objects within an image. Additionally, Duan et al. proposed an adversarial camouflage method, AdvCam [11], which conducts physical-world attacks by disguising target objects, resulting in adversarial examples that are more covert and less detectable.

Beyond artificially added adversarial patches, there also exist naturally occurring adversarial examples that are not human-made. Hendrycks et al. [12] argued that the current ImageNet test set is overly simplistic and does not represent the more complex images found in the real world, introducing the term "Natural Adversarial Examples" for the first time. Natural adversarial examples expose common blind spots in current convolutional networks, such as over-reliance on texture and excessive generalization, while being more aligned with real-world applications due to their origins in actual scenes. As a form of authentic, unmodified, and naturally occurring adversarial examples, natural adversarial examples inherently represent classification scenarios that models may encounter in real-world settings. Therefore, researching the defenses of image classification models against natural adversarial examples can not only lead to the development of robust network models for natural environments but also facilitate the practical deployment of models beyond laboratory settings, while simultaneously advancing theoretical research into the interpretability of deep learning.

This paper concentrates on convolutional neural network (CNN) image classification models. While transformer-based architectures have recently achieved notable success in this domain, the development of the ConvNeXt [6] model within CNNs indicates that CNNs continue to be the optimal choice for processing image signals. Consequently, our research exclusively focuses on CNNs. This paper focuses on natural adversarial examples and several common convolutional network classification models, conducting exploratory research on the defenses against natural adversarial examples from three perspectives: adversarial examples, models, and datasets. The primary research contributions are as follows:

(1) Conducting feature analysis on natural adversarial examples to uncover their attack mechanisms and study common attack patterns;

(2) Investigating the defensive effectiveness of different model architectures and varying sizes of models within the same architecture against natural adversarial example attacks;

(3) Examining the impact of the number of classes in the dataset on the model's defense effectiveness from the perspective of training and predicted class counts. The findings of this study lay the groundwork for subsequent research on natural adversarial examples and provide guidance for the practical deployment of image classification models in operational contexts, thereby contributing to the advancement of robustness and interpretability in deep learning.

2 Materials & Methods

2.1 Attack Patterns of Natural Adversarial Examples

2.1.1 Natural Adversarial Examples

In the study by Hendrycks et al. [12], a natural adversarial example test set for ImageNet classifiers, named ImageNet-A, was developed. This dataset includes 200 classes and 7500 natural adversarial examples. To construct these examples, 200 classes were initially selected from the 1000 classes of ImageNet-1k. A large number of images related to these categories were then downloaded from other databases. Images correctly classified by ResNet-50 were removed, and among the remaining misclassified samples, those with low confidence (<15%) were chosen to maximize classification errors for the selected classes. Each image was manually reviewed to ensure validity, singularity in category, and high quality. This process resembles difficult sample mining.

Existing convolutional network classifiers exhibit an overreliance on image color, texture, and background information, while underestimating the importance of the external shape and internal arrangement of objects themselves. Natural adversarial examples exploit this flaw within classifiers. Consequently, the natural adversarial examples may include images that lead to erroneous classifications due to factors such as uneven brightness distribution, object deformation, excessively small object size, blurriness, or occlusion of objects. Fig. 1 presents two typical natural adversarial examples, in which the squirrel and the dragonfly are misclassified by the convolutional network classifier with high confidence as a sea lion and a manhole cover, respectively.

Figure 1: Typical natural adversarial examples

Compared to the ImageNet dataset, the images in the natural adversarial example dataset, ImageNet-A, are richer in information, more complex, and closer to real-world scenarios. This poses a significant challenge for classification models. Experimental results indicate that the robustness of nearly all widely-used convolutional network classification models is reduced on ImageNet-A. Investigating the attack patterns of natural adversarial examples involves identifying and summarizing the fundamental attack mechanisms, with the goal of mitigating such attacks in practical applications.

2.1.2 Typical Attack Patterns

Image classification models based on deep convolutional networks operate similarly to a "black box" system. The input data is an image, and the output is the target class label. The processes of image data handling and feature learning are unknown, i.e., they function in an "end-to-end" manner. Class Activation Mapping (CAM) [13] technology can visualize the features of experimental results. Through heatmaps, it allows for intuitive comparisons, revealing internal information such as the activation results of the network model's intermediate layers, the types of features extracted by the network, and the critical image regions that affect the output. This facilitates further exploration of how various factors influence the operational mechanisms and decision-making criteria of convolutional neural networks.

Zhou et al. [13], through a series of experiments, demonstrated that the CAM technique can highlight the discriminative object parts detected by the convolutional neural network. However, the use of CAM relies on global average pooling of the model, necessitating modifications to the model architecture and retraining. To address this issue, Selvaraju et al. [14] proposed the Grad-CAM technique, which can visualize convolutional neural networks of any structure without modifying the network architecture or retraining. The core idea of Grad-CAM is to use the weights of the target feature map to express gradients and, through guided backpropagation, remove gradients less than zero, retaining feature regions that positively impact classification. Its implementation process is shown in Fig. 2.

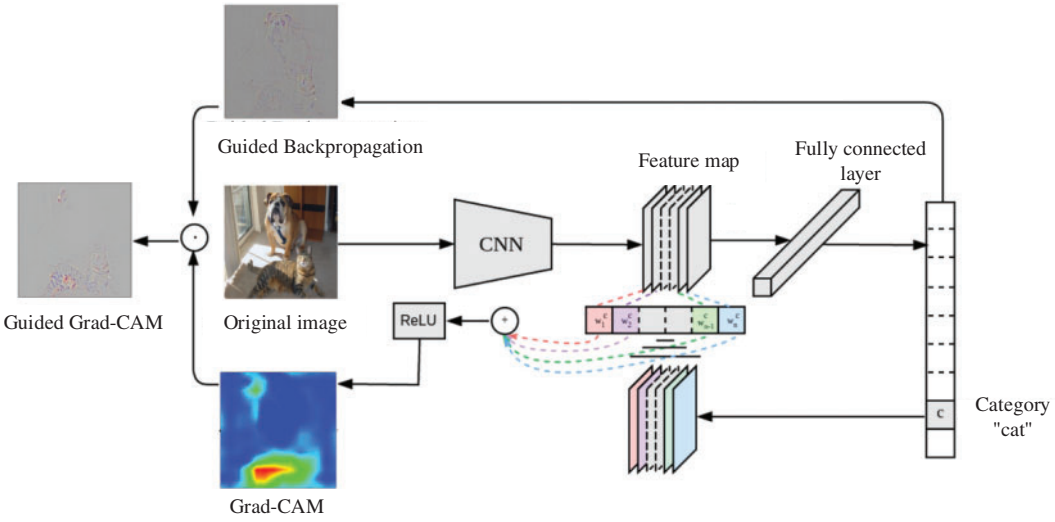

Figure 2: Grad-CAM flowchart

Grad-CAM generates a heatmap the same size as the original image. On this image, each pixel's value ranges from 0 to 255, indicating the contribution distribution of each pixel to the class prediction output. Brighter areas signify higher network responses and greater contributions from the corresponding regions of the original image. The theoretical calculation process of Grad-CAM is as follows:

(1) Calculate the partial derivative of the class probability y^c with respect to A_{ij}^k:

$$\frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

where y is the probability vector output by the Softmax layer, c represents the class index, A_{ij}^{k} is the object feature map, k is the channel dimension index of the feature map, i and j represent the pixel width and height, respectively.

