Familiarity-Based Open-Set Recognition Under Adversarial Attacks

Philip Enevoldsen Christian Gundersen Nico Lang Serge Belongie Christian Igel
Department of Computer Science, University of Copenhagen

Abstract

Open-set recognition (OSR), the identification of novel categories, can be a critical component when deploying classification models in real-world applications. Recent work has shown that familiarity-based scoring rules such as the Maximum Softmax Probability (MSP) or the Maximum Logit Score (MLS) are strong baselines when the closed-set accuracy is high. However, one of the potential weaknesses of familiarity-based OSR are adversarial attacks. Here, we present gradient-based adversarial attacks on familiarity scores for both types of attacks, False Familiarity and False Novelty attacks, and evaluate their effectiveness in informed and uninformed settings on TinyImageNet.

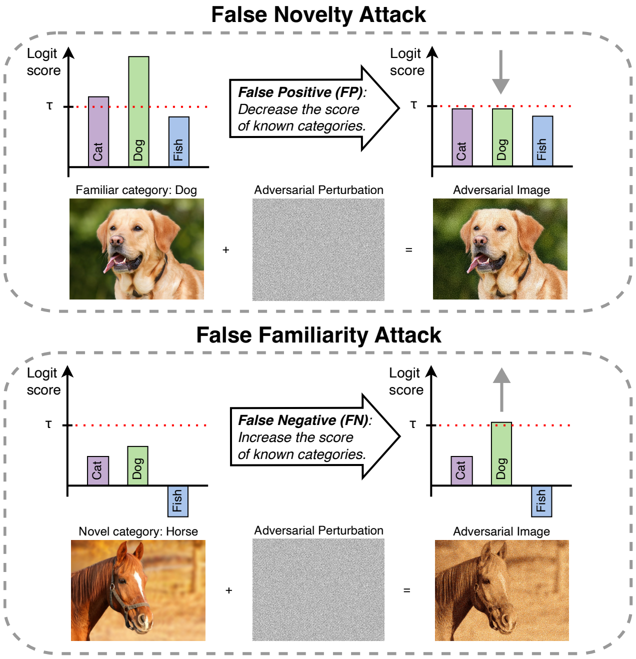


Figure 1: **Adversarial attacks on OSR familiarity scores.** Considering novel categories as positives, the top box depicts a *false positive (FP)* attack that lowers the familiarity of the known category leading to a *false novelty*. In contrast, the bottom box indicates a *false negative (FN)* attack that increases the familiarity of a known category leading to a *missed novelty* or *false familiarity*.

1. Introduction

In many real-world applications of machine learning models, it is crucial to understand the models' limitations and the trustworthiness of their predictions in novel situations. Thus, we investigate open-set recognition (OSR) [18], which can be seen as a special case of out-of-distribution (OOD) detection [23], where the task is to identify novel categories at test time, which were not included in the training dataset. Recently, Vaze et al. [22] have demonstrated that the progress in OSR performance over the past years is not necessarily due to advancement in OSR approaches, but is correlated with improved performance on the closed-set categories, i.e. the classification of categories included in the training dataset. With this observation, simple baseline scoring rules such as the Maximum Softmax Probability (MSP) [8] and the Maximum Logit Score (MLS) [22, 7] are competitive and perform on par with—or even outperform—more dedicated approaches such as ARPL and ARPL+CS [2], OSRCI [15], and OpenHybrid [24]. At the same time, Dietterich & Guyer [3] have proposed the Familiarity Hypothesis, stating that such familiarity-based scoring rules identify novel categories by measuring the absence of familiar features instead of actively recognizing the presence of novel features. They investigated occlusions as one of the weaknesses of familiarity-based OSR, which can cause false novelty detections. Adversarial attacks pose another potential weakness to familiarity-based OSR, which we study in this work. Dietterich & Guyer [3] mention the risks of adversarial vulnerability in their outlook discussion:

> *"By applying existing attack algorithms (e.g., the FGSM [6]), we predict that it will be very easy to raise the logit score of at least one class and thereby hide a novel class image from novelty detection. It may also be possible to depress the logit scores of enough classes to create false anomaly alarms."*

In other words, this prediction states that it might be trivial to compute adversarial perturbations that amplify familiar features to cause a *false familiarity*, but it might be harder to hide (all) familiar features to yield a *false novelty* (See Fig. 1). While it has been shown that the OSR approach OpenMax [1] is vulnerable to adversarial attacks [19, 20], we study the vulnerability of familiarity-based OSR approaches to gradient-based adversarial white-box attacks (i.e., the model parameters are given) by formulating three main questions:

1. **False Familiarity vs. False Novelty:** What type of attack is more effective?
2. **FGSM vs. iterative attacks:** Is it worth exploring more flexible iterative attacks to improve upon the fast gradient sign method (FGSM)?
3. **Uninformed vs. informed attacks:** How can adversarial attacks profit when the type of input is given (i.e., closed-set or open-set sample)?

2. Methodology

2.1. Familiarity-based open-set recognition (OSR)

We consider an input space $\mathcal{X}$ and a set $\mathcal{F}$ of *familiar* categories, i.e. the closed-set. In closed-set recognition (CSR), the objective is to model the probability $p(y \mid \boldsymbol{x},\, y \in \mathcal{F})$, where y is a label that is associated with the input $\boldsymbol{x} \in \mathcal{X}$. The model is trained on a training dataset $\mathcal{D}_{\text{train}} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{F}$ and evaluate on a non-overlapping closed-set test set, $\mathcal{D}_{\text{test-csr}} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i)\}_{i=1}^{M} \subset \mathcal{X} \times \mathcal{F}$ that contains the categories given at train time. We consider a deep neural network $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^{|\mathcal{F}|}$ parameterized by $\boldsymbol{\theta}$ for modelling $p(y \mid \boldsymbol{x},\, y \in \mathcal{F})$. Here, $f_{\boldsymbol{\theta}}$ maps an input to a vector of logits that are normalized using the softmax function $\sigma : \mathbb{R}^{|\mathcal{F}|} \to (0,1)^{|\mathcal{F}|}$ to obtain pseudo-probabilities for the familiar categories.

In the open-set recognition (OSR) setting a set $\mathcal{N}$ of *novel* categories is additionally considered and a test set containing inputs from both novel and familiar classes is used to evaluate the OSR performance: $\mathcal{D}_{\text{test-osr}} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i)\}_{i=1}^{M} \subset \mathcal{X} \times (\mathcal{F} \cup \mathcal{N})$. A balanced test set containing an equal number of familiar and novel samples is typically used to evaluate the OSR performance. To decide whether $y \in \mathcal{F}$, a familiarity score, $\mathcal{S}(y \in \mathcal{F} \mid \boldsymbol{x})$, is modelled to rank the test samples in $\mathcal{D}_{\text{test-osr}}$. Familiarity-based scoring rules include the *Maximum Softmax Probability (MSP)* score [8]:

$$\mathcal{S}_{\text{MSP}}(y \in \mathcal{F} \mid \boldsymbol{x}) := \max_{y} \sigma(f_{\boldsymbol{\theta}}(\boldsymbol{x}))_y \tag{1}$$

and the *Maximum Logit Score (MLS)* [22, 3]:

$$\mathcal{S}_{\text{MLS}}(y \in \mathcal{F} \mid \boldsymbol{x}) := \max_{y} f_{\boldsymbol{\theta}}(\boldsymbol{x})_y, \tag{2}$$

which has outperformed the MSP score in prior work [22]. For both scoring rules, high scores indicate familiar and low scores indicate novel categories.

