Generalizable data-free objective for crafting universal adversarial perturbations

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—Machine learning models are susceptible to adversarial perturbations: small changes to input that can cause large changes in output. It is also demonstrated that there exist input-agnostic perturbations, called universal adversarial perturbations, which can change the inference of target model on most of the data samples. However, existing methods to craft universal perturbations are (i) task specific, (ii) require samples from the training data distribution, and (iii) perform complex optimizations. Also, because of the data dependence, fooling ability of the crafted perturbations is proportional to the available training data. In this paper, we present novel, generalizable and data-free objective for crafting universal adversarial perturbations. Independent of the underlying task, our objective achieves fooling via corrupting the extracted features at multiple layers. Therefore, the proposed objectives are generalizable to craft image-agnostic perturbations across multiple vision tasks such as object recognition, semantic segmentation and caption generation, etc.**

In the practical setting of black-box attacking scenario, we show that our objectives outperform the data dependent objectives to fool these models. Further, via exploiting simple priors related to the data distribution, our objective remarkably boosts the fooling ability of the crafted perturbations. Our significant fooling rates emphasize that the current deep learning models are now at an increased risk, since our objective generalizes across multiple tasks and there is no training data requirement for crafting the perturbations.

***Index Terms*—Adversarial perturbations, attacks on ML systems, data independent objectives.**

I. Introduction

SMALL but structured perturbations to the input, called adversarial perturbations, are shown ([1]–[3]) to significantly affect the output of machine learning systems. Neural network based models, despite their excellent performance, are observed ([4]–[6]) to be vulnerable to adversarial attacks. Particularly, Deep Convolutional Neural Networks (CNN) based vision models ([7]–[11]) can be fooled by carefully crafted quasi-imperceptible perturbations. Multiple hypotheses attempted to explain the existence of adversarial samples, viz. linearity of the models [5], finite training data [12], etc. More importantly, the adversarial perturbations generalize across multiple models. That is, the perturbations crafted for one model fools another model even if the second model has a different architecture or trained on a different subset of training data ([4], [5]). This property of adversarial perturbations enables potential intruders to launch attacks without the knowledge about the target model under attack: an attack model typically known as *black-box attacking*. On the contrast, an attack model where everything about the target model is known to the attacker is called a *white-box attacking*. Until recently, all the existing works assumed a threat model in which the adversaries can directly feed input to the machine learning system. However, Kurakin *et al.* [13] lately showed that the adversarial samples can remain missclassified even if they were constructed in physical world and observed through a sensor (e.g., camera). Given that the models are vulnerable even in physical world scenario [13], it poses serious issues about their deploy-ability (e.g., safety concerns for autonomous driving). Particularly, in case of critical applications that involve safety and security, reliable models need to be deployed to stand against the robust adversarial attacks. Thus, the effect of these structured perturbations has to be studied thoroughly in order to develop dependable machine learning systems.

Recent work by Moosavi-Dezfooli *et al.* [8] presented the existence of image-agnostic perturbations, called universal adversarial perturbations (UAP) that can fool the state-of-the-art recognition models on most of the images. Their method for crafting the UAPs is based on the DeepFool [7] attacking method. DeepFool [7] method involves solving a complex optimization problem (try to refer to the corresponding equation in this draft) to design a perturbation. The UAP [8] procedure utilizes a set of training images to iteratively update the universal perturbation with an objective of changing the predicted label for multiple images. Similar to [8], Metzen *et al.* [11] proposed UAP for semantic segmentation task. They extended the iterative FGSM [5] attack by Kurakin *et al.* [13] to change the label predicted at individual pixels and craft the perturbation.

However, these approaches to craft UAPs ([8], [11]) have the following important issues:
(i) Data dependency: It is observed that the optimization presented by [8] requires a minimum number of training samples for it to converge and craft an image-agnostic perturbation. Moreover, the fooling performance of the resulting perturbation is proportional to the available training data (refer to the corresponding figure in this draft). Similarly, the objective for crafting image-agnostic perturbation proposed by [11] also requires data. Therefore, as their optimization/objectives involve data, existing procedures can not craft perturbations when data is not provided.
(ii) Weaker black-box performance: Since information about the target models is generally not available for attackers, it is practical to perform black-box attacks. Also, black-box attacks

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332.
E-mail: see http://www.michaelshell.org/contact.html
J. Doe and J. Doe are with Anonymous University.
Manuscript received April 19, 2005; revised August 26, 2015.

reveal the true susceptibility of the models, while white-box attacks provide the upper bound on the achievable fooling. However, the black-box attacking performance of UAP [8] is poor compared to their white-box performance (refer to the corresponding section). Note that, in [8], authors have not analyzed the performance of their perturbations in the black-box attack scenario. They have assumed that the training data of the target models is known, which amounts to performing only *semi white-box* attacks. In case of semantic segmentation, as [11] worked with targeted attacks, they observed that the perturbations do not generalize to other models.