(2) After calculating the partial derivatives of y^c for all pixels in the feature map, take the global average feature weight:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

Here, w_k^c represents the sensitivity of the current class to the k-th channel of the feature map output by the last convolutional layer.

(3) Use w_k^c as the weight to perform a weighted linear combination of the final layer feature maps, processed by the ReLU (Rectified Linear Unit) activation function to get the final value:

$$L_{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_k w_k^c A^k \right) \tag{3}$$

where ReLU is used to filter out the heatmap distribution positively correlated with the class probability score c. To obtain a heatmap of the same size as the input image, a size transformation operation is performed on $L_{Grad-CAM}^c$.

We employed the latest ConvNext convolutional network as the classification model and introduced Grad-CAM technology to visualize the image classification process. By observing the heatmaps (Grad-CAM, Guided Backprop, Guided Grad-CAM), we analyzed how natural adversarial examples cause the network model's classification performance to deteriorate and categorized these natural adversarial examples to summarize patterns. The Grad-CAM heatmap visually shows the classification basis of the neural network classification model, the Guided Backpropagation heatmap represents the network model's guided backpropagation visualization, and the Guided Grad-CAM heatmap highlights the areas affecting the final classification.

Through experiments, we analyzed and identified seven typical attack patterns of natural adversarial examples. The classification failure modes of convolutional neural network models regarding these examples, from another perspective, represent the attack patterns of natural adversarial examples.

Image Distortion

In academic research on image acquisition and processing, image distortion is a significant factor that affects image quality and recognition performance. Particularly in dynamic scenes, the relative rapid movement between the target object and the camera, as shown in Fig. 3a, causes image blur distortion. This phenomenon is primarily due to the motion blur effect, where the movement of the object during the exposure time results in blurred contours in the image. Additionally, when the shooting equipment adjusts the focal length and the light fails to focus accurately, focal length diffusion occurs, as illustrated in Fig. 3b, leading to image distortion. These distortion phenomena not only reduce the sharpness of the image but also may obscure critical features of the object, posing challenges for image recognition and analysis. Although these distortions might not severely impact the human eye's perception and recognition of the foreground objects in the image when they are slight, they make distorted images more challenging for CNN-based image classification models to classify correctly compared to regular images.

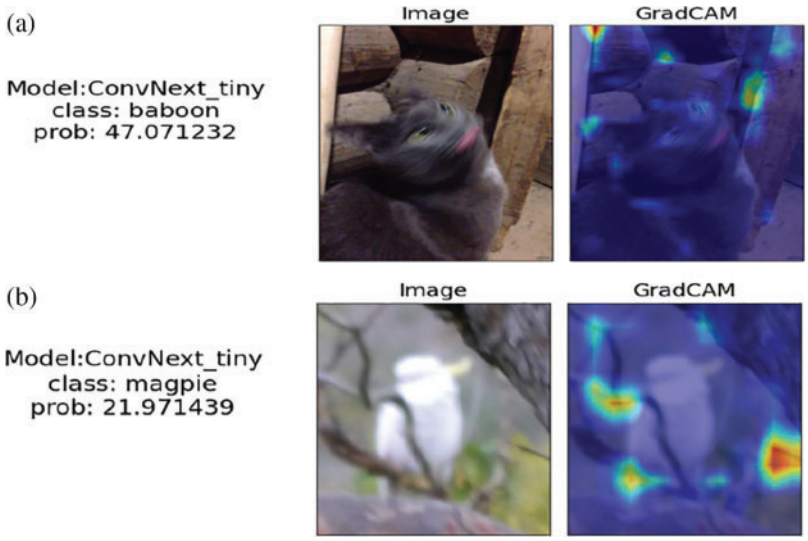

Figure 3: Image distortion attack pattern

The attack patterns of image distortion indicate that natural adversarial examples are more diverse, realistic, and deceptive compared to artificially created adversarial patches and other adversarial samples. These examples pose a significant threat to the credibility and robustness of recognition models. Therefore, comprehending and mitigating the impact of image distortion is essential for enhancing the reliability and robustness of these models.

Low Object-to-Background Ratio

In the image scene shown in Fig. 4, the target object occupies an extremely small proportion of the image area. In contrast, most of the image space is occupied by the background or irrelevant objects, presenting a significant challenge in the task of image classification. A detailed observation of the image reveals a high information density with numerous visual elements. In such a complex image background, the proportion of the target object in the overall image is minimal, posing a challenge not only to human visual recognition but also to the recognition capability of image classification models.

Figure 4: Low object-to-background ratio attack pattern

Heatmap analysis further confirms this phenomenon. Even when the target object is clear and its features are distinct, if its proportion in the image is too low, the model's attention is not focused

on the target object but rather dispersed onto other more prominent features in the image, such as the haystacks and utility poles shown in Fig. 4. These objects, due to their large spatial occupation and relatively salient features, become the primary focus of the model's detection. This phenomenon highlights the limitations of image classification models in handling complex images: when the object-to-background ratio is too low, the model's attention and the importance assigned to the target object decrease, leading to difficulties in accurately recognizing and classifying these small targets.

Grid Occlusion

When a target object is positioned behind a grid occlusion, as shown in Fig. 5 where a chick is divided by the bars of a birdcage or a pet dog is confined within an iron cage, the target object is segmented into several parts by the physical structure. Although humans can easily recognize these occluded objects through experience and intuition, for classification models that rely on visual features, the recognition process becomes complex and difficult. The model struggles to perceive the grid-occluded target as a whole entity, primarily because the grid occlusion disrupts the continuity and integrity of the target object, making it challenging for the model to capture complete object features.

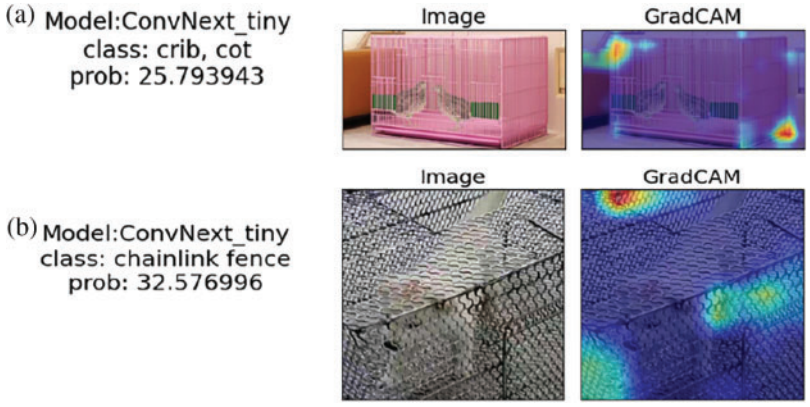

Figure 5: Grid occlusion attack pattern

Further analysis of the interference caused by grid occlusion in the model's recognition process can be explained from the perspective of the model's attention mechanism. The heatmap in Fig. 5 clearly reveals this phenomenon: when processing images of objects occluded by grids, the classification model's attention is significantly drawn to the features of the grid-like objects rather than to the actual features of the target in the image. This indicates that during feature extraction and classification decision-making, the model excessively focuses on the details of the grid occlusion and neglects the critical features of the occluded target object, which results in a significant decline in classification performance.

