Enhancing Adversarial Robustness with Autoencoder-Based Detection

Joey Mulé
jmule2@umbc.edu

Alex Shaner
shaner1@umbc.edu

Abstract

The increasing reliance on real-time object detection and image classification systems underscores the urgent need for more robust frameworks capable of handling adversarial attacks and unexpected inputs. Such inputs can significantly degrade model performance and erode trust in neural networks. In this work, we propose a novel framework that combines feature-based reconstruction techniques with softmax confidence scoring to enhance out-of-distribution (OoD) and anomaly detection. Our approach leverages feature extraction from a pretrained network, transformation-based reconstruction, and confidence scoring to effectively differentiate in-distribution (ID) and OoD samples. Experimental results demonstrate state-of-the-art performance in OoD detection and robust anomaly detection across multiple benchmarks. This framework provides an efficient strategy to improve the robustness of neural networks against adversarial conditions. https://github.com/joeduman/Upgrading-Autoencoders

1. Introduction

Real-time object detection and image classification technologies have become increasingly popular due to their wide-ranging applications and versatility. However, despite their widespread use, these technologies remain highly vulnerable to adversarial attacks. Adversaries can exploit even small changes in input data—such as Gaussian noise, image blur, or pixel-level corruptions—to drastically degrade model performance. These adversarial inputs are often hard to notice by the human eye but can lead to significant misclassifications or complete failure of the model to detect objects accurately.

One major challenge lies in the sensitivity of neural networks to minor alterations in input data. Even a seemingly small amount of noise can push a neural network towards an incorrect classification, highlighting a lack of robustness. This vulnerability becomes more pronounced and apparent in real-time applications, where time constraints and dynamic environments make it difficult to ensure clean input data is applied to the network. Furthermore, defending against these attacks requires substantial computational resources. Adversarial training, which exposes models to adversarial examples during training, often comes with the tradeoff of increased complexity and reduced efficiency. Relying on these training methods are often impractical for real-time systems where computational overhead must be minimized to maintain robustness.

Adversarial attacks often exploit the inherent weakness in the neural network's weight distribution. When adversarial noise is introduced, it distorts the activation patterns of the network, causing it to miscalculate and produce errors. As a result, a model's confidence significantly decreases, and the overall robustness of the system is compromised.

Without an effective defense or training regimen, adversarial attacks can significantly undermine the trustworthiness of neural networks in real life scenarios, and we intend to avoid this.

2. Background

Autoencoders are neural networks designed to compress input data into a lower-dimensional latent representation and then reconstruct it as accurately as possible. They consist of two main components: an encoder, which maps the input into a latent space, and a decoder, which reconstructs the input from this latent representation. Using autoencoders for out-of-distribution (OoD) and anomaly detection generally involves training the encoder on in-distribution data, measuring the reconstruction error, and then comparing that reconstruction error with the error from test samples. Since the autoencoder has not seen OoD data during training, intuition tells us that it would struggle to reconstruct such inputs accurately, leading to higher reconstruction errors for anomalies or OoD samples compared to in-distribution (ID) data. However, the effectiveness of this approach can vary, as some autoencoders may reconstruct OoD samples well if they share enough features with the training data or are just simple images.

2.1. Related Work

Hendrycks' and Gimpel's foundational work ”A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks” [1] introduced a softmax-based confidence scoring method for OoD detection. This approach computes confidence scores where the maximum softmax probability (MSP) serves as a measure of the model's confidence in its predictions. These confidence scores were used to distinguish between ID and OoD data. This work laid a foundational baseline for similar work and highlighted the need for better model calibration.

Considerable inspiration for this work was drawn from Beihang University, Yibo Zhou, and his recent publication: "Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection" [2]. In Zhou's work, he attempted to "formulate the essence of such approach as a quadruplet domain translation with an intrinsic bias to only query for a proxy of conditional data uncertainty." Additionally, "improvement direction was formalized as maximumly compressing the autoencoder's latent space..." His approach minimized a regularization loss to compress the latent space, ensuring that OoD samples fall outside the compact representation while maintaining sufficient reconstructive power by shifting the reconstruction target from raw input to activation vector features. Additionally, a framework for layerwise semantic reconstruction was developed, leveraging a simple encoder architecture with cross-entropy loss to refine domain translation and improve OoD detection robustness through normalized L2 distance evaluations. We implement these features in our method and as such readers are encouraged to refer to Zhou's original work with for more detail.

