

Enhancing Cross-task Transferability of Adversarial Examples with Dispersion Reduction

Yunhan Jia*, Yantao Lu*†, Senem Velipasalar†, Zhenyu Zhong, Tao Wei

Baidu X-Lab, †Syracuse University

{jiayunhan, yantaolu, edwardzhong, lenx}@baidu.com, svelipas@syr.edu

Abstract

Neural networks are known to be vulnerable to carefully crafted adversarial examples, and these malicious samples often transfer, i.e., they maintain their effectiveness even against other models. With great efforts delved into the transferability of adversarial examples, surprisingly, less attention has been paid to its impact on real-world deep learning deployment.

In this paper, we investigate the transferability of adversarial examples across a wide range of real-world computer vision tasks, including image classification, explicit content detection, optical character recognition (OCR), and object detection. It represents the cybercriminal's situation where an ensemble of different detection mechanisms need to be evaded all at once.

We propose practical attack that overcomes existing attacks' limitation of requiring task-specific loss functions by targeting on the "dispersion" of internal feature map. We report evaluation on four different computer vision tasks provided by Google Cloud Vision APIs to show how our approach outperforms existing attacks by degrading performance of multiple CV tasks by a large margin with only modest perturbations ($l_\infty \leq 16$).

1. Introduction

Recent research in adversarial learning has brought the weaknesses of deep neural networks (DNNs) to the spotlights of security and machine learning studies. Given a deep learning model, it is easy to generate adversarial examples (AEs), which are close to the original but are misclassified by the model [9, 24]. More importantly, their effectiveness sometimes *transfer*, which may severely hinder DNN based applications especially in security critical scenarios [16, 10, 25]. While such vulnerabilities are alarming, little attention has been paid on the realistic threat model of commercial or proprietary vision-based detection sys-

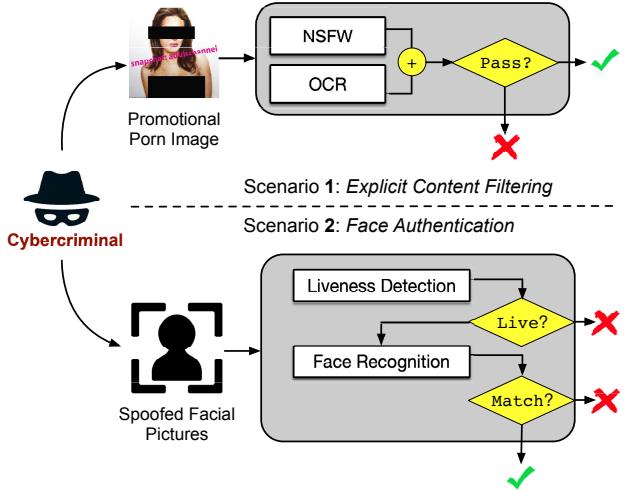


Figure 1. Real-world computer vision systems deployed in safety- and security-critical scenarios usually employ an ensemble of detection mechanisms that are opaque to attackers. Cybercriminals are required to generate adversarial examples that transfer across tasks to maximize their chances of evading the entire detection systems.

tems against real-world cybercriminals, which turn out to be quite different from those intensively studied by aforementioned research.

Deployment. Computer vision (CV) based detection mechanisms have been extensively deployed in security-critical applications such as content censorship and authentication with facial biometrics, and readily available services are provided by cloud giants through APIs (e.g., Google Cloud Vision [3], Amazon Rekognition [1]). The detection systems have long been targeted by evasive attacks from cybercriminals, and it has resulted in an arm race between new attacks and more advanced defenses.

Ensemble of different detection mechanisms. To overcome the weakness of deep learning in individual domain, real-world CV systems tend to employ an ensemble of different detection mechanisms to prevent evasions. As shown in Fig. 1, underground businesses embed promotional con-

*Equal contribution

tents such as URLs into porn images with sexual content for illicit online advertising or phishing. A detection system combines Optical Character Recognition (OCR) and image-based explicit content detection can thus drop posted images containing either suspicious URLs or sexual content to mitigate evasion attacks. Similarly, a face recognition model that is known to be fragile [22] is usually protected by a liveness detector to defeat spoofed digital images when deployed for authentications. Such ensemble mechanisms are widely adopted in real-world CV deployment.

