A Novel Defense Approach Against Natural Adversarial Examples in ImageNet Dataset

Amirreza Khoshbakht
a.khoshbakht@ug.bilkent.edu.tr
Bilkent University
Ankara, Turkey

Arshia Bakhshayesh
arshia@ug.bilkent.edu.tr
Bilkent University
Ankara, Turkey

Ömer Kagan Danacı
kagan.danaci@ug.bilkent.edu.tr
Bilkent University
Ankara, Turkey

Omar Hamdache
omar.hamdache@ug.bilkent.edu.tr
Bilkent University
Ankara, Turkey

Kerem Karzaoglu
kerem.karzaoglu@ug.bilkent.edu.tr
Bilkent University
Ankara, Turkey

ABSTRACT

Natural adversarial examples present a significant challenge to the robustness of deep neural networks (DNNs) in real-world applications. These examples arise from natural variations within datasets and are not artificially generated. This paper introduces a novel defense approach against natural adversarial examples in the ImageNet dataset by leveraging Salient Feature Extraction (SFE). Our method distinguishes between salient features (SF), which are robust and aligned with human perception, and trivial features (TF), which often mislead models. Utilizing a coupled generative adversarial network (GAN), we effectively extract and prioritize SFs, thereby enhancing the model's ability to accurately classify and defend against natural adversarial examples. Extensive experiments on ImageNet-A demonstrate that our approach significantly improves the robustness of DNNs, outperforming existing state-of-the-art techniques. The implementation and code are made publicly available to support further research in this critical area.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable success across various computer vision tasks, including image classification, object detection, and segmentation. However, these models are vulnerable to adversarial examples—inputs that cause incorrect predictions due to subtle, often imperceptible perturbations. While much research has focused on synthetic adversarial examples, natural adversarial examples pose an equally significant threat. These examples, which arise from the inherent complexities and variations within real-world datasets, can drastically reduce model performance.

Natural adversarial examples, as seen in datasets like ImageNet-A and ImageNet-O, expose shared vulnerabilities across models. ImageNet-A contains challenging in-distribution images that often lead to misclassifications, while ImageNet-O tests a model's ability to handle out-of-distribution (OOD) data, frequently causing high-confidence misclassifications. Current defense strategies, including adversarial training and feature squeezing, are often inadequate against these naturally occurring perturbations.

Our research introduces a novel defense mechanism that leverages Salient Feature Extraction (SFE) to combat natural adversarial examples. SFE focuses on differentiating between salient features (SF)—core, class-related features crucial for accurate model predictions—and trivial features (TF), which adversarial examples exploit to deceive models. By employing a coupled generative adversarial network (GAN) to separate and prioritize these features, our method significantly enhances the robustness and accuracy of DNNs against natural adversarial examples. We validate our approach on the ImageNet dataset, demonstrating superior performance in both detection and defense compared to existing methods. This work advances the field of adversarial defense by providing a robust solution to the challenges posed by natural adversarial examples in real-world applications.

2 BACKGROUND AND RELATED WORK

The vulnerability of deep neural networks (DNNs) to adversarial examples has been widely documented since the seminal work by Szegedy et al. (2013). Initially, research primarily focused on synthetic adversarial examples—artificially crafted perturbations designed to deceive models. However, recent studies have highlighted the equally concerning threat posed by natural adversarial examples, which arise from the inherent complexities and variations within datasets. Datasets such as ImageNet-A and ImageNet-O have been specifically curated to expose these vulnerabilities, offering challenging benchmarks that reveal the deficiencies in current model robustness. Existing defense strategies, including adversarial training, defensive distillation, and feature squeezing, have shown some success against synthetic attacks but often fall short against natural adversarial examples. This gap underscores the need for novel defense mechanisms that can effectively handle naturally occurring adversarial perturbations, ensuring the robustness and reliability of DNNs in real-world applications.

2.1 Imagenet-A and Imagenet-O

To address the challenge of natural adversarial examples, we consider two specific datasets: ImageNet-A and ImageNet-O, which provide valuable benchmarks for evaluating model robustness against naturally occurring adversarial perturbations.

ImageNet-A contains images from the same classes as the original ImageNet dataset but selected to be significantly more challenging. These images exploit the long tail of scene configurations and classifier blind spots to consistently cause misclassifications across various models. Despite being naturally occurring and unaltered, these examples induce substantial performance degradation, exposing shared weaknesses in contemporary models. For instance, a DenseNet-121 model, which typically achieves high accuracy on standard ImageNet data, drops to around 2% accuracy on ImageNet-A, reflecting a roughly 90% decrease .

