Circumventing and Hardening Stateful Detection of Query-Based Attacks

Eden Biran
Amnon Levi
edenbiran@mail.tau.ac.il
amnonlevi@mail.tau.ac.il
Blavatnik School of Computer Science, Tel Aviv University
Tel Aviv, Israel

ABSTRACT

Machine Learning models are known to be vulnerable to adversarial examples. These adversarial examples can be created even in a black-box setting without knowledge of the target model's architecture, weights, or parameters. The creation process is based on a series of queries that are highly similar in the input space. Therefore, in order to defend against these query-based attacks, defenders use an internal state to save historical queries and then detect and reject those that seem too similar.

In this paper, we introduce a novel improvement to existing query-based attacks. Our improvement helps reduce the similarity between consecutive queries, which in turn allows us to circumvent stateful defenses. We demonstrate a successful circumvention of the existing SoTA detection method while maintaining a high attach success rate.

Additionally, in order to harden stateful defenses, we propose a novel defense based on a deep neural network that classifies sequences of queries. Our defense differs from existing defenses that are based on the similarity between a single input and a set of historical queries. Our proposed defense achieves equal or better results when compared to existing stateful defenses on multiple different attack types and configurations.

1 INTRODUCTION

Over the past few years Machine Learning models have achieved and sometimes even surpassed human abilities in a variety of domains. As a result of these performances, such models are now being used in safety-critical applications such as autonomous driving, medical image classification, and identity verification.

Despite security and trustworthiness being paramount in these settings, such models are known to be vulnerable to adversarial examples. An adversarial example is maliciously modified input that is imperceptible from a benign input but is misclassified by the model. Adversarial attacks can be broadly divided into white-box and black-box attacks.

In the white-box setting the attacker has full access to the target model, including its architecture, weights, and parameters. This intimate access allows attackers to execute powerful and practical attacks [5, 13, 27, 28, 30]. In turn, these attacks inspired multiple defenses [10, 15, 24, 26, 29, 35] aiming to either prevent the generation of adversarial examples or hinder their success during inference time. Although a multitude of defenses has been explored and implemented, nearly all have been proven to be vulnerable to improved and adaptive attacks [1, 2, 4, 31].

In contrast to the white-box setting, the black-box setting assumes a more realistic attack scenario. In the black-box setting the attacker has access exclusively to the model's output, which could be either a distribution over the labels or only the output label. This kind of interaction is popular with Machine-Learning-as-a-Service (MLaaS) platforms [9, 14] where a model is hosted remotely and exposed via a limited API. Black-box attacks can further be divided into two main types: transferability-based and query-based.

In transferability-based attacks [11, 22, 36] an attacker creates adversarial examples against a local surrogate model (using a white-box attack) in the hope that they will transfer and be effective against the target model as is. This type of attack is addressed by defenses such as ensemble adversarial training [32, 34] that aim to reduce the likelihood of adversarial examples transferring.

In query-based attacks [3, 6, 7, 18, 19, 25] an attacker can use multiple queries to estimate the local loss landscape around a certain input. Using this loss estimation primitive, an attacker can then iteratively craft an adversarial example in a similar manner to white-box attacks.

Defenses against query-based attacks belong to two groups. The first, which has proven to be difficult [4], tries to statelessly detect whether any single input is malicious. The second is a stateful defense which takes advantage of the attacker's need to sample many similar points in order to estimate the local loss landscape. Therefore, defenders can use an internal state to save historical queries and then take advantage of the similarity between the adversary's consecutive queries in order to detect and reject those who seem too similar. These stateful defenses have seen partial success [8, 21], but ultimately fail to defend against a recently improved adaptive attack [12].

In this work, we first introduce a novel improvement to existing query-based attacks based on changing the image such that it is blended with a different image of the same class in every iteration. This approach drastically reduces the similarity between consecutive queries, thus circumventing current stateful defenses.

Second, we propose a novel stateful defense, which uses a deep neural network to classify sequences of queries that are part of the generation of adversarial examples. This approach differs from other defenses [8, 21] who instead check for the similarity between a single input and a set of historical queries. Our proposed defense achieves equal or better results when compared to two existing defenses (OSD [8] and Blacklight[21]) on two attack types (NES [18] and Boundary [3]) with multiple attack configurations each. We further evaluate the defense on our attack proposed in section

3, successfully defending against it while outperforming existing defenses.