In addition to the above issues, the current objectives to craft UAPs depend on the underlying task at hand. For each of the tasks, we generally have a separate objective to optimize, which is because, fooling has different meaning across multiple tasks.

In order to address the above shortcomings and to better analyze the stability of the models, we present novel data-free objective to craft universal adversarial perturbations. Our objective is to craft image-agnostic perturbations that can fool the target model without any knowledge about the data distribution, such as, the number of categories, type of data (e.g., faces, objects, scenes, etc.) or the data samples themselves. Since we do not have access to any data, instead of an objective to flip the predicted label (as in [4], [7], [8], [11]), we propose objectives to learn perturbations that can fool the features learned by the models. Our proposed objectives attempt to over-fire the neurons at multiple layers in order to fool the learned features. During the inference time, the added perturbations misfire the neuron activations in order to contaminate the representations and eventually change the predicted label.

This paper is an extension of our conference paper [9]. In this paper we propose the following new contributions to the conference version: (1) A better data-free objective for crafting image-agnostic perturbations (refer to eqn. **??**) that maximizes the energy of the resulting activations caused by perturbation, (2) An improved optimization procedure that makes the learning process efficient, (3) Demonstrate the generalizability of our data-free objectives to multiple vision tasks (refer to sections **??**) and (4) Provide comprehensive discussion and analysis on the proposed objectives and the crafted perturbations.

The rest of this paper is organized as follows: section II presents detailed account of related works, section III discusses the proposed data independent objectives to craft image agnostic adversarial perturbations, section IV demonstrates the effectiveness of our perturbations via a series of experiments, section V hosts comprehensive discussion on the proposed method and finally section VI concludes the paper.

II. Related Works

Szegedy *et al.* [4] demonstrated that despite their superior recognition performance, neural networks are susceptible to adversarial perturbations. Subsequently, multiple other works [5]–[7], [14]–[17] studied this interesting and surprising property of the machine learning models. Though it is first observed with recognition models, the adversarial behaviour is noticed with models trained on other tasks such as semantic segmentation [11], [18], object detection [18] and deep reinforcement learning tasks [19], etc. There exist multiple methods to craft these malicious perturbations for a given data sample. For recognition tasks, they range from performing simple gradient ascent([5]) on cost function to solving complex optimizations ([4], [7], [20]). Simple and fast methods such as FGSM [5] find the gradient of loss function to determine the direction in the image space to perturb the input image. An iterative version of this attack presented in [13] achieves better fooling via performing the gradient ascent multiple times. On the other hand, complex approaches such as [4], [7] find minimal perturbation that can move the input across the learned classification boundary in order to flip the predicted label. More robust adversarial attacks have been proposed recently that transfer to real world [13] and are invariant to general image transformations [21].

Moreover, it is observed that the perturbations exhibit transferability property. That means, perturbations crafted for one model can fool other models with different architectures and trained on a disjoint training set as well ([4], [5]). Further, Papernot *et al.* [22] introduced a practical attacking setup via model distillation to understand the *black-box* attacking. Black-box attacking assumes no information about the target model and its training data. They proposed to use a target model's substitute to craft the perturbations.

The common underlying aspect of all these techniques is that they are intrinsically data dependent. The perturbation is crafted for a given data sample independently of others. However, recent works by Moosavi-Dezfooli *et al.* [8] and Metzen *et al.* [11] showed the existence of input-agnostic perturbations that can fool the models over multiple images. In [8], authors proposed an iterative procedure based on Deepfool attacking [7] method to craft a universal perturbation to fool classification models. Similarly, in [11], authors craft universal perturbations that can result in target segmentation outputs. However, both these works optimize for different task specific objectives. Also, they require training data to craft the image-agnostic perturbations. Unlike the existing works, the proposed method presents data-free objective that can craft perturbations across multiple computer vision tasks. Particularly, our objective is generalizable across various models in spite of differences in terms of architectures, regularizers, underlying tasks, etc.