3. Method

The goal of this work is to reconstruct images using their extracted features, derived from a pretrained network or encoder, to effectively separate ID and OoD samples. The process begins by freezing all parameters of a pretrained network, which acts as a feature extractor. During training, a transformation matrix (W) is optimized to map extracted features for reconstruction. Additionally, two decoders are trained: D_1, which directly reconstructs the features, and D_2, which refines logits derived from W through a softmax transformation. A customized loss function is used to guide this training. The output includes the trained components D_1, D_2 and W, along with fitted parameters (μ and σ) to facilitate anomaly detection.

Zhou's algorithm, shown in Alg.1 provides a structured training pipeline for OoD detection using auxiliary (AV) features and decoders. The pre-trained network, $M(\cdot)$, serves as a feature extractor, deriving AV features v_i from the input data x_i. A transformation matrix, W, is applied to the extracted AV features to facilitate the reconstruction process. The decoders, D_1 and D_2, are trained to reconstruct AV features and logits, respectively, enabling a more refined representation. The training process incorporates two reconstruction losses, $\mathcal{L}_1$ and $\mathcal{L}_2$, where $S(\cdot)$ denotes the softmax function, and T is a temperature scaling factor used for logits. To ensure the AV features are well-structured in the latent space, a regularization term, $\mathcal{L}_{\text{regularizer}}$, is added to the loss function. After training, Gaussian distribution parameters (μ_0, σ_0), (μ_1, σ_1), and (μ_2, σ_2) are fitted on validation data. These parameters are subsequently used to calculate the final anomaly scores, which distinguish ID samples from OoD samples.

Algorithm 1 Zhou's training pipeline

```
Require: ID training set $\{(x_i, y_i)\}_{i=1}^{k}$, and ID validation set $\{(x_i, y_i)\}_{i=k+1}^{n}$
Require: Network $M(\cdot)$ fully trained on ID training set for classification of ID classes
1: Freeze all parameters of network $M(\cdot)$, and jointly train $W \in \mathbb{R}^{H \times C}$ ($C$ is the number of ID classes) and two decoders $D_1$ & $D_2$ to minimize the loss $\mathcal{L}$:

       $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda \cdot \mathcal{L}_{\text{regularizer}}$

       $\mathcal{L}_1 = \sum_{i=1}^{k} \|v_i - D_1(Wv_i)\|, \quad \mathcal{L}_2 = \sum_{i=1}^{k} \left\| \frac{Wv_i}{T} - D_2\left(S\left(\frac{Wv_i}{T}\right)\right) \right\|$

       $\mathcal{L}_{\text{regularizer}} = -\sum_{i=1}^{k} \sum_{j=1}^{C} \nVdash(j = y_i) \log S(Wv_i)_j$

2: where $v_i$ is $x_i$'s AV feature extracted in $M(\cdot)$ and $\lambda$ is the weight of the regularization loss
3: After training, compute:

       $(\mu_0, \sigma_0) = \text{norm.fit}\left(\left\{ S\left(\frac{Wv_i}{T}\right)_{\bar{y}_i} \right\}_{i=k+1}^{n}\right)$

       $(\mu_1, \sigma_1) = \text{norm.fit}\left(\left\| \frac{v_i}{\|v_i\|} - \frac{D_1(Wv_i)}{\|v_i\|} \right\|_{i=k+1}^{n}\right)$

       $(\mu_2, \sigma_2) = \text{norm.fit}\left(\left\| \frac{Wv_i}{T} - D_2\left(S\left(\frac{Wv_i}{T}\right)\right) \right\|_{i=k+1}^{n}\right)$

4: return $D_1$, $D_2$, $W$, $(\mu_0, \sigma_0)$, $(\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2)$
```

Secondly, this work submerses decoder based feature extraction with Hendrycks' softmax confidence calculation in Alg.2. Hendrycks' softmax confidence score can be calculated by extracting logits from a pretrained network, applying the softmax function to obtain class probabilities, and selecting the maximum probability as the confidence score. A threshold τ then separates the two domains.