Background Interference

When the color and texture information of the target object is similar to that of the background, the object may be perceived as part of the background and fail to be classified successfully. In this scenario, there are two main reasons for the misclassification: first, the target object merges into the background, making it difficult for the model to locate the target; second, even if the target object is located, the classifier may incorrectly classify the target due to the highly similar information between the background and the target object. As shown in Fig. 6, a lynx was ignored by the classification model due to its color and texture being similar to the tree trunks and soil in the background. This is a

clear manifestation of the image classification model's excessive reliance on image color, texture, and background information.

Figure 6: Background interference attack pattern

An additional aspect of background interference in classification is the association between the background and the foreground object. Xiao et al. [15] demonstrated through experiments that even without the foreground, models can still make reasonable classifications based solely on the background. In deep learning-based image classification, models sometimes rely on associative information rather than the objects themselves, leading to misclassification. During training, certain objects frequently appear with specific backgrounds, such as skis with snow or flowers with green leaves. This causes the classifier to associate targets with these backgrounds. During inference, the model's over-reliance on background cues can disrupt classification results, leading to errors.

Deep Background Objects

When recognizing objects in images, the relative position of objects within the image and their relationship to the depth of field have a crucial impact on the performance of recognition systems. This is especially challenging when the target object is located at a deeper depth of field and there are other distracting objects in the foreground. The challenge mainly arises from the focusing technology of the capturing device and the image acquisition conditions, which include but are not limited to failure to precisely focus on the target object or setting the focus range too broadly, resulting in multiple objects being clearly presented in the image. As shown in Fig. 7, the frog, which is the target object to be recognized, is located at a deeper depth of field in the image. Although it is partially obscured by a lotus leaf, its key morphological features are not entirely concealed. However, due to the focusing technology during capture, the lotus leaf is also included within the clear focus range. As a result, the contrast and clarity of the lotus leaf in the image are higher, and the frog, located at the deeper depth of field, appears weaker in the image. Therefore, the lotus leaf is more likely to become the primary focus of the classification model.

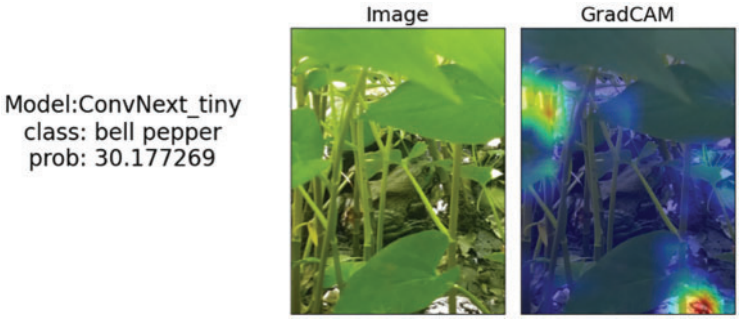

Figure 7: Deep background objects attack pattern

This pattern fundamentally differs from the background interference pattern. In the background interference pattern, the target object is typically positioned in front of the distracting objects and may be partially or completely obscured, making its features blurred or difficult to recognize in the image. In the deep background object attack pattern, the target object is primarily located behind the distracting objects, and its main features are not completely obscured.

Significant Edge Interference

The edge information of objects and regions with abrupt changes in color and brightness play a significant role in the attention distribution of models. When there are objects with clearer edge information in the image, such as fences, rooftops, roads, or regions with significant changes in color and brightness, these objects and regions are more likely to attract the attention of the image classification model due to their strong visual contrast and distinct contour features. This phenomenon is driven by the model's sensitivity and preference for prominent features in the image. As shown in Fig. 8, when there are structures with prominent edge information, such as eaves, fences, or cave outlines around the target object, the model may ignore or misclassify the target object despite its own distinctive features because of the strong contrast from the surrounding objects.

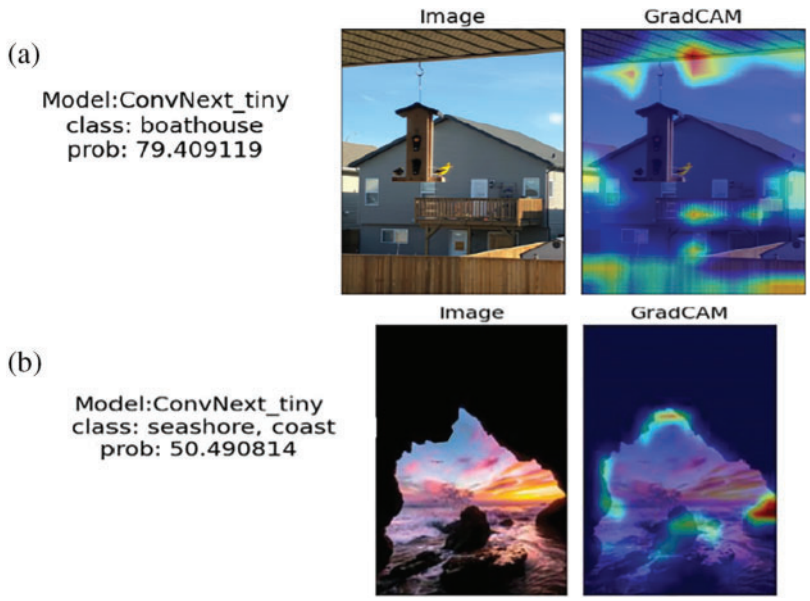

Figure 8: Significant edge interference objects attack pattern

This type of attack pattern conceptually overlaps with background information interference, but its core distinction lies in the nature and mechanism of the interference source. Background information interference typically refers to the disturbance caused by irrelevant background areas in the image on target detection. In contrast, the pattern discussed here involves interference objects that are not necessarily part of the background but actively attract the model's attention due to their prominent edge and gradient information, indirectly affecting the recognition of the target object. These interference sources might be located around the target object, becoming the primary focus of the model's recognition process through their clear edges or strong color and brightness contrasts, thus weakening the features of the target object in the model's decision-making process.

To further illustrate this attack pattern, a computer-generated image was created as shown in Fig. 9, where a gradient from white to black is juxtaposed. The Guided Backprop heatmap in the image represents the network model's guided backpropagation visualization, while the Guided Grad-CAM heatmap represents the areas of the network model that influence the final classification. From the last three images, it is evident that the model strongly focuses on the boundary region.