III. Proposed Approach

In this section we discuss the proposed data-free objective to craft UAPs in detail.

First, we introduce the notation followed throughout the paper. $\mathcal{X}$ denotes the distribution of images in $\mathbb{R}^d$. f denotes the function learned by the CNN that maps an input image $x \sim \mathcal{X}$ from the distribution to its output $f(x)$. Note that the output is task dependent, for example, it is a label for classification task and segmentation map for semantic segmentation task. Though the proposed objective is task independent, for ease of understanding we explain the proposed approach in the context of image recognition.

A. Data-free objective for fooling

The objective of this paper is to craft an image-agnostic perturbation $\delta \in \mathbb{R}^d$ that fools the CNN f over images from $\mathcal{X}$ without utilizing any samples from it. In other words, we seek a universal adversarial perturbation δ that significantly alters the prediction of the CNN f. That is, we synthesize a δ such that

$$f(x+\delta) \neq f(x), \quad \text{for } x \in \mathcal{X} \tag{1}$$

In order for the δ to be called adversarial perturbation, it has to be imperceptible when added to the images. Therefore the pixel intensities of δ are restricted by an imperceptibility constraint. Typically, it is realized as a max-norm (l_∞) constraint (e.g. [5], [8], [9], [11]). Thus, the aim is to find a δ such that

$$\begin{aligned} f(x+\delta) &\neq f(x), \quad \text{for } x \in \mathcal{X} \\ \| \delta \|_\infty &< \xi \end{aligned} \tag{2}$$

However, the focus of the proposed work is to craft the image-agnostic perturbations without requiring any samples from the training dataset $\mathcal{X}$ on which the target model f is learned. The data-free nature of our approach prohibits us from the first part of the eqn. 2 while learning δ. That is our approach does not contain the data term x in the proposed objective. Therefore, we propose to fool the CNN by over-firing the features extracted at multiple layers, as opposed to the "flipping the label" objective. That is, we make the activations at each layer to fire for the perturbation δ and thereby misleading the features (filters) at the following layer.

The perturbation essentially causes filters at a particular layer to spuriously fire and abstract out ineffectual/inefficacious information. Note that in the presence of data (during attack/testing), in order to mislead the activations from retaining useful discriminative information, the perturbation has to be highly effective. Also, the imperceptibility constraint (second part of eqn. 2) on δ makes it a difficult problem.

Hence without utilizing any data (x), we seek for an image-agnostic perturbation δ that can produce maximal spurious activations at each layer of a given CNN. In order to craft such a δ we start with a random perturbation and optimize for the following objective

$$Loss = -\log\left(\prod_{i=1}^{K} \|l_i(\delta)\|_2\right) \quad \text{such that} \quad \| \delta \|_\infty < \xi \tag{3}$$

where $l_i(\delta)$ is the activation in the output tensor at layer i when δ is fed to the network f. Note that the activations are considered after the non-linearity (typically ReLU). K is the total number of layers in f at which we maximize the activations for perturbation δ. ξ is the maxnorm limit on δ.

The proposed objective computes product of activation magnitude at all the individual layers in order to simultaneously maximize the interference at all layers. We observed product resulting in stronger δ than other forms of combining (e.g. sum) individual layer activations. This is understandable, since product can force the individual activations to increase for reducing the loss. To avoid working with extreme values (≈ 0), we apply log on the product. Note that the objective is open-ended as there is no optimum value to reach. We would ideally want δ to cause as much strong perturbation at all the layers as possible within the imperceptibility constraint.

B. Implementation Details

We begin with a target network f which is a trained CNN whose parameters are frozen and a random perturbation δ. We then perform the proposed optimization to update δ for causing strong activations at multiple layers in the given network. We typically consider all the convolution $(conv)$ layers before the fully connected (fc) layers. This is because, the $conv$ layers are generally considered to learn required features to extract information over which a series of fc layers perform classification. Also, we empirically found that it is efficient to optimize at $conv$ layers. Therefore, we restrict the optimization to feature extraction layers. In case of advanced architectures such as GoogLeNet [23] and ResNet [24]

Note that the optimization updates the perturbation image δ but not the network parameters and no image data is involved in the optimization process. We update δ with the gradients computed for loss in eqn. (3).