To evade detections with uncertain mechanisms, attackers turn to generate adversarial examples that transfer across CV tasks. Many adversarial techniques on enhancing transferability have been proposed [26, 25, 16, 10]. However, most of them are designed for image classification tasks, and rely on task-specific loss function (e.g., cross-entropy loss), which limits their effectiveness when transferred to other CV tasks.

In this paper, we propose a simple yet effective approach to generate adversarial examples that transfer across a broad class of CV tasks, including classification, object detection, explicit content detection and OCR. Our approach called *Dispersion Reduction (DR)* as shown in Fig. 2, is inspired by the impact of “contrast” on an image’s perceptibility. As lowering the contrast of an image would make the objects indistinguishable, we presume that reducing the “contrast” of internal feature map would also degrade the recognizability of the subjects in the image, and thus could evade CV-based detections. We use *dispersion* as a measure of “contrast” in feature space, which describes how scattered a set of data is. We empirically validate the impact of dispersion on model predictions, and find that reducing the dispersion of internal feature map would largely affect the activation of subsequent layers. Based on another observation that lower layers detect simple features [15], we hypothesis that the low level features extracted by early convolution layers share many similarities across CV models. Thus the distortions caused by dispersion reduction in feature space, are ideally suited to fool any CV models, whether designed for classification, object detection, OCR, or other vision tasks.

We evaluate our proposed DR attack on both popular open source models and commercially deployed detection models. The results on four Google Cloud Vision APIs: classification, object detection, SafeSearch, and OCR (see §4) show that our attack causes larger drops on the model performance than state-of-the-art attacks (MI-FGSM [10] and DIM [25]) by a big margin of 11% on average across different tasks. We hope that our finding to raise alarms for real-world CV deployment in security-critical applications, and our attacks to be used as benchmarks to evaluate the robustness of DNN-based detection mechanisms. Code is available at: https://github.com/jiayunhan/dispersion_reduction.

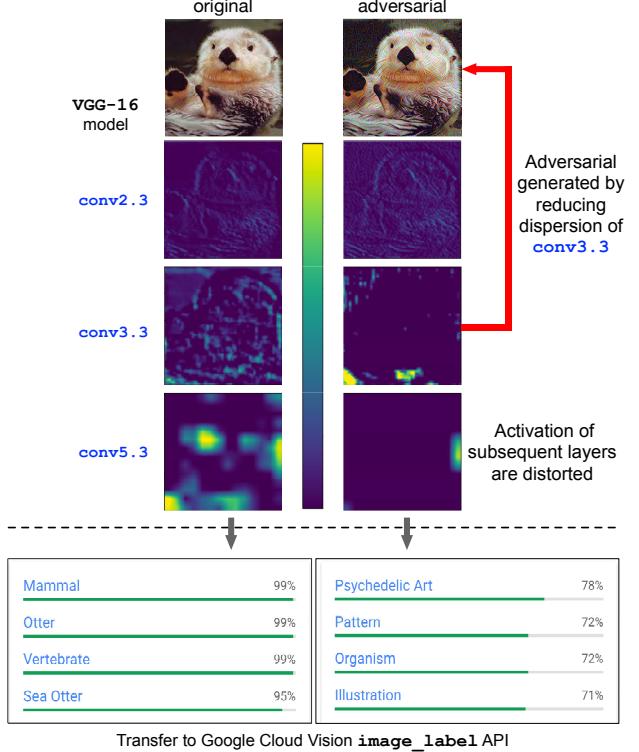


Figure 2. DR attack targets on the dispersion of feature map at specific layer of feature extractors. The adversarial example generated by minimizing dispersion at `conv3 . 3` of VGG-16 model also distorts feature space of subsequent layers (e.g., `conv5 . 3`), and its effectiveness transfers to commercially deployed GCV APIs.