3.1. Overview

Our work builds on Zhou's approach by incorporating a confidence-based softmax scoring mechanism inspired by Hendrycks. Specifically, we integrate Hendrycks' softmax confidence scores into Zhou's feature extraction methodology to improve the encoding, classification, and decoding processes. The softmax confidence scores help distinguish between ID and OoD samples by analyzing score distribu-

Algorithm 2 Hendrycks' Softmax Confidence Score

```
Require: Pretrained network $M(\cdot)$, input data $X$, threshold $\tau$
Ensure: Predicted labels for ID and OoD samples
Initialize:
    $S \leftarrow \emptyset$    ▷ Set to store softmax confidence scores
Forward Pass:
for each input $x_i \in X$ do
    $\mathbf{z}_i \leftarrow M(x_i)$    ▷ Extract logits $\mathbf{z}_i$ from the network
    $\mathbf{p}_i \leftarrow \text{softmax}(\mathbf{z}_i)$    ▷ Compute softmax probabilities
    $S_i \leftarrow \max(\mathbf{p}_i)$    ▷ Confidence score is the max softmax probability
    Append $S_i$ to $S$
end for
OoD Detection:
for each score $S_i \in S$ do
    if $S_i > \tau$ then
        Label $x_i$ as ID (In-Distribution)
    else
        Label $x_i$ as OoD (Out-of-Distribution)
    end if
end for
return Predicted labels for all $x_i \in X$
```

tions, providing a measure of classification certainty. This enables both anomaly detection and enhanced feature extraction. Meanwhile, Zhou's feature-based decoding emphasizes image reconstruction through feature transformations, focusing on latent space analysis. For our implementation, we utilized the pre-trained ResNet18 model as the backbone for feature extraction.

3.2. Scope of Application

Amongst our methodology, the proposed approach is designed to be domain-agnostic and theoretically applicable to various data modalities other than images, like text. While our implementation and evaluation were conducted solely on image data, we had considered its broader applicability during the design process. In theory, with appropriate modifications for feature extraction (e.g., using LSTMs or temporal CNNs) and decoding, the approach could be extended to text, or possibly other domains, enabling the approach's diversity.

4. Experiments

Our experiments compare the three phases of our discussed methodologies: The '**base**' method which illustrates the performance of a standard-trained autoencoder; '**Zhou's**' method where we implement his regularization, loss minimization, and feature-level extraction; and our '**combined**' method where we apply Hendrycks' softmax confidence scoring to the feature-level extraction. We compare the methods' anomaly detection Area Under the Receiver Operating Curve (AUROC), OoD detection AUROC, and true OoD detection rate to evaluate performance.

4.1. Experimental Setup

For the ID data, we utilized the CIFAR-10 dataset, while the OoD data was represented by a resized version of Tiny-ImageNet. Additionally, for our anomaly detection, we created a custom dataset derived from CIFAR-10 by applying various adversarial perturbations; including Gaussian noise, salt-and-pepper noise, FGSM attacks, brightness and contrast adjustments, and pixel dropout. These modifications were designed to simulate anomalous data for enhanced robustness testing.

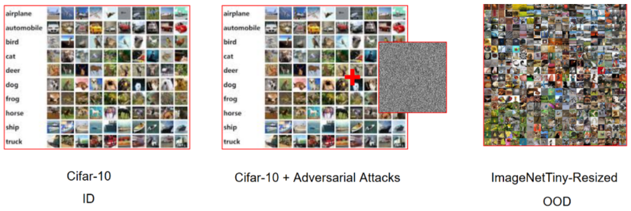

Figure 1. Datasets used in experiments

4.2. Results

The results of our combined method demonstrate comparable OoD detection rates to Zhou's method, which already exhibited a high level of success in identifying OoD samples. However, it is important to note that Zhou's work was primarily focused on OoD detection and did not address anomaly detection. In contrast, our combined method extends this framework to anomaly detection, achieving a significant improvement in detecting anomalous samples (see Table 1). We also optimize Zhou's method to work seamlessly within our anomaly detection pipeline. This integration resulted in enhanced robustness, with the combined approach outperforming the individual methods in anomaly detection while maintaining SOTA performance in OoD detection.

Table 1. OoD and Anomaly Detection Performance

Method	**OoD Acc. (%)**	**Anomaly Acc. (%)**
Simple Autoencoder	15.89	70.84
Zhou's Method	99.93	85.05
Combined Method	99.90	95.94

The metrics we computed include OoD accuracy (detected OoD examples / total OoD examples), anomaly accuracy (detected anomalous examples / total anomalous examples), and AUROC (Area under the TPR vs. FPR curve).