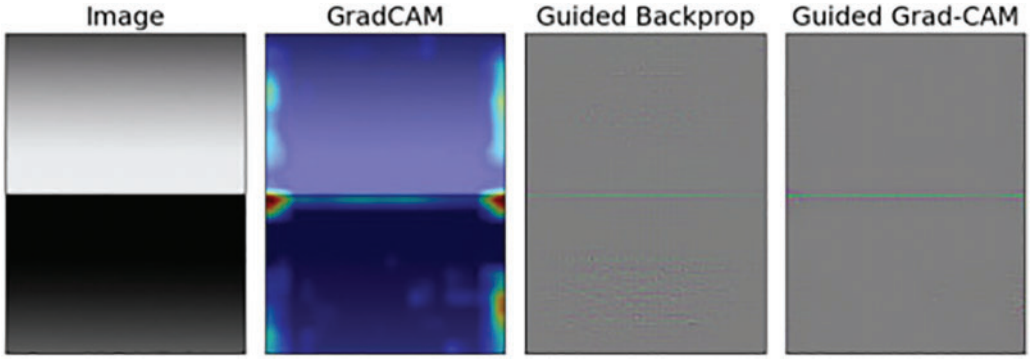

Figure 9: Interference effect of significant edge information

Environmental Factors

During the image acquisition stage, the complexity of external natural environmental factors poses significant challenges to subsequent image classification tasks. These factors primarily include lighting conditions, weather conditions, and shooting angles, which together profoundly impact the final image quality and the performance of the classification model.

First, lighting conditions are a key variable that affects image quality. Excessive lighting can lead to overexposure of the target objects in the image, potentially causing the loss of detailed features, thereby reducing the classification model's recognition accuracy. Conversely, when lighting is insufficient, the contrast between the object and the background diminishes, making the object contours blurry, which similarly increases the difficulty of classification. Secondly, the diversity of weather conditions introduces additional obstacles to image acquisition. For instance, adverse weather conditions such as rain, snow, and fog not only reduce image clarity, resulting in blurriness or partial occlusion of the target objects, but may also introduce additional noise, interfering with the correct judgment of the classification model. Finally, the choice of shooting angle should not be overlooked. Even the same object can appear significantly different when captured from different angles, presenting additional challenges for the model's detection and recognition. Specifically, certain special angles, such as top-down, bottom-up, or side views, may cause some critical features of the object to be obscured or distorted, thereby affecting the model's classification accuracy.

2.2 Selection of Classification Models

Compared to traditional machine learning algorithms, CNN algorithms based on deep learning encompass a more diverse range of models. Each model within this category has subtle differences in its network architecture, and it is precisely these differences that distinguish them from conventional machine learning algorithms. These variations enable the derivation and iteration of various network models, allowing them to leverage advantages that traditional machine learning algorithms do not possess when addressing a wide array of practical problems. In the study conducted by Lee et al. [10], the network model employed in their experiments was primarily ResNet. However, there are numerous

existing network models, and in recent years, frameworks with excellent classification performance, such as EfficientNet and ConvNeXt, have emerged. Therefore, in our research, we selected these popular convolutional networks for comparative study to analyze the differences in the effectiveness of various model structures in defending against natural adversarial examples. We will not only consider the differences in defensive capabilities between image classification models with different architectures but also the differences in defensive capabilities among models with varying numbers of network layers within the same architecture.

While transformer-based architectures have recently achieved notable success in this domain, the development of the ConvNeXt model within CNNs suggests that CNNs remain the optimal choice for processing image signals. Consequently, our research focuses exclusively on CNNs.

2.3 Configuration of Dataset Scale

In previous studies, the network models used pre-trained weights to predict natural adversarial examples, without addressing the size of the training and testing datasets. In the field of deep learning, dataset work has always been a focal point. Therefore, in this section, the research direction will shift to examining the impact of datasets on the defense performance of models.

2.3.1 Impact of the Number of Training Set Categories

In 2017, Sun et al. [16] conducted experiments using the JFT-300M dataset to verify the impact of dataset expansion on network accuracy. The experiments demonstrated that large-scale data enhances representation learning, thereby optimizing the performance of all studied visual tasks. Additionally, as the magnitude of training data continues to increase, task performance improves logarithmically. Even when the scale of training images reached 300 million, no performance plateau was observed. Based on these experimental observations, if data volume is unlimited and precisely annotated, and assuming the model has sufficient complexity, deep learning models can fit functions that achieve accurate classification. In other words, the more data and the more complex the model, the better the predictive performance. However, models can easily give incorrect results with high confidence when identifying natural adversarial examples, indicating that the model may have extracted deep features corresponding to other categories in these adversarial samples. It is reasonable to believe that the more object categories a model learns, the more prone it is to such errors. Therefore, this subsection will shift the research focus to reducing the number of categories in the training dataset. By selecting a dataset corresponding to 200 classes from ImageNet-1k that are associated with natural adversarial examples, we will train network models to verify whether the number of dataset categories affects the model's performance on natural adversarial examples.

The basic experimental approach in this section is as follows: After altering the number of categories in the training set, train the model and test its classification performance on natural adversarial examples to examine how different numbers of training set categories impact the model's classification performance. The basic steps are:

a) Randomly select N categories from the 200 classes in the ImageNet-A dataset;
b) Extract all images (including natural adversarial and non-adversarial images) of the N categories from the ImageNet-1k dataset to construct a new dataset, ImageNet-N;
c) Train a ConvNeXt model using the ImageNet-N dataset, and denote the trained model as ConvNeXt (N);
d) Test the prediction accuracy of ConvNeXt (N) on the selected N categories of natural adversarial examples;

e) Repeat the random selection of N categories three times, and repeat Steps (a) through (d) to complete Experiments 1 to 3;
f) Sequentially choose $N = \{20, 80, 140, 200\}$ and repeat Steps (a) through (e).

The pseudocode for implementing the above steps is shown in Algorithm 1.

Algorithm 1: Impact of the number of training set categories

Input: ImageNet-A dataset, ImageNet-1k dataset, N values {20, 80, 140, 200}
Output: Prediction accuracy
1: **for** each N in {20, 80, 140, 200} **do**
2: **for** experiment = 1 to 3 **do**
3: selected_categories = RANDOM_SELECT(N, 200)
4: ImageNet_N = EXTRACT_IMAGES(selected_categories, ImageNet_1k)
5: ConvNeXt_N = TRAIN_MODEL(ImageNet_N)
6: Prediction accuracy = TEST(ConvNeXt_N, selected_categories, ImageNet-A)
7: **end for**
8: **end for**
9: **return** Prediction accuracy

2.3.2 Impact of the Number of Test Set Categories

During the model training phase, the number of categories in the training set leads to different model weights, which in turn affects the model's test accuracy. Similarly, during the model inference phase, the number of predicted categories also impacts the classification accuracy.