2. Background & Related Work

2.1. Transferability of Adversarial Examples

Since the seminal finding of Szegedy *et al.* [24], the transferability of adversarial examples between different models trained over same or disjoint datasets have been discovered. Followed by Goodfellow *et al.* [11], this phenomenon was attributed to the reason that adversarial perturbations is highly aligned with the weight vector of model. More recently, Papernot *et al.* [21] investigated attacks against black-box models by training substitute models. They also demonstrated attacks against machine learning services hosted by Amazon, and Google.

Our work differs from Papernot *et al.* [21] in two main aspects. First, the GCV APIs we attack in this work is not the same as the Cloud Prediction API [2] (now the Google Cloud Machine Learning Service) attacked in Papernot *et al.* [21]. Both systems are black-box, but the Prediction API is intended to be trained by user’s own data, while the GCV APIs are trained on Google’s data and are provided “out-of-box”. Second, we study transferability over black-box commercial models assuming no feedback on testing samples. Our proposed DR attack do not query the sys-

tems for constructing substitute model [21, 20] nor running score or decision based attacks [8, 12, 19, 23], and as Liu *et al.* [16] demonstrated, it is more difficult to transfer adversarial examples to commercial models that are trained on large dataset, and are potentially ensemble.

2.2. Adversarial Attacks

Several methods have been proposed recently to find AEs and improve transferability. A single-step attack, called fast gradient sign method (FGSM) was proposed by Goodfellow *et al.* [11]. In a follow up work, Kurakin *et al.* [13] proposed a multi-step attack, called iterative fast gradient sign method (I-FGSM) that iteratively searches the loss surface. Generally iterative attack achieves higher success rate than single-step attack in white-box setting, while performs worse when transfer to other models [25].

Fueled by the NIPS 2017 adversary competition [14], several adversarial techniques that enhance transferability have been introduced, among them we given an overview of the most notable ones.

MI-FGSM. Momentum Iterative Fast Gradient Sign Method (MI-FGSM) proposed by Dong *et al.* [10] integrates momentum term into the attack process to stabilize update directions and escape poor local maxima. The update procedure is as follow:

$$\begin{aligned} x'_{t+1} &= x'_t + \alpha \cdot \text{sign}(g_{t+1}) \\ g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x'_t, y)}{\|\nabla_x J(x'_t, y)\|_1} \end{aligned} \quad (1)$$

The strength of MI-FGSM can be controlled by the momentum and the number of iterations.

DIM. Momentum Diverse Inputs Fast Gradient Sign Method (DIM) combines momentum and input diversity strategy to enhance transferability [25]. DIM applies image transformation($T(\cdot)$) to the inputs with a probability p at each iteration of iterative FGSM to alleviate the overfitting phenomenon. The updating procedure is similar to MI-FGSM, with the only replacement of Eq.1 by:

$$x'_{t+1} = \text{Clip}_x^\epsilon \{x'_t + \alpha \cdot \text{sign}(\nabla_x L(T(x'_{t+1}; p), y^{true}))\} \quad (2)$$

where $T(x'_t, p)$ is a stochastic transformation function that performs input diversion on input with a probability of p .

The major difference between dispersion reduction (DR) with existing attacks is that DR doesn't require task-specific loss functions (e.g., cross-entropy used by the family of FGSM attacks). It targets on the numerical property of low level features that is task-independent, and presumably similar across CV models. Our evaluation in §4 demonstrate good transferability of adversarial examples generated by DR across real-world CV tasks.

Algorithm 1 Dispersion reduction attack

Input: A classifier f , original sample x , feature map at layer k ; perturbation budget ϵ

Input: Attack iterations T , learning rate ℓ .