2.2. Fast gradient sign method (FGSM)

A simple and effective method for generating adversarial inputs is the *Fast Gradient Sign Method* (FGSM) which was first described by [6]. The FGSM generates an adversarial input, $\boldsymbol{x}^{\text{adv}}$, using the following equation:

$$\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \varepsilon \,\text{sign}[\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}, y)] \tag{3}$$

Here $\boldsymbol{x}$ represents the unmodified input and the second term is known as the *adversarial perturbation*, where ε controls the magnitude of the perturbation.. Initially, $\mathcal{L}$ is set to the training objective [6], but can be any objective function that an adversary aims to optimize. The adversarial perturbation is constrained such that $\|\boldsymbol{x}^{\text{adv}} - \boldsymbol{x}\|_\infty \leq \varepsilon$.

2.3. Iterative attacks

Iterative approaches can generate more diverse perturbations compared to the FGSM by optimizing the objective function in a more flexible manner but at higher computational costs. The *Basic Iterative Method (BIM)* [12] applies the FGSM update iteratively and is described by:

$$\boldsymbol{x}_0^{\text{adv}} = \boldsymbol{x} \tag{4}$$

$$\boldsymbol{x}_{n+1}^{\text{adv}} = \text{Clip}_{\boldsymbol{x},\varepsilon}\{\boldsymbol{x}_n^{\text{adv}} + \alpha \,\text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_n^{\text{adv}}, y))\}. \tag{5}$$

In this method, the step size α and the number of iterations can be adjusted to get the desired trade-off between runtime and performance. Alternative approaches are inspired by gradient-based optimizers using momentum to improve performance [4]. We investigate an iterative approach using RPROP [16, 5] that relaxes the fixed step size α of BIM with an adaptive step size:

$$\boldsymbol{x}_{n+1}^{\text{adv}} = \text{Clip}_{\boldsymbol{x},\varepsilon}\{\boldsymbol{x}_n^{\text{adv}} + \text{Step}(\mathcal{L}, \boldsymbol{\theta}, \boldsymbol{x}_n^{\text{adv}}, y)\} \;\;, \tag{6}$$

where $\text{Step}(\mathcal{L}, \boldsymbol{\theta}, \boldsymbol{x}_n^{\text{adv}}, y)$ denotes the update step computed by some iterative optimization method. RPROP adjusts the step size separately for every optimizable parameter while iterating—in the case of adversarial attacks on images for every pixel per channel. Adversarial perturbations created with RPROP can be sparse and may therefore be less noticeable. For a fair comparison with FGSM, the perturbations are clipped to ε.