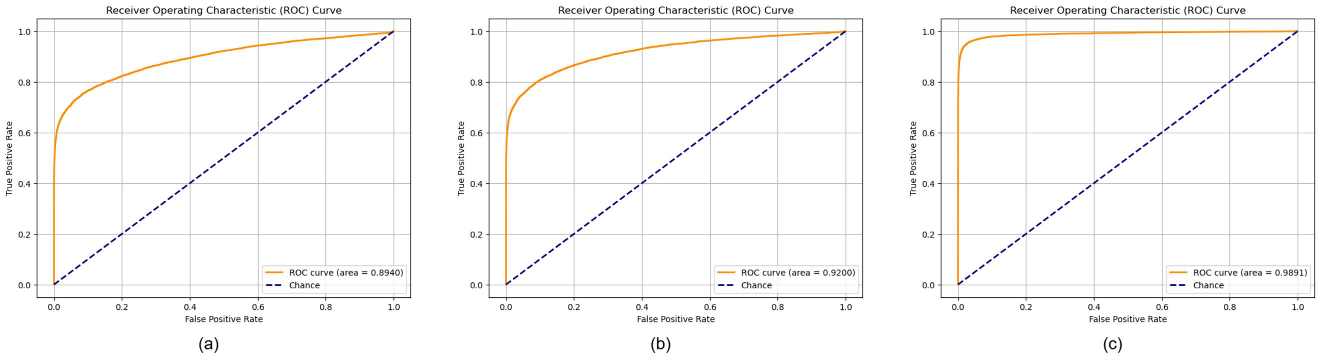

Figure 2. **Higher value is better** (a): AUROC for standard-trained autoencoder anomaly detection method (baseline). (b): AUROC for Zhou's method, which shows improvement over the baseline. (c): Our combined approach, which reflects the highest AUROC performance in our experiment.

4.3. Ablation Study

To evaluate the contributions of the individual components in our combined OoD and anomaly detection framework, we conducted an ablation study. The study focused on understanding the roles of feature extraction and reconstruction within autoencoders. This led towards the inspiration of building off of Zhou's design process, and also applying softmax confidence scoring in improving each tasks' performance. The framework was implemented and tested using both TensorFlow and PyTorch to ensure code reproducibility.

The experimental configurations in this study explored the individual and combined effects of feature extraction and confidence-based scoring mechanisms on OoD detection—three configurations and models were tested. The first involved a baseline autoencoder trained with mean-squared-error (MSE) loss to reconstruct input data (pixel-wise) without additional enhancements. The second configuration introduced a pre-trained network for feature extraction, followed by decoders trained to reconstruct extracted features and logits, with a transformation matrix facilitating the reconstruction process. Finally, the third configuration combined feature reconstruction with a softmax confidence-based scoring mechanism to detect both structural anomalies and prediction uncertainties.

The results highlighted the complementary nature of the two approaches. Feature reconstruction, based on the extracted features rather than raw inputs, improved the separation between ID and OoD samples by emphasizing structural inconsistencies in unfamiliar data. The confidence-based scoring mechanism further enhanced detection by penalizing uncertain or overconfident predictions through entropy-based analysis of softmax outputs. When these two mechanisms were combined, the model achieved the best performance, as the feature reconstruction addressed semantic anomalies while the confidence scoring captured uncertainty in classification.

The ablation study also explored the effects of varying hyperparameters, such as the temperature scaling factor T in the softmax confidence calculation, as well as the weighting coefficients λ and β for the loss components. These parameters were found to significantly influence model performance. For instance, a lower temperature value led to sharper softmax distributions, which improved confidence-based detection but at the cost of increased sensitivity to noise. Similarly, the choice of λ and β affected the balance between reconstruction quality and confidence regularization. By systematically tuning these parameters, we observed that the combined approach provided robust OoD detection across a range of settings, further validating the adaptability of the framework. Additionally, the study revealed that the reconstruction component was particularly effective at identifying structural inconsistencies, whereas the confidence-based scoring focused on detecting ambiguous or borderline cases. This demonstrates the complementary nature of these components and underscores the need to balance their contributions within the framework.

4.4. Visual Comparison

Upon analyzing the reconstruction error and/or feature reconstruction error, we can determine between non-corrupted and corrupted data.

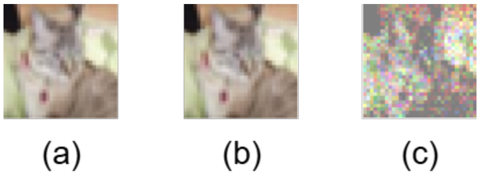

Figure 3. (a): Original CIFAR-10 image. (b): Autoencoder reconstructed image. (c): Autoencoder reconstructed corrupted image.