By analyzing the experimental methods of Lee et al. [10], it is evident that when using a network model to predict the categories of natural adversarial examples, they manually selected 200 categories from the initial 1000 category probability values predicted by the model that corresponded to the natural adversarial examples. Only the predicted classifications belonging to these 200 categories were considered valid, while the remaining 800 categories from ImageNet-1k were discarded as invalid results. The final classification category was determined by selecting the maximum value among the probabilities of these 200 categories. The model's initial 1000 category probability values were obtained using the pre-trained weights of ImageNet-1k, and then 200 categories corresponding to the natural adversarial examples were extracted. Further, some categories were selected from these 200 categories to obtain the corresponding classification accuracy.

The basic idea of the experiment in this section is as follows: using the ImageNet-1k dataset, which contains all 1000 image categories, to train the model, and during the testing phase, selecting M categories of natural adversarial examples from the 200 categories in ImageNet-A to test the trained model, thereby examining the impact of different numbers of test set categories on the model's classification performance. The basic steps are:

a) Train the VGG16 (Visual Geometry Group 16), ResNet101 (Residual Neural Network 101), and ConvNeXt classification models using the ImageNet-1k dataset, with the resulting models denoted as VGG16(1000), ResNet101(1000), and ConvNeXt(1000).
b) Randomly select M categories of adversarial sample images from the 200 categories in the ImageNet-A dataset to form a new natural adversarial example test set, ImageNet-A-M.
c) Test the models VGG16(1000), ResNet101(1000), and ConvNeXt(1000) using Image Net-A-M.

d) Repeat the random selection of M categories three times (resulting in three test sets: ImageNet-A-M_1, ImageNet-A-M_2, ImageNet-A-M_3 and repeat Steps (a)–(c) to complete experiments one through three.
e) Sequentially choose M ={20, 40, 60, 80, 100, 120, 140, 160, 180, 200} and repeat Steps (a)–(d).

The pseudocode for implementing the above steps is shown in Algorithm 2.

Algorithm 2: Impact of the number of training set categories

Input: ImageNet-A dataset, ImageNet-1k dataset, M values {20, 40, 60, 80, 100, 120, 140, 160, 180, 200}
Output: Prediction accuracy

```
Define models = ["VGG16", "ResNet101", "ConvNeXt"]
Define category_counts = {20, 40, 60, 80, 100, 120, 140, 160, 180, 200}
for each model in models do
    model_1000 = TRAIN_MODEL(model, ImageNet_1k)
end for
for each M in category_counts do
    for experiment = 1 to 3 do
        ImageNet_A_M = RANDOM_SELECT(M, 200)
        for each model in models do
            Prediction accuracy = TEST_MODEL(model_1000, ImageNet_A_M)
        end for
    end for
end for
return accuracy
```

3 Results

3.1 Impact of Classification Models on Defense Capability

3.1.1 Network Architecture

In the experimental section of this study, we meticulously selected a series of advanced deep learning models, including VGG16, ResNet101, EfficientNet_b0, MobileNet_v3, and ConvNeXt, to comprehensively evaluating their classification performance on natural adversarial examples. To ensure the comprehensiveness and objectivity of the evaluation, we adopted three key performance metrics: Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and Area Under the Precision-Recall Curve, known as Average Precision (AP). Together, these metrics depict the robustness and accuracy of the models when handling natural adversarial examples.

The ROC (Receiver Operating Characteristic) curve, with its unique coordinate system of False Positive Rate (FPR) and True Positive Rate (TPR), visually demonstrates the performance variations of the models across different thresholds. The AUC value, which quantifies the area under the ROC curve, provides a comprehensive measure of model performance, particularly important when dealing with imbalanced datasets. On the other hand, the PR (Precision-Recall) curve offers in-depth insights into model performance within specific application domains through the comparison of recall and precision, while the AP value further reinforces this insight by accurately measuring the average performance of the model based on the area under the PR curve.

The experimental results, as shown in Table 1, reveal the classification performance of each model on the ImageNet-1k and ImageNet-A datasets. Notably, all models exhibited a significant decline in classification performance when dealing with natural adversarial examples, highlighting the challenges posed by these samples to deep learning models and the necessity of considering adversarial robustness in model design. Among the models compared, ConvNeXt stood out for its exceptional performance. Compared to other models, ConvNeXt not only achieved higher accuracy on the ImageNet-1k dataset but also demonstrated significant advantages in classifying natural adversarial examples. This model was inspired by ResNet50 but incorporated a series of innovations, including macro design optimizations, depthwise separable convolutions, the introduction of inverted bottleneck layers, the use of large convolutional kernels, and design details borrowed from the Swin Transformer, resulting in a breakthrough in performance.

Table 1: Index of different architecture network models

Model	Acc	AUC	AP
VGG16	0.0264	0.834	0.021
ResNet	0.0473	0.852	0.029
EfficientNet	0.0708	0.851	0.041
MobileNet	0.0261	0.819	0.019
ConvNeXt	0.2347	0.901	0.154

The core innovation of ConvNeXt lies in its application of grouped convolutions, a strategy that allows each convolutional kernel to focus on processing individual channels while mixing information only in the spatial dimensions. This approach achieves performance comparable to self-attention mechanisms while maintaining computational efficiency. The incorporation of self-attention mechanisms enables the model to learn the relationships between individual pixel points and those at other locations, including long-range dependencies. This characteristic endows the ConvNeXt model with stronger defensive capabilities and finer feature capturing ability when handling natural adversarial examples, significantly enhancing the model's performance in the face of such challenges.

3.1.2 Parameter Scale

In both academia and industry, the number of parameters in a model has long been regarded as a key indicator of model complexity. It not only reflects the computational requirements of the model but is also closely related to the model's ability to fit the training data. Traditional machine learning theory [17] posits that as model complexity increases, its capacity to fit the training data improves; however, this also heightens the risk of overfitting, where the model performs excellently on the training set but exhibits a sharp decline in performance on unseen test data. However, when we shift our focus to the specific domain of natural adversarial examples, this classical rule seems to lose its direct applicability.

Through a detailed analysis of the classification accuracy of a series of deep learning models with different architectures on natural adversarial examples, our research reveals an intriguing phenomenon: the relationship between model parameter count and classification accuracy is far more complex than initially anticipated. Specifically, we calculated the number of parameters for several models, including EfficientNet, ConvNeXt, ResNet, and VGG, and compared their performance on natural adversarial examples, as shown in Fig. 10.