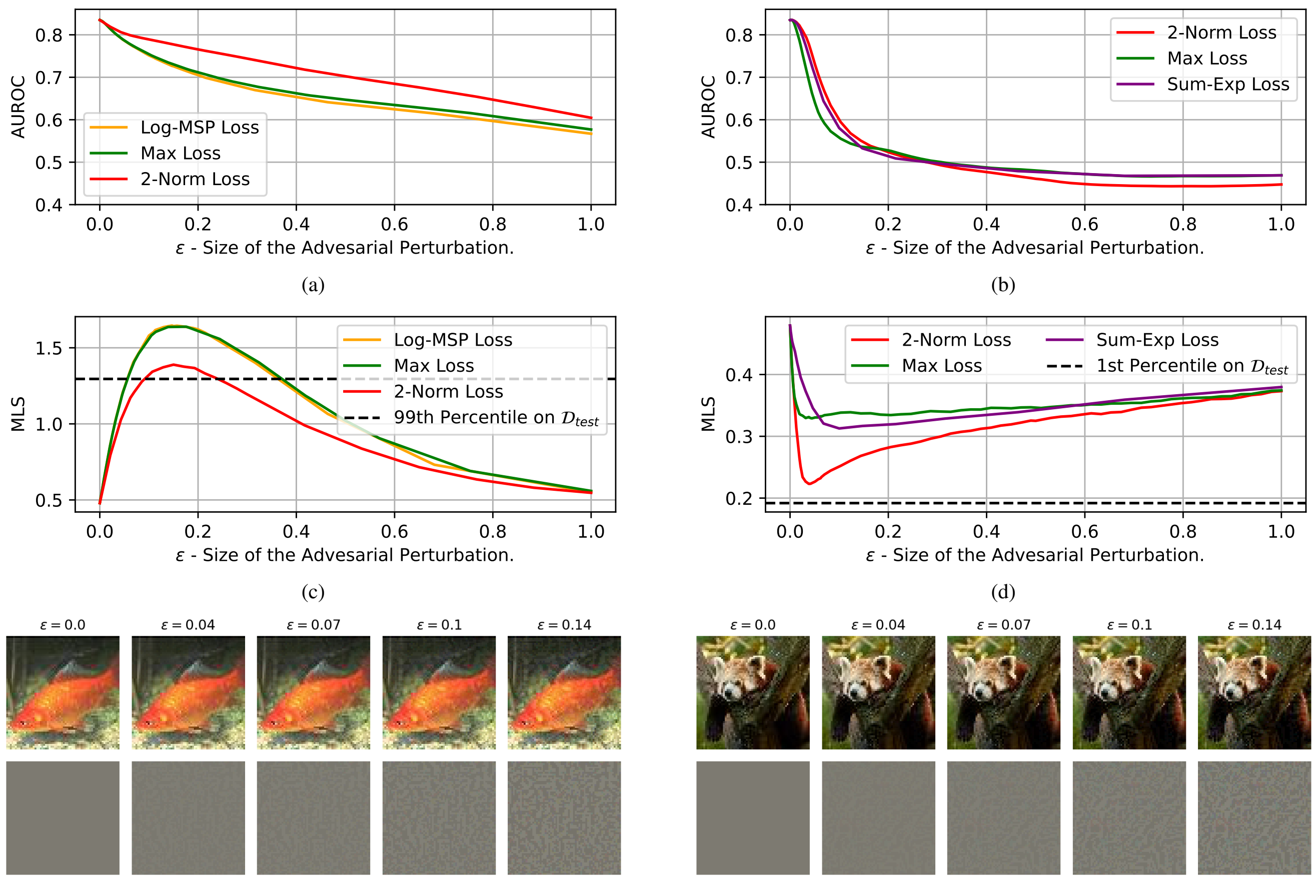
2.4. Adversarial attacks on familiarity-based OSR

False Familiarity (False Negative, FN). False Familiarity attacks aim to *increase* the logit (or softmax probability) of an arbitrary familiar category, which is similar to targeted attacks in closed-set recognition [11]. We investigate three objective functions to achieve this attack:

$$\mathcal{L}_{\text{max}}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \max_{y'} f_{\boldsymbol{\theta}}(\boldsymbol{x})_{y'} \tag{7}$$

$$\mathcal{L}_{\text{2-norm}}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \|f_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 \tag{8}$$

$$\mathcal{L}_{\text{log-MSP}}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \log\, \max_{y'} \sigma(f_{\boldsymbol{\theta}}(\boldsymbol{x}))_{y'} \tag{9}$$



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: **Uninformed FGSM attacks.** Fast gradient sign method (FGSM) attacks on TinyImageNet. Left: False Familiarity (false negative, FN) attacks. Right: False Novelty (false positive, FP) attacks. (a,b) The OSR ranking measured by AUROC. (c,d) Median Maximum Logit Score (MLS) w.r.t. original scores. (e,f) Qualitative examples of adversarial perturbation.

The log-MSP loss has been proposed in the ODIN approach [14] (which was refined in the generalized ODIN [9]) to preprocess images with adversarial perturbations to improve OOD detection using the MSP score.[1]

False Novelty (False Positive, FP). In this likely more challenging setting, we may have to decrease the logits of multiple categories either with a single FGSM step or multiple iterative steps. Objective functions rewarding only the decrease of the largest logit might fail, thus, besides the $\mathcal{L}_{\text{max}}$, we investigate a the $\mathcal{L}_{\text{2-norm}}$ and the sum-exp loss:

$$\mathcal{L}_{\text{sum-exp}}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \sum_{y' \in |\mathcal{F}|} e^{f_{\boldsymbol{\theta}}(\boldsymbol{x})_{y'}} \tag{10}$$

The 2-norm encourages reducing non-maximum logits while still prioritizing the max logit. However, one limitation of the 2-norm is that it is non-negative. Since logits are unnormalized and can be negative, it would be preferable if the objective function also rewarded making the logits negative. This led us to propose the sum-exp loss, which continues to decrease if a logit becomes negative.