2 RELATED WORK

2.1 Black-Box Attacks

NES. Ilyas et al. [18] introduce a score-based black-box attack based on two stages. First, gradients are estimated using finite differences over samples near the current image (using a process inspired by natural evolution strategies [33]). Second, the image is moved in the direction of this estimated gradient via projected gradient descent in order to increase the loss. These stages are then repeated until a sufficient adversarial image is found.

Boundary. Brendel et al. [3] introduce a label-based black-box attack. The attack starts from a randomly initialized adversarial image and then iteratively adjusts the image using two stages. First, the decision boundary around the current image is estimated by sampling nearby images. Second, sampled images that are still adversarial are moved in the direction of the original image. Finally, the adversarial image which is closest to the original image is selected for the next iteration. These stages are then repeated until a sufficient adversarial image is found.

2.2 Stateful Defenses

OSD. Chen et al. [8] introduce a new paradigm for defending against adversarial examples by shifting from stateless to stateful defenses. They propose training an encoder network to lower the dimensionality of the inputs such that adversarial variations of the same input have similar latent vectors. The encoded vector is then compared to a buffer of encodings of previous queries by the same user, to calculate the mean k nearest neighbors distance. If the distance is lower than some (training data-based) threshold, the user is flagged as malicious and counter-measures can be taken (e.g. banning the user). They demonstrate the robustness of the defense against both NES and boundary attacks, by providing a lower bound for the number of MLaaS accounts that need to be created in order to circumvent it. Due to it pioneering the field of stateful defenses, it has been referred to as the Original Stateful Defense (OSD).

Blacklight. Li et al. [21] observed some weaknesses in OSD, the most fundamental of which is the measured presence of Sybil attacks, which completely nullifies the effectiveness of OSD's per-user approach. They additionally make note of the computational infeasibility of computing the k nearest neighbors across all users of an MLaaS server. To address these issues, they propose Blacklight, a stateful defense scheme that efficiently detects highly similar adversarial queries, while avoiding falsely flagging benignly similar queries, such as different frames of the same video.

For the detection phase, they transform the input by adding a random salt image of the same dimensionality, followed by quantization. the quantization process serves to both convert the inputs to a hashing-friendly integer vector, as well as to increase the similarity between adversarial queries, which typically only feature minor variations. The added salt image serves the purpose of mitigating attempts by the attacker to reverse engineer the defense parameters.

The quantized input is then flattened, and a large fingerprint vector is created by an overlapping sliding window of a one-way hash function. The large fingerprint vector is then reduced to the top S hash values, a process that reduces its dimension significantly, in an unpredictable (due to the nature of hash values), yet purely input-based fashion. This final vector serves as the query's fingerprint. For each query, the calculated fingerprint is stored into a buffer, and compared to all previously stored fingerprints. If sufficient overlap is detected with a past fingerprint, the query is flagged as an attack. The Authors claim that Blacklight is even resistant to two evasion methods: Gaussian noise adding and image augmentation.

Radial Attacks. Feng et al. [12] show that both OSD and Blacklight can be circumvented by simply adjusting the hyperparameters of the attacks, and propose a generalization of both that successfully defends against this simple attack variation. Additionally, they propose an improvement to Blacklight that calculates the fingerprint from a learned neural hash rather than a pixel-based hash function. A key assumption of current stateful detection defenses is the similarity between attack queries. Radial attacks challenge this assumption by finding the similarity "radius" and adaptively making changes large enough to avoid entering it. Feng et al. introduce a couple of new attacks that are not successfully defended against by either OSD or Blacklight. NESRadial and BoundaryRadial are variations on NES and boundary attacks (respectively). NESRadial circumvents the defenses by employing two adaptive tactics. *I)* Searching for the minimal standard deviation of added Gaussian noise necessary to avoid detection. *II)* Adaptively changing the gradient descent, doubling it when multiple detections are observed, and halving it when no actions are observed. BoundaryRadial employs its own pair of adaptive tactics. *I)* Adjusting the magnitude of the Gaussian sampling step, increasing it when detections are frequent and vice versa. *II)* If detections are observed, the final step of moving the adversarial samples closer to the original image is skipped.

3 CIRCUMVENTING STATEFUL DETECTION

In this section, we introduce a method of masking an attack and demonstrate its capabilities using a label-only NES attack, as laid out by Ilyas et al. [18].