C. Exploiting additional priors

Though the proposed method is a data-free optimization for crafting image-agnostic perturbations, it can exploit very simple additional priors about the data distribution $\mathcal{X}$. In this section we demonstrate how our method can utilize simple priors such as (i) mean value and dynamic range of the input, and (ii) minimal target data.

1) Mean and dynamic range of the input: Note that the proposed optimization (eqn. (3)) does not consider any information about $\mathcal{X}$. We present only the norm limited δ as input and maximize the resulting activations. That is, during the optimization, input to the target CNN has a dynamic range of $[-10, 10]$. However, during the inference time, input lies in $[0, 255]$ range. Therefore, we provide this useful information about the data $(x \in \mathcal{X})$, and let the optimization better explore the space of perturbations. Thus, we slightly modify our objective to craft δ relative to the dynamic range of the data. We create pseudo data d via randomly sampling from a Gaussian distribution whose mean (μ) is equal to the mean of training data and variance (σ) that covers 99.9% of density to lie in $[0, 255]$, the dynamic range of input. Essentially, we solve for the following

$$\begin{aligned} Loss &= -\log\left(\prod_{i=1}^{K} \|l_i(d+\delta)\|_2\right) \\ \text{such that } &\|\delta\|_\infty < \xi, \text{ and } d \sim \mathcal{N}(\mu, \sigma) \end{aligned} \tag{4}$$

2) Minimal target data: In this subsection, we demonstrate that our data-free objective can utilize samples from the target distribution $\mathcal{X}$ and benefit to improve the fooling ability of the crafted perturbations. Note that in the case of data availability, we can design better objectives such as reducing confidence for the predicted label or changing the predicted label, etc.

However, we show that our data-free objective of over-firing the activations, though is not designed to utilize data, crafts better perturbations when data is presented to the optimization. Note also that, our objective does not utilize data to manipulate the predicted confidences. Rather, the optimization benefits from limited prior information about the data distribution such as the dynamic range, local patterns, etc., which can be provided with minimal samples. Therefore, with minimal data samples we solve for the following problem

$$Loss = -\log\left(\prod_{i=1}^{K} \|l_i(x+\delta)\|_2\right) \quad \text{such that } \|\delta\|_\infty < \xi, \text{ and } x \sim \mathcal{X} \tag{5}$$

D. Improved Optimization

In this subsection we present the improvements we propose to our optimization first proposed in [9].

1) Re-scaling the dynamic range: : The proposed the objective in [9] is observed to quickly accumulate δ beyond the imposed max-norm constraint (ξ). Because of the clipping performed after each iteration, the updates will be futile after δ reaching the constraint. In order to tackle this saturation, δ is re-scaled to half of its dynamic range (i.e. $[-5,\ 5]$) in regular time intervals of 300 iterations. Though this re-scaling helps to learn better δ, it is inefficient since it performs blind re-scaling without verifying the scope for updating the δ. For example, as the learning progresses, magnitude of updates decreases and during the interval of 300 iterations, the values of δ might not reach the extreme values of ± 10. Performing another re-scaling badly affects the learning.

Therefore, we propose an adaptive re-scaling of δ based on the rate of saturation in its pixel values. During the optimization, at each iteration we compute the proportion (p) of the pixels in δ that reached the max-norm limit ξ. As the learning progresses, since our objective is open ended, more number of pixels reach the max-norm limit and because of the clipping get saturated at ξ. Hence, the rate of increase in p decreases as δ saturates. For consecutive iterations, if increase in p is not significant, we perform a re-scaling to half the dynamic range. Note that the proposed criterion for re-scaling is similar to the typical usage of validation performance to stop training.

IV. Experiments

In this section we present the experimental evaluation to demonstrate the effectiveness of the proposed data-free objective. We consider three different vision tasks to demonstrate the generalizability of our objective, namely, object recognition, semantic segmentation and place-holder. We explain each of the tasks separately in the following subsections.

A. Object recognition

We have worked with models trained on ILSVRC [25] and Places [26] datasets, viz. Caffenet [27], VGG-F [28], GoogLeNet [23], VGG-16 [29], VGG-19 [29], ResNet-152 [24]. Since our approach does not involve training the models, for all the experiments we work with available trained models. Also, unlike UAP [30], as we do not use training data in the data-free case (sec. III-A), no training data is used. However, in case of exploiting additional data prior (sec. III-C), we use very limited data from the corresponding training set. However, for evaluating the fooling ability of the crafted perturbations, 50000 images from the ILSVRC validation set and $20000(?)$ images form Places validation set are used.