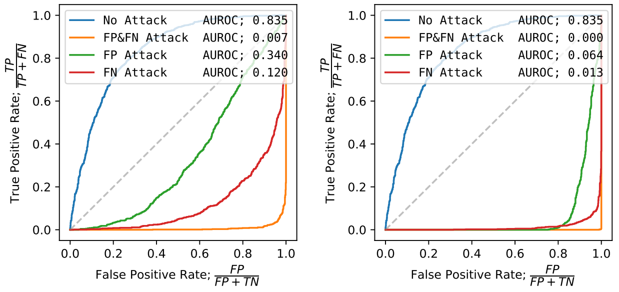
Importantly, while False Familiarity attacks *maximize* these objectives, False Novelty attacks *minimize* them.

Uninformed vs. informed attacks. We call an attack *informed* if the adversary has access to the binary set-labels of the input, i.e. closed-set vs. open-set, and *uninformed* if that information is not available [10]. In the uninformed setting, either a FP or FN attack is applied on all images, disregarding whether an image is novel or familiar. For informed adversaries, FN attacks are performed on novel images and FP attacks on familiar images only.

3. Experimental results

We experiment with the TinyImageNet dataset [13], described as one of the most challenging benchmarks used in the OSR literature [22]. Here we use the open-set split presented in Vaze et al. [22] and follow their experimental setup. TinyImageNet consists of a subset of 200 ImageNet

[1] While this is not further investigated in this work, our OSR experiments did not confirm an improvement of the MSP score as also mentioned in other independent work [3].



(a) FGSM

(b) Iterative attack

Figure 3: **Informed attacks.** False Positive (FP) attacks are performed on familiar samples (2-norm loss) and False Negative (FN) attacks on novel samples (max loss). (a) Fast gradient sign method (FGSM) using $\varepsilon = 0.07$ for FN and $\varepsilon = 0.06$ for FP. (b) Our iterative method with $\varepsilon = 0.07$ for FN and $\varepsilon = 0.04$ for FP attacks.

categories [17], whereas 20 classes are used as the closed-set training dataset and 180 classes as the open-set. The CNN architecture used is a VGG32[2], a lightweight version of the VGG architecture [21]. This results in a reproduced closed-set accuracy of 84.2% averaged over five class splits.

We report the OSR performance of the MLS for the first of the five splits with the area under the Receiver-Operator curve (AUROC). The AUROC is a threshold-less metric that evaluates the ranking from open-set to closed-set samples. As a higher AUROC means better OSR performance, adversarial attacks aim to lower the AUROC.

What type of attack is more effective? It depends. In the *uninformed* FGSM experiments, False Novelty (False Positive, FP) attacks are more effective in destroying the ranking, i.e. decreasing the AUROC, than False Familiarity (False Negative, FN) attacks at the same magnitude ε of adversarial perturbation (Fig. 2a, 2b). However, we observe the opposite in the *informed* setting (Fig. 3), where the AUROC of FN attacks is lower than FP attacks. To understand this behaviour, we look at the distribution of scores before and after the attacks.

It is too easy to raise the logit score. Or in other words, to amplify familiar features. While FN attacks aim to amplify familiar features of the open-set to cause a missed novelty, in contrast, FP attacks aim to hide familiar features to reduce the familiarity of closed-set categories. We recall that uniformed attacks are performed on both novel and familiar images. Even though FN attacks can increase the median MLS above the 99th percentile of the original test data scores (Fig. 2c), the AUROC is rather preserved (Fig. 2a). Hence, the FN attacks not only increase the familiarity (i.e.; MLS) of the novel but also of the familiar samples, which preserves the ranking. In contrast, FP attacks cannot decrease the median MLS below the 1st percentile of the original test scores, but the ranking (AUROC) is effectively destroyed. This suggests that FP attacks tend to decrease the scores of the closed-set more than the scores of the open-set. Our experiments confirm the prediction of Dietterich & Guyer [3] that it is very easy to raise the logit score, which only leads to effective FN attacks in the informed setting. However, our results reveal that for uninformed attacks the ability to easily raise the logit score is not the key to attack the ranking of familiarity-based OSR approaches.