As stated, the key assumption of the stateful defenses is the similarity between queries. We show that an attack can be masked such that it retains its success rate within a margin of error while circumventing a defense that would otherwise detect it. This masking decreases the similarity between the queries, thus challenging said key assumption.

3.1 Attack Method

We formulate an attack for a tensor X of class c as an iterative process searching for an adversarial tensor X_{adv} identified as class $c' \neq c$ (untargeted) by using $\{Q_k\}$ queries.
Given an attack query Q_k and a query retention rate $\rho \in [0, 1]$, instead of querying with Q_k, we instead use:

$$Q'_k = M_{\rho,k} \cdot Q_k + (1 - M_{\rho,k}) \cdot N_k$$

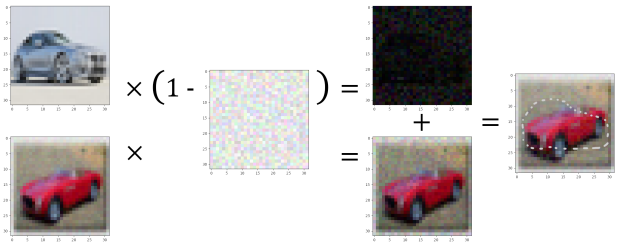

Figure 1: interpolation of benign training data N_k (top left) with the adversarial query Q_k (bottom left) using a noisy mask. The area darkened by the benign image is marked on the final query Q'_k (right-most) for better clarity.

Where $M_{\rho,k}$ is the masking rate tensor and N_k is the masking noise. both are of the same dimensions as Q_k. The gradient estimate is also multiplied by $M_{\rho,k}$ and normalized accordingly.

We found that the best overall results are obtained for:

$$(M_{\rho,k})_{i,j} \sim Uniform(2\rho - 1, 1)$$

$$N_k \sim CIFAR\text{-}10_{Train}(c)$$

i.e. the noise is sampled from the same domain as X. In our case, from the same dataset and the same class class (but from the training data rather than the testing data). The reasoning behind the choice is that if some elements of the adversarial attack are generalized across the class, they will be less noised by this method.

See Figure 1 for an illustration of the method for a given query Q_k.

When using antithetic sampling, we use what we call antithetic noising. For a pair of queries Q_k^+, Q_k^-:

$$Q_k'^{+} = M_{\rho,k} \cdot Q_k^+ + (1 - M_{\rho,k}) \cdot N_k$$

$$Q_k'^{-} = M_{\rho,k} \cdot Q_k^- - (1 - M_{\rho,k}) \cdot N_k$$

Or equivalently:

$$N_k^- = -N_k^+;\ M_{\rho,k}^- = M_{\rho,k}^+$$

3.2 Evaluating The Attack

In order to evaluate the attack, we used images from CIFAR-10 [20]. The adversarial examples were taken from the test dataset and created against a RESNET-20 [16] model pre-trained on the CIFAR-10 training dataset. The noise images used in the masking were taken from the training set as well.

We tested the capabilities of the attack against Blacklight [21], which is the current state-of-the-art detection method, as explored by Feng et al. [12].

We test on a small-scale attack (160,000 queries) with different Blacklight detection threshold values to demonstrate the effect of the retention rate (denoted as ρ) on the trade-off between the circumvention and attack success rates (Table 1). Keeping up with Feng et al. [12], all of the attacks were made with $\sigma = 0.05$. It's important to note that the tested τ values are unrealistic for a deployed Blacklight detector, as they cause an undesired amount of false positives [12, 21] and are only used to simulate a very strict similarity-based detector.

Table 1: Detection rates of a small-scale NES attack for different Blacklight τ values.

Attack	Hyperarameters	Success rate	$\tau = 0.9$	$\tau = 0.95$	$\tau = 0.975$
NES	$\sigma = 0.05$	89.4%	0.28%	94.3%	93.9%
Masked NES	$\rho = 0.975$	88.1%	0.01%	5.2%	71.1%
	$\rho = 0.95$	82.2%	0.0%	0.3%	19.2%
	$\rho = 0.925$	63.0%	0.0%	0.1%	0.9%

Table 2: Detection rates of a large-scale NES attack with and without masking

Attack	Hyperparameters	Success rate	Detection rate
NES	$\sigma = 0.001$	90%	82.2%
Masked NES	$\rho = 0.99$	75%	0.78%

We also test for a large-scale attack, with the label-only parameters used in the original NES paper (50,000,000 query pairs and $\sigma = 0.001$)[18] against the original Blacklight configuration [21]. We use $\rho = 0.99$ for the masked attack, demonstrating the effect of even small masking in such an attack.