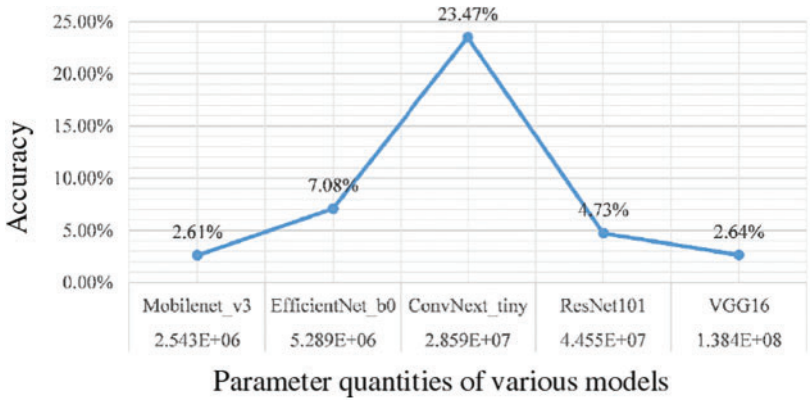

Figure 10: Relationship between models with different parameters and accuracy

The EfficientNet model stands out with its streamlined parameter count; although its classification accuracy on natural adversarial examples is slightly lower than that of the ConvNeXt model, it still outperforms the more parameter-heavy ResNet and VGG models. This finding suggests that the relationship between parameter count and classification accuracy on natural adversarial examples is neither linear nor monotonic, but is profoundly influenced by the design of the model architecture.

To further explore this phenomenon, we focused on models of the same type—keeping the model architecture constant while varying the parameter count—to observe its impact on classification accuracy for natural adversarial examples. Through carefully designed experiments, we obtained comparative results shown in Fig. 11. These results indicate that within the same category of models, an increase in parameter count does not necessarily correlate with an improvement in classification accuracy, further confirming the significant role of model architecture in handling natural adversarial examples.

This series of findings holds important implications for the design of deep learning models and the development of defense strategies against natural adversarial examples. It emphasizes the importance of model architecture design rather than merely pursuing an accumulation of parameters. Future research should delve deeper into how different model architectures influence feature learning from data, and how optimizing model architecture can enhance the generalization ability and robustness of models when dealing with natural adversarial examples. Exploring this field not only enriches our understanding of the essence of deep learning models but may also provide new insights for developing more efficient and secure machine learning systems.

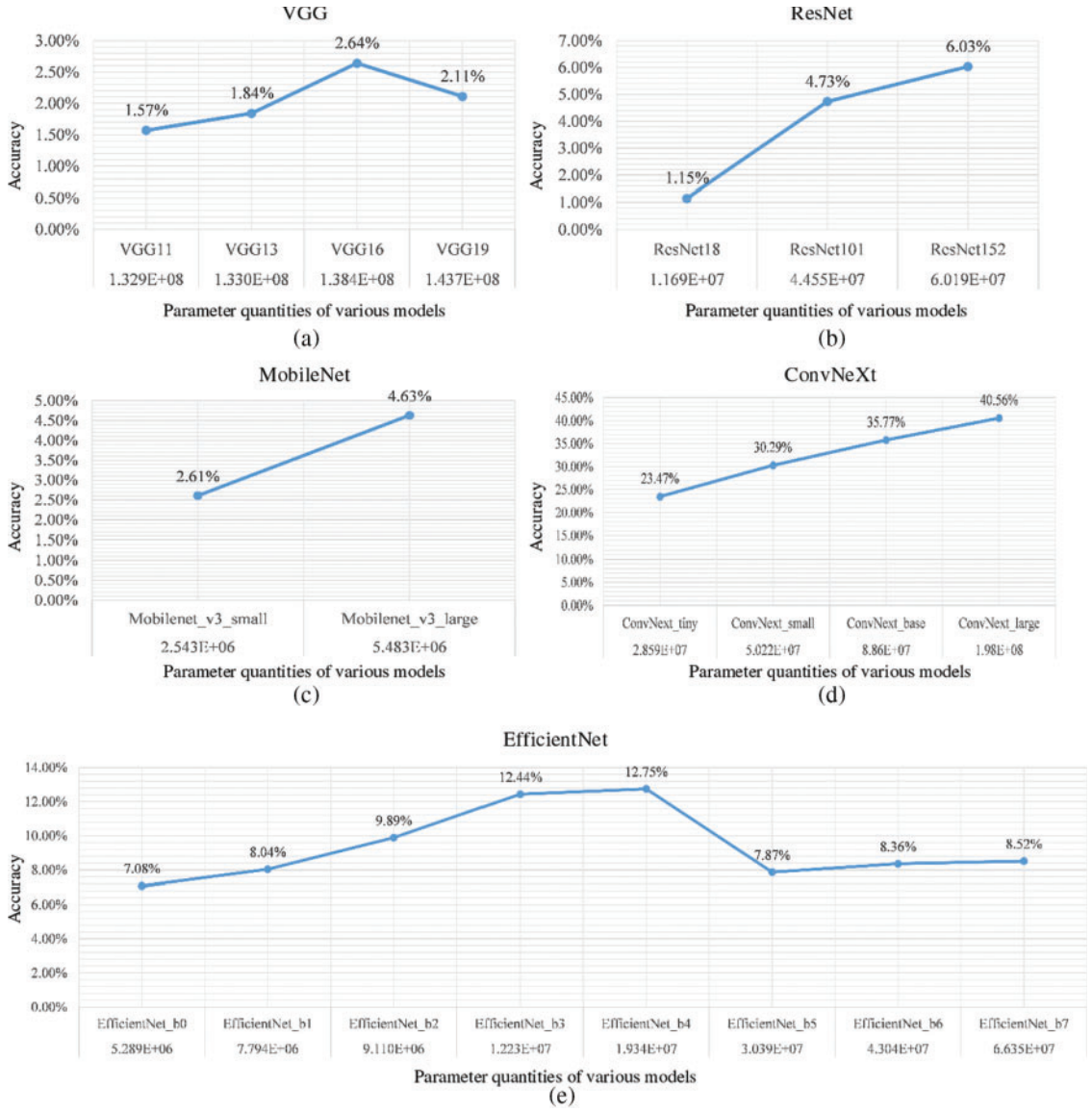

Figure 11: The relationship between parameter quantity and accuracy under the same model framework

3.2 Impact of Datasets on Defense Capability

3.2.1 Number of Categories in the Training Set

To investigate whether the number of training dataset samples affects classification performance, we selected 20, 80, and 140 samples from the 200 classes corresponding to natural adversarial examples. The ConvNeXt model, which demonstrated the best defense performance in the previous section, was chosen for this experiment. Each sample size underwent three trials to ensure the reliability of the results. Classification accuracy on natural adversarial examples was used as the evaluation metric, with the experimental data presented in Table 2.

Table 2: Defensive effectiveness under different numbers of training set categories (%)

Number of categories	20	80	140	200	1000
1st test	43.64	23.44	16.06	9.15	23.47
2nd test	44.08	20.63	15.82	9.73	
3rd test	45.19	24.49	15.97	11.43	
Mean value	44.19	22.85	15.95	11.43	

Comparative analysis of the experimental data reveals that, for the same number of training dataset samples, the classification accuracy of the model on natural adversarial examples does not vary significantly. This is partly because random preprocessing is applied to the data samples during the experiments, and partly because the distribution of adversarial samples across different categories is uneven, leading to varying numbers of adversarial samples in the validation set. Consequently, there are differences in accuracy across different trials, but these differences remain within the allowable error margin.