Which objective function performs best? While some objective functions are able to perform both types of attacks by swapping the sign, no objective is clearly best for both FN and FP attacks (Fig. 2a, 2b). FGSM FN attacks achieve lowest AUROC using the Log-MSP loss and second lowest with the Max loss. For FP attacks the Max loss achieves the lowest AUROC with $\epsilon < 0.1$. Whereas at $\epsilon \approx 0.3$ all objective functions achieve an AUROC of $\approx$0.5, for $\epsilon > 0.3$ the 2-Norm achieves even lower AUROC.

FGSM vs. iterative attacks. Informed iterative attacks are able to decrease the AUROC by an order of magnitude compared to informed FGSM attacks using the same or even smaller ε (Fig. 3). The AUROC for FP attacks is decreased from 0.34 (FGSM) to 0.06 (iterative) and for FN attack from 0.12 (FGSM) to 0.01 (iterative).

Informed attacks reverse the ranking almost perfectly. Informed FGSM attacks can improve substantially over uninformed attacks. Informed FGSM and iterative attacks are able to reverse the ranking of novel and familiar images almost perfectly when using both FP attacks on familiar and FN attacks on novel samples together (Fig. 3).

4. Conclusion

We have studied the vulnerability of familiarity-based OSR approaches to adversarial attacks. Our MLS experiments confirm Dietterich & Guyer's [3] prediction that the logit score can be easily increased with an adversarial perturbation. However, this ability leads only to effective False Familiarity (FN) attacks in the informed setting. In an uninformed setting, FN attacks are less effective than FP attacks that, in contrary, are able to successfully destroy the ranking by hiding familiar features of closed-set categories. The uninformed setting may be informative for the development of new scoring rules. It remains to be tested if the observed adversarial robustness holds for alternative familiarity scores, such as the MSP. We hope that our findings can contribute to the design of better scoring rules in the future and to make familiarity scores robust to adversarial attacks.

[2] We use the model weights published on: https://github.com/sgvaze/osr_closed_set_all_you_need (accessed 2023-05-23).

Acknowledgement

N.L., S.B., and C.I. acknowledge support by the Pioneer Centre for AI, DNRF grant number P1.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 2

[2] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 1

[3] Thomas G. Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, 2022. 1, 2, 3, 4

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193. IEEE, 2018. 2

[5] Ciprian Florescu and Christian Igel. Resilient backpropagation (Rprop) for batch-learning in TensorFlow. In *International Conference on Learning Representations (ICLR) Workshop*, 2018. 2

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 8759–8773, 17–23 Jul 2022. 1

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[9] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10951–10960, 2020. 3

[10] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011. 3

[11] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations (ICLR)*, 2017. 2

[12] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[13] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. In CS231N, 2015. 3

[14] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations (ICLR)*, 2018. 3

[15] Lawrence Neal, Matthew L. Olson, Xiaoli Z. Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision (ECCV)*, volume 11210 of *LNCS*, pages 620–635, 2018. 1

[16] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks (ICNN)*, pages 586–591. IEEE, 1993. 2

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 4

[18] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. 1

[19] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. In *European Conference on Computer Vision (ECCV)*, pages 682–698, 2020. 2

[20] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense with clean-adversarial mutual learning. *International Journal of Computer Vision*, 130(4):1070–1087, 2022. 2

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015. 4

[22] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 3

[23] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021. 1

[24] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, volume 12348 of *LNCS*, pages 102–117, 2020. 1