Output: An adversarial example x' with $\|x' - x\|_\infty \leq \epsilon$

1: **procedure** DESPERSION REDUCTION
2: $x'_0 \leftarrow x$
3: **for** $t = 0$ to $T - 1$ **do**
4: Forward x'_t and obtain feature map at layer k :

$$\mathcal{F}_k = f(x'_t)|_k \quad (3)$$

5: Compute standard deviation of \mathcal{F}_k : $g(\mathcal{F}_k)$
6: Compute its gradient w.r.t the input: $\nabla_x g(\mathcal{F}_k)$
7: Update x'_t by applying Adam optimization:

$$x'_t = x'_t - \text{Adam}(\nabla_x g(\mathcal{F}_k), \ell) \quad (4)$$

8: Project x'_t to the vicinity of x :

$$x'_t = \text{clip}(x'_t, x - \epsilon, x + \epsilon) \quad (5)$$

9: **return** x'_t

3. Methodology

Existing attacks perturb input images along gradient directions $\nabla_x J$ that depend on the ground-truth label y and the definition of the task-specific loss function J , which limits their cross-task transferability. We propose *dispersion reduction* (DR) attack that formally define the problem of finding an AE as an optimization problem:

$$\begin{aligned} \min_x g(f(x', \theta)) \\ s.t. \|x' - x\|_\infty \leq \epsilon \end{aligned} \quad (6)$$

where $f(\cdot)$ is a DNN classifier with output of intermediate feature map and $g(\cdot)$ calculates the dispersion. Our proposed DR attack in Algorithm 1 takes a multi-step approach that creates an adversarial example by iteratively reducing the dispersion of intermediate feature map at layer k . Dispersion describes the extent to which a distribution is stretched or squeezed, and there can be different measures of dispersion such as the variance, standard deviation, and gini coefficient [18]. In this work, we choose standard deviation as the dispersion metric and denote it as $g(\cdot)$ due to its simplicity.

Given a target feature map, DR applies Adam optimizer to iteratively perturb image x' along the direction of reducing standard deviation, and projects it to the vicinity of x by clipping at $x \pm \epsilon$. Denoting the feature map at layer k as $\mathcal{F}_k = f(x')|_k$, DR attack solves the following formula:

$$\begin{aligned}
x'_{t+1} &= x'_t - \nabla_{x'} g(\mathcal{F}_k) \\
&= x'_t - \frac{dg(t)}{dt} \cdot \frac{df(x'_t)|_k}{dx'} \\
&= x'_t - \frac{t - \bar{t}}{\sqrt{N-1} \cdot \sqrt{\sum_i (t_i - \bar{t})^2}} \cdot \frac{df(x'_t)|_k}{dx'} \quad (7)
\end{aligned}$$

From Eq.7, we state that given the targeted intermediate feature map, the optimized adversarial example x'_t is achieved when all feature map elements $t_j \in t$ have the same value. Table 4 compares the transferability of AEs generated on different layers (shallow to deep) of off-the-shelf feature extractors across different classification and object detection models. The result on 1000 randomly chosen samples from ImageNet validation set shows that targeting on middle layers, i.e. conv3.3 of VGG-16 and conv3.8.3 of Resnet-152 provides better transferability.

4. Experiments

In this section, we compare DR with state-of-the-art adversarial techniques to enhance transferability on commercially deployed Google Cloud Vision (GCV) tasks:

- Image Label Detection (**Labels**) ¹ classifies image into broad sets of categories.
- Object Detection (**Objects**) ² detects multiple objects with their labels and bounding boxes in an image.
- Image Texts Recognition (**Texts**) ³ detects and recognize text within an image, which returns their bounding boxes and transcript texts.
- Explicit Content Detection (**SafeSearch**) ⁴ detects explicit content such as adult or violent content within an image, and returns the likelihood.

Datasets. We use ImageNet validation set for testing **Labels** and **Objects**, and the NSFW Data Scraper [7] and COCO-Text [4] dataset for evaluating against **SafeSearch** and **Texts** respectively. We randomly choose 100 images from each dataset for our evaluation, and Fig. 3 shows sample images in our testing set.

Experiment setup. We choose normally trained VGG-16 and Resnet-152 as our target models, from which the AEs are generated, as Resnet-152 is commonly used by MI-FGSM and DIM for generation [25, 10]. As DR attack targets on specific layer, we choose conv3.3 for VGG-16 and conv3.8.3 for Resnet-152 as per the profiling result in Table 1.