The experimental results indicate that in classifying natural adversarial examples, the more concentrated the training dataset, the better the model's classification performance. Theoretically, under complete training conditions, the results for 200 classes and 1000 classes should be relatively close, yet there is a significant discrepancy between these datasets. Examination of the training logs for the 200 classes reveals that the training accuracy is still increasing, and the loss function continues to decrease. This suggests that ConvNeXt has not been fully trained on the 200-class dataset. Considering the complexity of the ConvNeXt model, it is analyzed that insufficient epochs have led the model to learn only the shallow features of the images, failing to learn the deep features that determine the final predicted category of the image. As a result, the model has not captured the correct data features and is unable to classify natural adversarial examples effectively.

3.2.2 Number of Categories in the Test Set

To investigate whether the number of test set categories affects classification performance, we selected 20, 40, 60, 80, 100, 120, 140, 160, and 180 subcategories from the 200 classes corresponding to natural adversarial examples and determined the classification results within these subcategory sets. During the experiment, the VGG16, ResNet101, and ConvNeXt models were chosen, and three random selections were performed for each category count to ensure the reliability of the results. Classification accuracy on natural adversarial examples was used as the evaluation metric, with the experimental data shown in Tables 3–5.

Table 3: The defensive effectiveness of VGG16 under different numbers of predicted categories (%)

Number of categories	20	40	60	80	100	120	140	160	180	200
1st test	2.96	4.01	3.95	3.48	3.77	3.76	3.59	3.28	2.89	2.64
2nd test	2.67	3.52	3.44	2.87	3.31	4.16	3.31	3.49	2.95	

(Continued)

Table 3 (continued)

Number of categories	20	40	60	80	100	120	140	160	180	200
3rd test	2.97	2.96	3.80	3.93	3.83	3.32	3.68	3.59	2.99	
Mean value	2.87	3.50	3.73	3.42	3.64	3.75	3.53	3.54	2.94	

Table 4: The defensive effectiveness of ResNet101 under different numbers of predicted categories (%)

Number of categories	20	40	60	80	100	120	140	160	180	200
1st test	4.29	4.81	5.35	5.51	5.84	6.36	5.81	5.31	4.96	4.73
2nd test	3.53	5.20	5.97	6.36	6.40	4.83	5.19	5.36	5.20	
3rd test	3.72	5.60	5.64	5.99	5.67	5.13	5.45	5.32	5.01	
Mean value	3.85	5.20	5.65	5.95	5.97	5.44	5.48	5.33	5.06	

Table 5: The defensive effectiveness of ConvNeXt under different numbers of predicted categories (%)

Number of categories	20	40	60	80	100	120	140	160	180	200
1st test	6.93	8.51	14.36	14.67	17.41	19.01	20.53	21.36	23.17	23.47
2nd test	4.52	8.36	14.33	16.71	15.21	18.93	21.03	21.93	22.15	
3rd test	7.57	7.01	13.69	16.63	17.31	17.92	20.05	21.23	22.27	
Mean value	6.34	7.96	14.13	16.00	16.64	18.62	20.54	21.51	22.53	

The mean data from the above tables are represented as a curve, as shown in Fig. 12, where the horizontal axis represents the number of selected categories and the vertical axis represents classification accuracy (%). As shown in Fig. 12, for VGG16 and ResNet101, the prediction accuracy initially increases with the number of selected categories (reflected in the ascending part of the curve on the left side). However, after the number of categories increases beyond a certain value, the prediction accuracy begins to decrease with further increases in the number of categories (reflected in the descending part of the curve on the right side). In other words, there is an optimal number of selected categories for the VGG16 and ResNet101 models, corresponding to the best classification performance on natural adversarial examples. On the other hand, the prediction accuracy of the ConvNeXt model monotonically increases with the number of selected categories, and no optimal number of selected categories can be observed. It is also possible that the optimal number of selected categories for the ConvNeXt model lies outside the 200-class range of the ImageNet-A dataset.

This analysis provides guidance for applying classification models. Before deployment, a classification model should be trained on target data from all categories. After deployment, the number of predicted categories can be adjusted to an optimal level based on the model's characteristics and task requirements to achieve optimal performance. For the latest ConvNeXt model, retaining the maximum number of predicted categories is advisable. For instance, a classifier trained to recognize 1000 classes may have unlikely categories removed in practical applications, retaining only those most likely to be encountered.

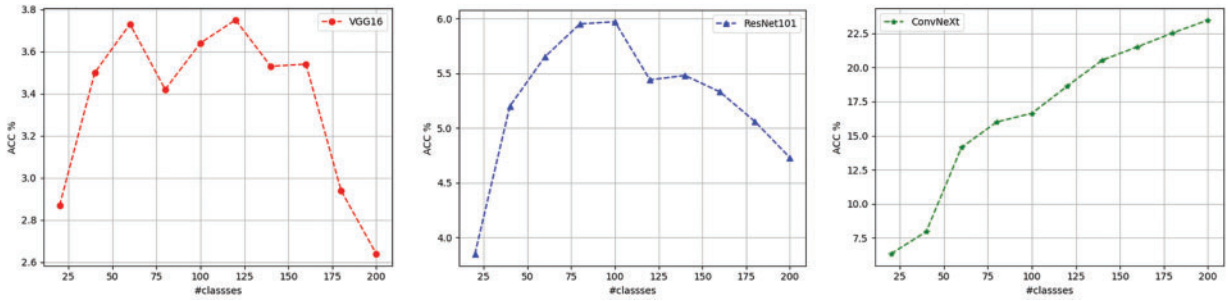

Figure 12: The variation curve of classification accuracy with respect to the number of predicted categories for typical CNN classification models

4 Discussion

The research conducted by Su et al. [18] found that the adversarial robustness and accuracy of image classification models cannot be reconciled. Although his study was focused solely on digital adversarial samples, it inspired the necessity to investigate the impact of datasets on model defense in the context of natural adversarial example attacks.

Focusing on defense strategies against natural sample attacks, we carried out in-depth and systematic academic research, primarily from three core dimensions: the analysis of adversarial sample characteristics, optimization and upgrading of model structures, and the comprehensiveness and diversity of datasets.

4.1 Adversarial Samples

Based on Grad-CAM technology, we identified and summarized seven typical patterns of natural adversarial example attacks on CNN classifiers. These patterns are all evident from the perspective of human visual understanding. Considering that the natural adversarial example dataset includes 200 categories and 7500 samples, these seven patterns should possess a certain degree of universality and have promotional value. It is important to emphasize that a specific attack may contain various attack patterns; therefore, CNN classifiers face a high risk of natural adversarial attacks in practical image classification applications.