3.3 Discussion

In this section, we have shown that without changing the attack itself, it can still work to different extents when the queries are partially noised. This undermines the current paradigm of stateful defenses, regardless of the choice of defense.

While we presented the *overall* best method we found, some variations yield better results in different aspects. For example, different use cases call for different detection tolerances: a single-user attack against a banning policy detector might require perfect avoidance, even at the price of a significantly lowered success rate.

We have also experimented with using the same iterative method used for finding the adversarial image, to learn the best mask for it. The method requires vastly augmenting the dataset. For that reason, it is very slow and thus was not thoroughly tested. It is accessible through the project code. It showed a qualitative promise.

4 HARDENING STATEFUL DETECTION

In this section, we introduce a novel stateful defense paradigm against black-box query-based attacks. Our defense is based on a Deep Neural Network trained to classify if a sequence of queries is benign or part of the generation of an adversarial example. The proposed defense's classification is based on a set of historical queries, setting it apart from defenses such as OSD and Blacklight, whose classification is based on the similarity of a single new query to a set of historical queries. The proposed defense paradigm allows the classifier a higher degree of freedom, which we show translates in to a robust and accurate stateful defense. We name our defense Video Classifier.

4.1 Attack Dataset

In order to train and evaluate the video classifier, a dataset containing benign and adversarial sequences of queries was created. Each sequence is encoded as a video file where each frame corresponds to an image query. A benign sequence contains a series of (sometimes partially) random images, while an adversarial sequence contains a series that is part of the generation of an adversarial example. All images were taken from CIFAR-10 [20] and the adversarial examples were created against a RESNET-20 [16] model pre-trained on CIFAR-10.

The benign half of the dataset is split in to two subsets. The first subset contains sequences of completely random images, simulating unrelated queries from multiple users. The second subset contains sequences of random images which all have the same class, simulating a user attempting to predict a specific class. This second subset also aims to aid the model in correctly identifying adversarial attacks while achieving a low false positive rate on sequences of similar images belonging to the same class.

The adversarial half of the dataset is split in to multiple subsets, where each subset is created using a unique attack configuration. An attack configuration is defined as a specific black-box attack (such as NES or Boundary) accompanied by the attack specific parameters (such as σ for NES or ϵ, δ for Boundary). Moreover, each subset is created with a certain spacing between adversarial images. We define a spacing of n to mean that between each pair of adversarial images a random amount of $0 - n$ benign images are inserted.

4.2 Video Classifier

Our classifier is a DNN trained to classify sequences of images encoded as a video file. The length of each sequence was chosen to be 300 images which are formatted as a 10 second video at 30 frames per second. This amount of images was chosen as it is significantly less than the amount of queries needed to fully create an adversarial image, which can amount to anywhere between thousands and millions of queries [3, 18].

The architecture of the classifier is made up of three main building blocks. First, a ConvNeXt-tiny [23] model fine-tuned on CIFAR-10 is applied to each frame of the input. Then, the output of the previous stage is passed through a 2 layer LSTM [17] with a hidden size of 128. Finally, the LSTM's output is fed through a classification head composed of a single linear layer with softmax activation.

The classifier was trained for 50 epochs on a subset of the attack dataset containing 1000 benign sequences and 1000 adversarial sequences. The adversarial sequences were all created using NES with $\sigma = 0.001$. As the input of the video classifier is a sequence of 300 images, which is less than the length of a full attack, each full generation of an adversarial example was randomly sampled for a single short and continuous sequence of queries.

4.3 Evaluating The Video Classifier

The trained video classifier was evaluated on multiple distinct subsets of the attack dataset. Each subset contains 500 sequences made up of 250 benign sequences and 250 adversarial sequences.

The specific attack configurations can be seen in Table 3. The base attack configurations are the same as in the original work on OSD and Blacklight. Additionally, we evaluate the video classifier on the attack configurations used by Feng et al. [12], as these configurations were shown to circumvent the aforementioned defenses. Finally, we also evaluate the video classifier against the Masked NES attack described in section 3.