¹<https://cloud.google.com/vision/docs/detecting-labels>

²<https://cloud.google.com/vision/docs/detecting-text>

³<https://cloud.google.com/vision/docs/detecting-safe-search>

⁴<https://cloud.google.com/vision/docs/detecting-objects>

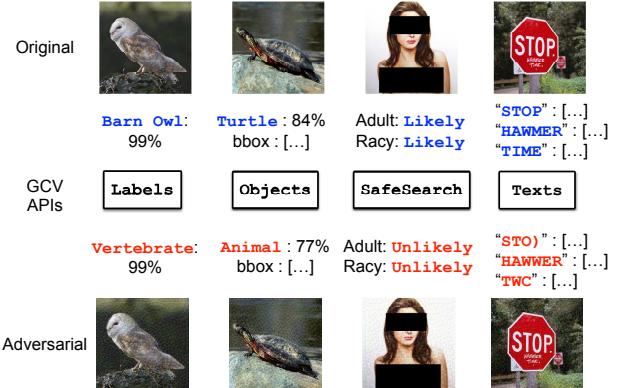


Figure 3. Visualization of images chosen from testing set and their corresponding AEs generated by DR. All the AEs are generated on VGG-16 conv3.3 layer, with perturbations clipped by $l_\infty \leq 16$, and they effectively fool the four GCV APIs as indicated by their outputs.

Attack parameters. We follow the default settings in [10] with the momentum decay factor $\mu = 1$ when implementing the MI-FGSM attack. For the DIM attack, we set probability $p = 0.5$ for the stochastic transformation function $T(x; p)$ as used in [25], and use the same decay factor $\mu = 1$ and total iteration number $N = 20$ as in the vanilla MI-FGSM. For our proposed DR attack, we don't rely on FGSM method, and instead we use Adam optimizer ($\beta_1 = 0.98$, $\beta_2 = 0.99$) with learning rate $5e^{-2}$ to reduce the dispersion of target feature map. The maximum perturbation of all attacks in the experiment are limited by clipping at $l_\infty = 16$, which is still considered less perceptible for human observers [17].

Evaluation metrics. We perform adversarial attacks only on single network and test them on the four black-box GCV models. The effectiveness of attacks are measured by the model performance under attacks. As the labels from original datasets are different from labels used by GCV, we use the prediction results of GCV APIs on the original data as the ground truth, which gives a baseline performance of 100% accuracy or 100.0 mAP and AP respectively. We also provide state-of-the-art results on each CV tasks as references (Table 4).

Figure 3 shows example of each GCV model's output for original and adversarial examples. The performance of **Labels** and **SafeSearch** are measured by the accuracy of classifications. More specifically, we use *top1* accuracy for **Labels**, and use the accuracy for detecting our given porn images as LIKELY or VERY_LIKELY being adult for **SafeSearch**.

The performance of **Objects** is given by the mean average precision (mAP) at $IoU = 0.5$. For **Texts**, we follow the bi-fold evaluation method of ICDAR 2017 Chal-

	Target Layer	Classification - acc.		Detection - mAP(IoU=0.5)	
		Inception-v3	DenseNet	RetinaNet	YOLOv3
VGG-16	conv1.2 (shallow)	52.5%	29.3%	31.8	42.3
	conv3.3 (mid)	28.7%	34.6%	18.3	33.8
	conv5.1 (deep)	35.5%	44.8%	34.0	41.5
Resnet-152	conv1 (shallow)	53.7%	63.1%	28.3	57.3
	conv3.8.3 (mid)	25.8%	34.7%	29.5	41.6
	conv5.3.3 (deep)	28.4%	41.5%	20.5	38.5

Table 1. **The performance of classification and object detection models (columns) when attacked by adversarial examples generated on VGG-16 and Resnet-152.** The profiling result suggests that AEs generated by targeting middle layers degrade performance of both classification and detection models by a larger margin.