ImageNet-O, on the other hand, is designed to test a model's out-of-distribution (OOD) detection capabilities. It comprises images from classes not included in the ImageNet-1K dataset. These images often cause models to mistakenly classify them with high confidence as in-distribution examples. ImageNet-O thus evaluates a model's ability to handle semantic shifts in the data distribution. Models that perform well on ImageNet-1K tend to struggle with these OOD images, highlighting the models' susceptibility to overconfidence in unfamiliar contexts .

Together, these datasets provide a comprehensive evaluation framework for testing and improving the robustness of deep neural networks against natural adversarial examples, making them crucial for developing and assessing new defense mechanisms. Our novel approach aims to enhance model performance on these datasets by effectively differentiating between robust, human-aligned salient features and misleading trivial features .

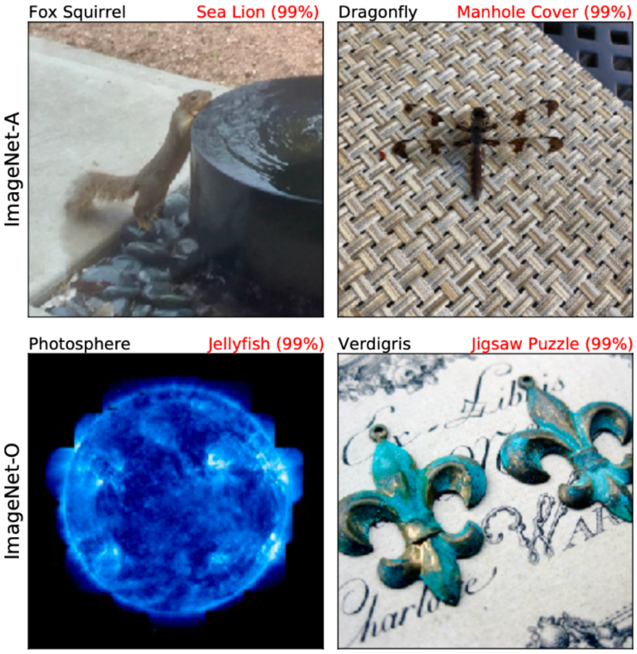

Figure 1: Natural Adversarial Examples [2]

2.2 Threat model

Our threat model focuses on natural adversarial examples within the ImageNet dataset, specifically leveraging ImageNet-A and ImageNet-O as benchmarks. These examples arise from natural variations in the dataset and do not involve synthetic perturbations. The threat model is defined by the following characteristics:

Adversarial Nature: The adversarial examples in question are naturally occurring and not artificially generated. They exploit inherent complexities, rare configurations, and edge cases in the dataset to deceive models.

ImageNet-A: This subset consists of images that are particularly challenging for standard classifiers, often causing significant misclassifications. These images exploit the long tail of visual configurations that are less represented in the training data, revealing vulnerabilities in the model's feature extraction and decision-making processes.

ImageNet-O: This subset includes out-of-distribution (OOD) images that belong to classes not present in the ImageNet-1K training set. These images test the model's ability to distinguish between in-distribution and out-of-distribution data, exposing the model's tendency to confidently misclassify unfamiliar images.

Salient Feature Extraction (SFE): Our approach leverages Salient Feature Extraction (SFE) to enhance the robustness of DNNs against natural adversarial examples. SFE focuses on distinguishing between salient features (SF), which are robust and aligned with human perception, and trivial features (TF), which adversarial examples often exploit to mislead models. By emphasizing SF and minimizing the influence of TF, our method aims to improve the model's ability to correctly classify and defend against natural adversarial examples.

Evaluation Metrics: The effectiveness of our defense mechanism will be evaluated using metrics such as accuracy, robustness, and the model's capability to detect and correctly classify natural adversarial examples. Performance will be measured on the ImageNet-A dataset to assess robustness to difficult in-distribution examples.

In summary, our threat model targets naturally occurring adversarial examples that expose the inherent weaknesses of DNNs. By applying Salient Feature Extraction, we aim to enhance the robustness and reliability of these models in real-world scenarios, providing a robust defense against natural adversarial perturbations.