5. Limitations

OoD detection using autoencoders has long presented challenges, particularly due to the poor generalization of these techniques. Fine-tuning often becomes essential to address these limitations, a concern that this work aimed to mitigate. However, it is important to acknowledge that identical performance of the proposed detection framework cannot be guaranteed when applied to vastly different OoD or anomaly datasets, highlighting an inherent limitation of current methodologies.

To progress toward domain-agnostic OoD and anomaly detection, it is crucial to employ datasets that encompass a wide range of features, both similar and distinct, to rigorously test the decoders under diverse scenarios. One of the technical challenges encountered during this work was the dimensionality of the data, which required careful consideration and adaptation during the experimental setup to ensure optimal performance.

Additionally, the reliance on a fully trained neural network posed significant time constraints. Retraining the autoencoder following implementation adjustments and calculating reconstruction errors, as well as determining appropriate thresholds for OoD and anomaly detection, were time-intensive tasks. These computational demands consistently limited the pace of experimentation and iterative refinement.

6. Future Work

Future work could entail the exploration of using a wider variety of datasets as a short-term or context-specific solution to address the lack of generalizability in current methods. By incorporating diverse datasets, models can be evaluated and refined to perform better across a broader range of scenarios, reducing biases tied to specific domains or limited data distributions.

While we conducted some basic, well-used metrics within the field, future work should emphasize the inclusion of additional standardized metrics for evaluating anomaly and OoD detection. Metrics such as False Positive Rate at 95% True Positive Rate (FPR@95%TPR) can provide deeper insights into the model's robustness under high true-positive regimes by assessing its tendency to generate false alarms. Similarly, incorporating metrics like the Area Under the Precision-Recall Curve (AUPR) would complement AUROC by offering a nuanced perspective, particularly valuable for imbalanced datasets. Additionally, calibration metrics such as Expected Calibration Error (ECE) [3] could assess how well the confidence levels of softmax-based predictions align with actual outcomes. Exploring OoD-specific approaches like Mahalanobis distance [4] [5] or Energy-Based [6] Outlier Scores would also establish meaningful baselines to better compare the robustness of our proposed methods against SOTA techniques.

Additionally, introducing different types of adversarial attacks could further stress-test and refine the system's resilience. By simulating a wide array of adversarial scenarios, such as those leveraging subtle perturbations or more complex, targeted attack strategies, the model can be challenged to detect anomalies effectively under adversarial conditions.

Exploring the use of variational autoencoders represents another promising direction. VAEs, with their probabilistic latent space, offer the advantage of modeling data distributions explicitly, making them well-suited for quantifying uncertainty which could be utilized as we have done in this work with reconstruction error.

7. Conclusion

In this work, we proposed a novel OoD and anomaly detection framework by integrating Zhou's feature-based reconstruction and Hendrycks' softmax scoring. Through this combination, our approach leverages the strengths of both works, and effectively addresses challenges of OoD and Anomaly detection.

Our experiment demonstrates that our method not only maintains SOTA performance in OoD Detection but also improves on baseline approaches for anomaly detection. This increase in performance demonstrates our method's increased robustness and reliability, which is of high value in safety-critical scenarios.

Despite our success, this approach still faces limitations in generalizability and computational efficiency during implementation. We acknowledge these weaknesses and believe that they are areas with room for improvement. Overall this work provides a strong foundation for integrating feature reconstruction with probabilistic confidence measures, contributing to more robust and versatile neural network systems for real-world applications.

8. Contributions

- Project Research and Conceptualization: **J. Mulé** and **A. Shaner** collaboratively developed the research idea and conceptualized the experimental framework.

- Implementation and Coding: **J. Mulé** led the coding efforts, including the implementation of algorithms and experimental pipelines.

- Writing and Presentation: **J. Mulé** and **A. Shaner** jointly contributed to the project report and presentation materials, with **J. Mulé** taking the lead on intensive writing.

References

[1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl 1

[2] Y. Zhou, "Rethinking reconstruction autoencoder-based out-of-distribution detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7369–7377. 2

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, 2017. 5

[4] P. C. Mahalanobis, "On the generalised distance in statistics," *Sankhyā: The Indian Journal of Statistics*, vol. 2, no. 1, pp. 49–55, 1936. 5

[5] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf 5

[6] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 464–21 475. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf 5