Table 3: Accuracy of our video classifier compared to existing defenses on different attack configurations. Higher accuracy implies the defense was more successful in detecting the attack. The accuracy of OSD and Blacklight is taken from previous research [8, 12, 21].

Attack	Hyperparameters	OSD	Blacklight	Video Classifier
NES	$\sigma = 0.001$	15.6%	82.2%	99.8%
	$\sigma = 0.01$	6.6%	6.4%	99.8%
	$\sigma = 0.05$	5.6%	6.8%	99.8%
Boundary	$\epsilon = 1, \delta = 0.1$	–	100%	100%
	$\epsilon = 0.01, \delta = 0.01$	100%	91.0%	100%
Masked NES	$\rho = 0.95$	–	0.00%	100%
	$\rho = 0.925$	–	0.00%	99.4%

The main drawback of our proposed type of defense is its vulnerability to the spacing out of images that are part of the generation of an adversarial example. This is due to the classification of a set of successive queries together, unlike defenses that compare queries in a pairwise fashion making them indifferent to the spacing between adversarial images.

In order to evaluate the effect of spacing the adversarial images, multiple additional video classifiers were trained and evaluated on attacks with different spacing values. These video classifiers have an identical architecture to that described in section 4.2. Likewise, they were all trained on adversarial sequences created using NES with $\sigma = 0.001$. The only difference between them is the spacing value of the adversarial sequences in their training data. An additional model was trained on data containing an equal mixture of sequences of the different spacings.

4.4 Results

The results of the evaluation and a comparison to existing defenses can be seen in Table 3. We note the high success rate (above 99%) of our proposed defense, which is better or equal to that of existing defenses. Additionally, the results show that while existing defenses are vulnerable to the modification of attack parameters, the video classifier is robust to such changes.

Moreover, the video classifier exhibits high transferability and generalization. Recall that the video classifier was trained solely on adversarial sequences generated from NES attacks with $\sigma = 0.001$. Despite this relatively narrow training data, the video classifier achieves high accuracy both on the same attack type with different parameters and on different and unseen attack types.

Furthermore, in the performed evaluation, the video classifier has a single opportunity to classify an adversarial sequence. This single chance consists of classifying a randomly sampled sequence of length 300 out of the complete generation of an adversarial example. Even under this extremely restricting condition, the video

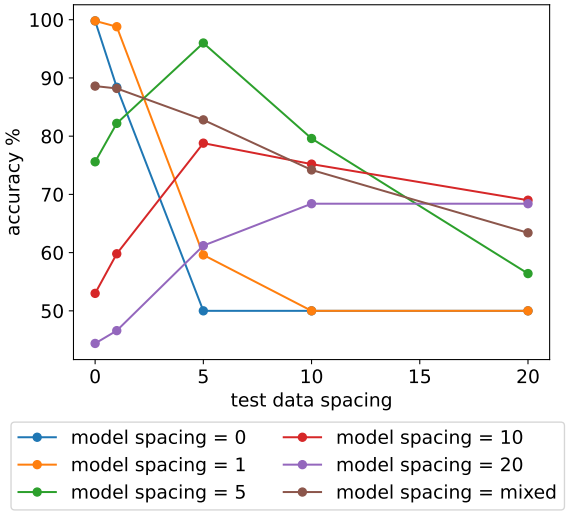

Figure 2: Accuracy of multiple video classifiers trained on different adversarial sequence spacings. Each model is evaluated on test data subsets with different adversarial sequence spacings. Higher accuracy implies the defense was more successful in detecting the attack.

classifier outperforms existing defenses that inspect the complete generation of an adversarial example.

The results also show that the video classifier successfully differentiates adversarial attacks from potential "false positive" sequences containing images of the same class. Therefore, we can conclude that the video classifier is not classifying solely on the fact that a sequence contains similar images, but that it has managed to develop some notion of the creation process of an adversarial image.

Figure 2 depicts the relation between the video classifier's accuracy and the spacing of the adversarial images. Unsurprisingly, models generally achieve the highest accuracy on test data with the same spacing as their training data. Additionally, a model's accuracy is higher the closer the test spacing is to the train spacing, showing some generalization beyond the training data. Unfortunately, the accuracy does drop relatively quickly as the spacing difference grows larger. Despite this behavior, we note that the model trained on a mixture of spacings shows a relatively consistent accuracy, ranking in the top-3 models across all spacings. This gives cause to believe that with a larger training set and a careful mixture of spacings, a single model could be trained to achieve high accuracy across all spacings.