3 METHODOLOGY

3.1 Overview

Our methodology centers on the extraction, separation and analysis of salient features and trivial features from images to detect and defend against adversarial examples. The approach involves a multi-stage process including feature extraction, adversarial detection and re-identification of the correct classification labels through an SFE framework. This framework leverages coupled Generative Adversarial Networks and a dedicated adversarial detector (AdvD) to enhance model robustness.

3.2 Framework

The framework of our proposed method, as shown in Figure 2, comprises several key components: feature extraction and separation, adversarial example detection, and defense via re-identification of SF. Initially, benign and adversarial examples are fed into a pre-trained target model. The high-dimensional features obtained from the last fully connected layer of this model serve as the input to the SFE. The SFE employs a coupled GAN structure to separate and extract SF and TF. An adversarial detector (AdvD) is subsequently trained using these extracted features. Finally, adversarial examples are detected by evaluating the difference between SF and TF, while correct labels are reassigned based on the re-identified SF.

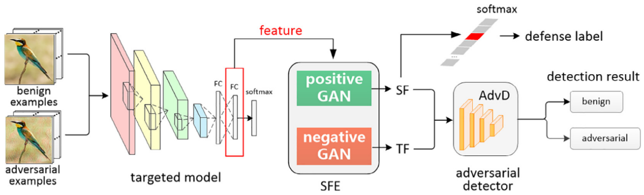

Figure 2: SFE Framework [1]

In the framework, the SFE comprises two coupled GAN structures: a positive GAN (responsible for SF) and a negative GAN (responsible for TF). The positive GAN includes a positive generator (PG) and a discriminator (D), which are designed to learn and generate salient features. Conversely, the negative GAN consists of a negative generator (NG) and the same discriminator (D), but focuses on trivial features.

The generators (PG and NG) within the SFE are structurally identical but are trained with different data to perform their distinct functions. Each generator consists of stacked fully connected layers tailored to handle high-dimensional image features. The input layer size of the generators is [H, W, C], corresponding to the dimensions of the input image, while the output layer size is [H x W x C, 1]. The discriminator (D) plays a crucial role in classifying the generated features as true or false. It shares parameters across both the positive and negative GANs to maintain consistency in decoding high-level features, thereby effectively capturing the relationship between SF and TF. This shared parameter setup helps in generating features that closely resemble the ground truth distribution.

The training process of the SFE involves inputting benign and adversarial examples into the target model and extracting the high-dimensional features from the last fully connected layer. These features are then used to train the SFE. The goal is to make the SF of both benign and adversarial examples indistinguishable, while the TF should match the high-dimensional feature layer for both input types. The optimization objective is to minimize the mean square error (MSE) between the input features and the generated features, ensuring that the generated data closely approximates real data. The parameters of the PG and D are updated alternately during the training process to refine the model.

The AdvD in the framework is designed to identify adversarial examples by analyzing the separated SF and TF. Since SF and TF are similar in benign examples but differ in adversarial examples, this characteristic is leveraged to train the AdvD.

The AdvD consists of five fully connected layers. The input layer size is [H x W x C, 1], matching the output of the SFE generators, while the output layer size is [1, 1]. During training, the output of the PG (SF) and NG (TF) are concatenated to form the training set for the AdvD. The AdvD is trained to output a binary classification: 0 for benign examples and 1 for adversarial examples. The parameters of the AdvD are optimized by minimizing the classification loss (lossAdvD).

After the training of the SFE is completed, its outputs are used to train the AdvD. Specifically, the SF of benign and adversarial examples, generated by the PG, and the TF generated by the NG, are concatenated to create the training set for the AdvD. During the training phase, the concatenated features are fed into the AdvD, which then outputs the detection result: ’benign’ marked as 0 and ’adversarial’ marked as 1. The AdvD’s parameters are updated by minimizing the detection loss, ensuring that the AdvD can effectively distinguish between benign and adversarial examples.

For re-identifying the correct labels of adversarial examples, the SF reconstructed by the PG in the SFE is utilized. These salient features retain critical information necessary for accurate classification. When the well-trained SF is input to the target model, it ensures the correct classification of adversarial examples, thus providing a robust defense mechanism.

As defined earlier, the SF of adversarial examples are the same as those of their corresponding benign examples and closely related to the output of the hidden layer in the model. High-dimensional image features contain significant information that is essential for classification. The positive GAN in the SFE reconstructs these high-dimensional features and enhances the important features. Consequently, for a well-trained SFE, the SF generated by the PG still contains the crucial information needed for accurate classification. These reconstructed SFs are then input into the target model, which subsequently provides the correct classification results for adversarial examples, effectively defending against them.