4.2 Model Structure

In exploring model structures, we particularly focused on the development history of CNNs. Through comparative analyses of CNN classification models proposed from early to modern stages, we revealed the critical role of model optimization in enhancing defense capabilities. Specifically, experimental results clearly indicate that as CNN models evolve, their classification performance on the standard ImageNet-1k dataset significantly improves. Moreover, this enhancement positively impacts the processing of the ImageNet-A dataset of adversarial samples. This strongly demonstrates that optimizing and enhancing the classification capabilities of CNN models not only improves performance on conventional datasets but also serves as a robust means to effectively resist natural adversarial example attacks. This finding provides new perspectives for model designers and has significant practical implications for building more robust visual recognition systems.

4.3 Dataset

In our exploration of the dataset dimension, we focused on analyzing the influence of the number of categories in the training set and the test set on the model's ability to defend against natural adversarial example attacks.

In the experiments, we observed that the prediction accuracy of CNN models when facing natural adversarial examples is affected by the number of predicted categories, exhibiting an optimal number of extracted categories that allows the model's defense performance to reach its peak. Specifically, for the VGG16, ResNet101, and ConvNeXt models, their optimal extracted category numbers were 40, 120, and 200, respectively. This conclusion emphasizes that in practical applications, reasonable adjustments to the number of predicted categories can significantly enhance the model's defense capabilities against natural adversarial examples.

For instance, in actual military applications, considering the specific requirements of reconnaissance and identification tasks and the collected battlefield intelligence, we proposed a strategy: by excluding task-irrelevant categories, identifying the remaining categories in the dataset library. This process equates to selectively extracting certain categories from the initial training set. Experimental results confirmed that when the number of categories processed by the model approaches its optimal extracted category number, its defense performance against natural adversarial examples reaches its peak, providing strong technical support for rapid and accurate decision-making in battlefield environments, while also offering new perspectives for model customization and optimization.

5 Conclusion

This paper conducts exploratory research on the defense against natural adversarial examples from three perspectives: samples, models, and datasets. In terms of samples, we investigate the characteristics of natural adversarial examples and study their attack patterns. Regarding models, we examine the influence of different model architectures and the size of models with the same architecture on the defense against natural adversarial example attacks. In terms of datasets, we analyze the impact of the number of categories in the training dataset and the number of predicted categories during inference on the model's defense capabilities. The following conclusions were reached:

1. We summarized and identified seven typical attack patterns of natural adversarial examples: image distortion, low occupancy ratio, grid occlusion, blending of target objects with backgrounds, excessive depth of field for target objects, significant edge interference, and environmental factors.
2. CNN models equipped with self-attention mechanisms exhibit better defense performance against natural adversarial examples. For most common CNN models, the defense capabilities improve to a certain extent with an increase in the number of parameters.
3. The more concentrated the category distribution of the training dataset, the stronger the model's defense capabilities. Under the condition of a fixed number of categories in the training set, several common CNN models exhibit an optimal range of predicted categories that yields the best defense performance against natural adversarial examples.

Compared to artificially modified digital adversarial samples, natural adversarial examples originate from real images in nature, making them more complex and less controllable, thus aligning more closely with practical application needs. Studying their attack patterns and defense technologies not only enables the training of more robust network models and promotes advancements in deep learning but also facilitates their application in actual military operations, by reducing the enemy's

ability to detect and identify our military targets while enhancing our capability to search for and detect hidden enemy objectives. This paper presents exploratory research conducted after the emergence of natural adversarial examples; although some research conclusions have been drawn, there remains a lack of theoretical and foundational studies on the mechanisms by which natural adversarial examples affect CNN models. Building on prior research, ImageNet-A emerged as the most comprehensive dataset available for our study. Derived from the widely recognized ImageNet dataset, ImageNet-A serves as a robust foundation for our investigation into defenses against natural adversarial examples. Our subsequent efforts will involve generating natural adversarial examples using the MNIST and CIFAR datasets, thereby extending our methodologies to these datasets. Future work will also focus on strengthening research in this area and further investigating specific defense techniques against natural adversarial examples based on the findings of this paper.

Acknowledgement: We would like to express our gratitude to Hefei Zhongke Leinao Technology Co., Ltd., for providing GPU servers during our research, which saved time in our model training and ensured the smooth progress of our experiments.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yaoxuan Zhu, Bin Zhu; data collection: Yaoxuan Zhu; analysis and interpretation of results: Yaoxuan Zhu, Hua Yang; draft manuscript preparation: Yaoxuan Zhu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study, ImageNet-A, can be downloaded from https://github.com/hendrycks/natural-adv-examples (accessed on 12 November 2024, and valid indefinitely). The source code can be obtained by contacting the author at the email: zyx@nudt.edu.cn.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] A. Krizhevsky, I. Sutskever, and E. Hinton, "ImageNet classfication with deep convolutional neural networks," *Adv. Neural Inform. Process. Syst.*, pp. 1097–1105, 2012. doi: 10.1145/3065386.

[2] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015. doi: 10.48550/arXiv.1409.1556.

[3] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015. doi: 10.1109/CVPR.2015.7298594.

[4] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[5] G. Huang, Z. Liu, G. Pleiss, L. V. D. Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8704–8716, 2022. doi: 10.1109/TPAMI.2019.2918284.

[6] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. CVPR*, 2022. doi: 10.1109/CVPR52688.2022.01167.

[7] X. Hu, Y. Liu, and F. Yang, "PFCFuse: A poolformer and CNN fusion network for infrared-visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024. doi: 10.1109/TIM.2024.3450061.

[8] J. Lu, "Decision fusion of CNN and SRC with application to SAR target recognition," *Infrared Laser Eng.*, vol. 51, no. 3, pp. 520–526, 2022.

[9] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014. doi: 10.48550/arXiv.1312.6199.
[10] M. Lee and Z. Kolter, "On physical adversarial patches for object detection," *Statistics*, 2019. doi: 10.48550/arXiv.1906.11897.
[11] R. Duan *et al.*, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. CVPR*, 2020. doi: 10.1109/CVPR42600.2020.00108.
[12] D. Hendrycks, K. Zhao, S. Basar, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. CVPR*, 2021. doi: 10.1109/CVPR46437.2021.01501.
[13] B. Zhou, A. Khosla, A. Lapedriz, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016. doi: 10.1109/CVPR.2016.319.
[14] R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.
[15] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," in *Proc. Int. Conf. Learn. Rep.*, 2021. doi: 10.48550/arXiv.2006.09994.
[16] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. ICCV*, 2017. doi: 10.1109/ICCV.2017.97.
[17] Z. Xu and H. Zhou, "Deep frequency principle towards understanding why deeper learning is faster," in *Proc. AAAI Conf. Artif. Intell.*, 2021. doi: 10.1609/aaai.v35i12.17261.
[18] D. Su *et al.*, "Is robustness the cost of accuracy?— A comprehensive study on the robustness of 18 deep image classification models," in *Proc. ECCV*, 2018. doi: 10.1007/978-3-030-01258-8_39.