4.5 Discussion

In this section, we proposed what we believe to be a promising novel paradigm for the stateful detection of black-box attacks. In order to evaluate this paradigm, a specific video classifier architecture was chosen and trained on a specific composition of training data. We have shown that this instance of a video classifier performs better than previously known defenses. We also show that the classifier can be trained to detect attacks interspersed within benign data.

We note that these specifics are highly configurable and we hypothesize that changes to them could further increase performance, depending on the exact use case. In particular, we believe that changes to the DNN architecture, query sequence length, and training data composition (such as using multiple attack configurations and adversarial image spacings) would have the largest impact.

Such changes could also help address a potential drawback of the proposed method, which is the relatively high computational cost of the classifier compared to that of previously proposed defenses (such as OSD and Blacklight). Changes to the video classifier's architecture and parameters (such as input sequence length) would enable the creation of a video classifier tailor-made to the needs and computational power of a particular system.

In order to further strengthen the defense while also creating a system that could be deployed in a real-world scenario, the input to the video classifier would be a sliding window of the recent queries. This would allow the video classifier multiple chances to correctly detect an attack sequence, further strengthening the defense. Additionally, this window could be configured to be either unique per account or global across all accounts, in order to combat the adversarial creation of images by multiple accounts belonging to the same adversary.

5 CONCLUSION

Machine learning models are known to be vulnerable to black-box query-based adversarial attacks. These attacks are especially dangerous, as they assume a realistic attack scenario using only a model's output. These attacks led to the proposal of multiple stateful defenses aiming to recognize and block such attacks.

In this work, we first introduce a novel improvement to existing query-based attacks based on the dominance of adversarial hyperdirections and their resistance to noise. Our attack manages to circumvent existing defenses, showing that they are still lacking.

Second, we introduce a novel defense based on a deep neural network that classifies sequences of queries, unlike existing defenses which are based on observing the similarity between a single input and a set of historical queries. Our defense achieves equal or better results when compared to existing stateful defenses on multiple different attack types and configurations, including that introduced in this work. We believe that our proposed defense could be extended in order to further improve performance and tailor it to real-world use cases.

CONTRIBUTIONS

Amnon Levi led the research on circumventing stateful detection and Eden Biran led the research on hardening stateful detection. All other shared work was split equally.

REFERENCES

[1] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*. PMLR, 274–283.

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*. PMLR, 284–293.

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2017. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models.

arXiv preprint arXiv:1712.04248 (2017).

[4] Nicholas Carlini and David Wagner. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 3–14.

[5] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.

[6] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*. IEEE, 1277–1294.

[7] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 15–26.

[8] Steven Chen, Nicholas Carlini, and David Wagner. 2020. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*. 30–39.

[9] Clarifai. [n. d.]. *The world's AI: Clarifai computer vision AI and machine learning platform*. https://www.clarifai.com/

[10] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. 2018. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442* (2018).

[11] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4312–4321.

[12] Ryan Feng, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, Somesh Jha, and Atul Prakash. 2023. Investigating Stateful Defenses Against Black-Box Adversarial Examples. *arXiv preprint arXiv:2303.06280* (2023).

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[14] Google. [n. d.]. *Google Cloud Vision AI*. https://cloud.google.com/vision/

[15] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 630–645.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*. PMLR, 2137–2146.

[19] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978* (2018).

[20] Alex Krizhevsky et al. 2009. Learning multiple layers of features from tiny images. (2009).

[21] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. 2022. Blacklight: Scalable Defense for Neural Networks against {Query-Based}{Black-Box} Attacks. In *31st USENIX Security Symposium (USENIX Security 22)*. 2117–2134.

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[25] Seungyong Moon, Gaon An, and Hyun Oh Song. 2019. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *International Conference on Machine Learning*. PMLR, 4636–4645.

[26] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 582–597.

[27] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 1528–1540.

[28] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2019. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)* 22, 3 (2019), 1–30.

[29] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. 2017. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766* (2017).

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[31] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems* 33 (2020), 1633–1645.

[32] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* (2017).

[33] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. 2014. Natural evolution strategies. *The Journal of Machine Learning Research* 15, 1 (2014), 949–980.

[34] Eric Wong, Leslie Rice, and J Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994* (2020).

[35] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2017. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991* (2017).

[36] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2730–2739.