3.3 Other Implementation Details

In the implementation, we used a dataset consisting of 2640 benign images and 2640 adversarial images. The dataset was split into an 80% training set (2112 images) and a 20% test set (528 images) to evaluate the performance of our method. The classification model employed is InceptionV3 [3], which is pre-trained on the ImageNet-1k dataset [4], encompassing 1000 classes.

The training process involves multiple stages. Initially, the SFE is trained using the high-dimensional features extracted from the target model. Benign and adversarial examples are fed into the target model to obtain these features, which are then used to train the SFE. The MSE objective is employed to minimize the distance between the input features and the generated features, ensuring that the generated features closely resemble the real data.

Subsequently, the AdvD is trained using the concatenated outputs of the PG (SF) and NG (TF) from the SFE. The AdvD learns to distinguish between benign and adversarial examples based on the differences in SF and TF. The training set for the AdvD consists of these concatenated features, and the AdvD is optimized to correctly classify the examples as either benign or adversarial.

In the defense stage, the well-trained SFE reconstructs the SF from the input images. These SFs are then fed into the target model, which re-identifies the correct classification labels for the adversarial examples. This stage ensures robust defense against adversarial attacks by accurately classifying the input images based on the reconstructed salient features.

4 RESULTS AND ANALYSIS

After training the GAN model and conducting some evaluations on the InceptionV3, we recorded the accuracy of the model and compared it with its accuracy before using the SFE method. Our test dataset consists of 528 natural adversarial images, sampled from the adversarial dataset ImageNet-A.

4.1 Accuracy Without Defense

Feeding the InceptionV3 model with the 528 images that were sampled for testing, we obtained 3% accuracy. This shows that the model is unable to classify most of the images, missing 513 images out of 528. Since the dataset is meant to be an adversarial dataset, such results are expected. The reason behind this is as mentioned before, the model is taking in Trivial Features (TF) that are negatively affecting the results of the classification.

4.2 Defense Accuracy

After feeding the trained model with the test dataset described above, we compared the classified label with the true label of the image and calculated an overall accuracy. The accuracy of the defense model moved from 3% to 79%. The model now is able to correctly classify 415 images, increasing the number of correctly classified images by 400. This noticeable increase is due to the model now being fed Salient Features (SF) instead of Trivial Features (TF). Most of the adversarial examples are failing to trick the model with its TF dominance, since now those TF are eliminated.

4.3 Visualizing the Results

In order to view what the model is basing the classification on, we used Gradient-weighted Class Activation Mapping (GRAD-CAM) [5]. GRAD-CAM helps to make the decision of the model easier to understand by generating a heatmap that can be then overlaid on the original image. The heatmap will indicate the locations on the image where the model was affected the most during the classification. The process is as follows:

(1) The input image is passed through the CNN, and the final convolutional layer's feature maps are obtained.
(2) The gradients of the score for the target class with respect to the feature maps are computed. These gradients represent how changes in the feature maps affect the final prediction.
(3) The gradients are then weighted by the importance of each feature map, which is determined by the average gradient value of each feature map. This emphasizes the importance of feature maps that have higher gradients.
(4) The weighted combination of gradients is summed up over all the feature maps to obtain the final heatmap.
(5) Finally, the heatmap is overlaid on the input image to visualize which regions of the image are most relevant for the prediction of the target class.

We can see from Figure 3 that the output of GRAD-CAM before applying the defense highlights the areas around the object. This is due to the model being tricked by the TFs and misclassified the image. We can see that the true label of this image is dragonfly, but it is classified as manhole_cover. The natural adversarial examples that we used have the same effect on the model, tricking it with trivial features.

After applying the defense mechanism utilizing SFE, we can see from Figure 4 that the image is now correctly classified as dragonfly. We can also see that the heatmap generated is now focusing primarely on the dragonfly itself, and not on the area around it. This is due to the SFE feeding the SFs to the model, and eliminating the TFs that were previously tricking the model.