Model	Attack	Labels	Objects	SafeSearch	Texts
		acc.	mAP(IoU=0.5)	acc.	AP(IoU=0.5)
baseline (SOTA) ¹		82.5%	73.2	100%	69.2
VGG-16	MI-FGSM	41%	42.6	62%	38.2
	DIM	39%	36.5	57%	29.9
	DR (Ours)	23%	32.9	35%	20.9
Resnet-152	MI-FGSM	37%	41.0	61%	40.4
	DIM	49%	46.7	60%	34.2
	DR (Ours)	25%	33.3	31%	34.6

¹ The baseline performance of GCV models cannot be measured due to the mismatch between original labels and labels used by Google. We use the GCV prediction results on original images as ground truth, thus the baseline performance should be 100% for all accuracy and 100.0 for mAP and AP. Here we provide state-of-the-art performance [5, 6, 4, 7] for reference.

² Correctly recognized words (C.R.W) [4].

Table 2. **The degraded performance of four Google Cloud Vision models, where we attack a single model from the left column.** Our proposed DR attack degrades the accuracy of **Labels** and **SafeSearch** to 23% and 35%, the mAP of **Objects** and **Texts** to 32.9 and 20.9, the word recognition accuracy of **Texts** to only 4.1%, which outperform existing attacks.

lengen [4]. We measure text localization accuracy using average precision (AP) of bounding boxes at $IoU = 0.5$, and evaluate the word recognition accuracy with correctly recognized words (C.R.W) that are case insensitive.

Results. As shown in Table 4, DR outperforms other baseline attacks by degrading the target model performance by a larger margin. For example, the adversarial examples crafted by DR on VGG-16 model brings down the accuracy of **Labels** to only 23%, and **SafeSearch** to 35%. Adversarial examples created with the same technique also degrade mAP of **Objects** to 32.9% and AP of text localization to 20.9%, and with barely 4.1% accuracy in recognizing words. Strong baselines like MI-FGSM and DIM on the other hand, only obtains 38% and 43% success rate when attacking **SafeSearch**, and are less effective compared with DR when attacking all other GCV models. The results demonstrates the better cross-task transferability of dispersion reduction attack.

When comparing the effectiveness of attacks on different generation models, the results that DR generates adversarial

examples that transfer better across these four commercial APIs still hold. The visualization in Fig. 3 shows that the perturbed images with $l_\infty \leq 16$ well maintain their visual similarities with original images, but fools real-world computer vision systems.

5. Discussion & Conclusion

One intuition behind DR attack is that by minimizing the dispersion of feature maps, we are making images “featureless”, as few features can be detected, if neuron activations are suppressed by perturbing the input (Fig. 2). Further, with the observation that low level features bear more similarities across CV models, we hypothesis that DR attack would produce transferable adversarial examples when targeted on intermediate convolution layers. Evaluation on four different CV tasks shows that this enhanced attack greatly degrades model performance, and thus would facilitate evasion attacks against even an ensemble of CV-based detection mechanisms. We hope that our proposed attack can serve as benchmark for evaluating robustness of future defense.

References

- [1] Amazon Rekognition. [Link.](#) 1
- [2] Google Cloud Machine Learning Engine. [Link.](#) 2
- [3] Google Cloud Vision. [Link.](#) 1
- [4] ICDAR2017 Robust reading challenge on COCO-Text. [Link.](#) 4, 5
- [5] ImageNet Challenge 2017. [Link.](#) 5
- [6] Keras Applications. [Link.](#) 5
- [7] NSFW Data Scraper. [Link.](#) 4, 5
- [8] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 1
- [9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1
- [10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 1, 2, 3, 4
- [11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3
- [12] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018. 3
- [13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 3
- [14] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defenses competition. In *The NIPS’17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018. 3
- [15] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009. 2
- [16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 1, 2, 3
- [17] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015. 4
- [18] Chris A Mack. *NIST, SEMATECH e-Handbook of Statistical Methods*. 2007. 3
- [19] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016. 3
- [20] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 3
- [21] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM, 2017. 2, 3
- [22] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 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*, pages 1528–1540. ACM, 2016. 2
- [23] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019. 3
- [24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2
- [25] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. *arXiv preprint arXiv:1803.06978*, 2018. 1, 2, 3, 4
- [26] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 2