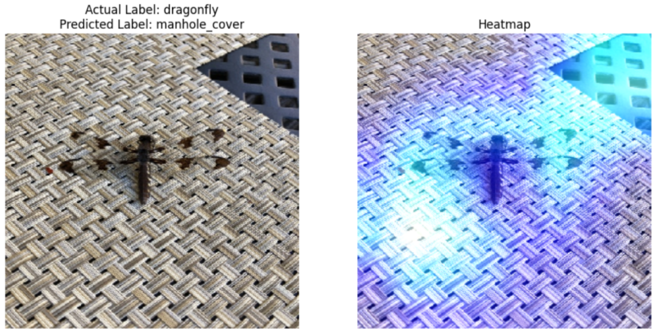

Figure 3: GRAD-CAM output before defense

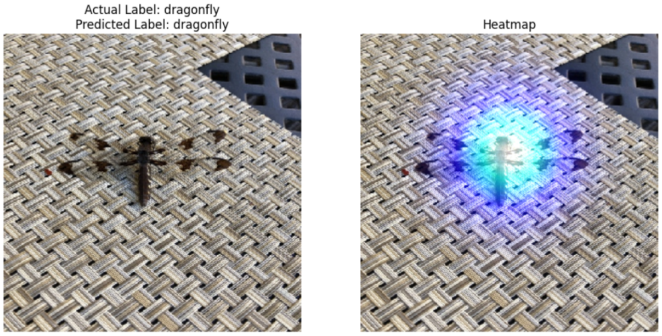

Figure 4: GRAD-CAM output after defense

5 DISCUSSION

The defense approach proposed in this paper addresses the critical challenge posed by natural adversarial examples in the ImageNet dataset. By leveraging Salient Feature Extraction (SFE), the method aims to enhance the robustness of deep neural networks (DNNs) against naturally occurring perturbations, thereby improving model performance in real-world applications.

The results presented demonstrate a significant improvement in model accuracy and robustness when defending against natural adversarial examples. Without defense, the model achieved only a 3% accuracy rate on a test dataset comprising adversarial images sampled from ImageNet-A. However, after implementing the defense mechanism, the accuracy substantially increased to 79%. This notable enhancement underscores the effectiveness of the proposed approach in mitigating the impact of adversarial perturbations on model predictions.

Visualizations using Gradient-weighted Class Activation Mapping (GRAD-CAM) provide further insights into the decision-making process of the model before and after applying the defense mechanism. Before defense, the model tends to focus on trivial features surrounding the objects in the images, leading to misclassifications. In contrast, after defense, the model's attention shifts towards salient features relevant to the objects, resulting in more accurate classifications.

This approach was not implemented before for defending against natural adversarial examples. The method's effectiveness was only investigated for the defence against adversarial examples generated by using perturbations in the paper we mentioned before [1]. Our results are valuable in the area of defense against adversarial examples, since natural adversarial examples are one of the most commonly seen types, since they occur naturally in most datasets, and they cannot be detected by an observer checking the images.

6 CONCLUSION

In conclusion, this paper introduces a novel defense mechanism leveraging Salient Feature Extraction (SFE) to combat the significant challenge posed by natural adversarial examples in the ImageNet dataset. The proposed approach demonstrates a substantial improvement in the robustness and accuracy of deep neural networks (DNNs) against naturally occurring perturbations, thereby enhancing model performance in real-world applications.

The results presented highlight a remarkable enhancement in model accuracy, with a notable increase from 3

It is noteworthy that this defense approach was not previously implemented specifically for defending against natural adversarial examples. Prior research primarily focused on synthetic adversarial examples, making this study a valuable contribution to the field. Natural adversarial examples pose a significant challenge as they occur naturally in most datasets and cannot be easily detected by human observers.

Overall, the findings presented in this paper contribute to advancing the field of adversarial defense, particularly in addressing the vulnerabilities of DNNs to naturally occurring perturbations. By effectively differentiating between robust salient features and misleading trivial features, the proposed approach enhances the reliability and robustness of DNNs, thereby paving the way for more secure and trustworthy applications in computer vision and related fields.

REFERENCES

[1] Jinyin Chen, Ruoxi Chen, Haibin Zheng, Zhaoyan Ming, Wenrong Jiang, and Chen Cui. 2021. Salient Feature Extractor for Adversarial Defense on Deep Neural Networks. (2021). arXiv:cs.CV/2105.06807

[2] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural Adversarial Examples. (2021). arXiv:cs.LG/1907.07174

[3] Keras Team. n.d.. Keras Documentation: Inceptionv3. (n.d.). https://keras.io/api/applications/inceptionv3/ Accessed: 15 May 2024.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. (2015). arXiv:cs.CV/1409.0575

[5] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* abs/1610.02391 (2016). arXiv:1610.02391 http://arxiv.org/abs/1610.02391