TABLE I
Fooling rates for the proposed data-free objective. Diagonal rates indicate the white-box attacking and off-diagonal ones represent the black-box attacking scenarios.

Model	Caffenet	VGG-F	GoogLeNet	VGG-16	VGG-19	Resnet-152
Caffenet	**87.02**	65.97	49.40	50.46	49.92	38.57
VGG-F	59.89	**91.91**	52.24	51.65	50.63	40.72
GoogLeNet	44.70	46.09	**71.44**	37.95	37.90	34.56
VGG-16	50.05	55.66	46.59	**63.08**	56.04	36.84
VGG-19	49.11	53.45	40.90	55.73	**64.67**	35.81
Resnet-152	38.41	37.20	33.22	27.76	26.52	**37.3**

TABLE II
Fooling Rates with Various Priors

Model	No prior	Noise prior	Data prior	UAP'17
Caffenet	84.88	87.02	91.54	93.1
VGG-F	85.96	91.81	92.64	93.8
Googlenet	58.62	71.44	83.54	78.5
VGG-16	45.47	63.08	77.77	77.8
VGG-19	40.68	64.67	75.51	80.8
Resnet-152	29.78	37.3	66.68	84.0

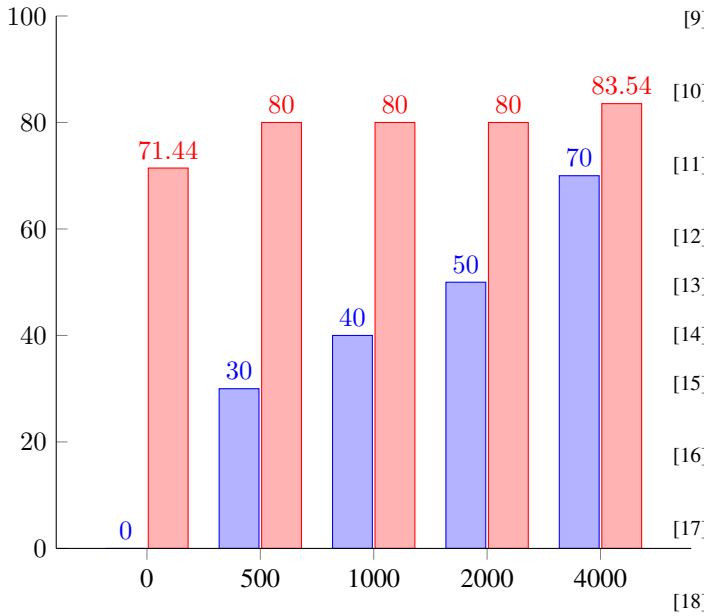
Table I presents the fooling rates achieved by our objective on various network architectures. Fooling rate is the percentage of test images for which our crafted perturbations δ successfully changed the predicted label. Higher the fooling rate, greater is the perturbation's ability to fool and lesser is the the classifier's robustness. Fooling rates in Table I are obtained using the mean and dynamic range prior of the training distribution (sec. III-C1). Each row in the table indicate one target model employed in the learning process and the columns indicate various models attacked using the learned perturbations. The diagonal fooling rates indicate the *white-box attacking*, where all the information about the model is known to the attacker. The off-diagonal rates indicate *black-box attacking*, where no information about the model under attack is revealed to the attacker. Our perturbations cause a mean white-box fooling rate of 69.24% and a mean black-box fooling rate of 45.13%. Note that, given the data-free nature of the optimization, the fooling rates are alarmingly significant.

TABLE III
COMPARISON OF OBJECTIVES

Model	Mean Objective	L2 Objective
Caffenet	88.35	91.54
VGG-16	72.68	77.77
Resnet-152	65.43	66.68

TABLE IV
TRANSFER ATTACK, ON IMAGENET, DATA OF PLACES205

Model	Ours	UAP
CaffeNet	87.02	73.09
GoogleNet	71.44	28.17



V. DISCUSSION

VI. CONCLUSION

The conclusion goes here.

REFERENCES

[1] B. Biggio, G. Fumera, and F. Roli, "Pattern recognition systems under attack: Design issues and research challenges," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 07, 2014.

TABLE V
TRANSFER ATTACK, ON PLACES205, DATA OF IMAGENET

Model	Ours	UAP
CaffeNet	88.61	77.21
GoogleNet	83.37	52.53

TABLE VI
TRANSFER ATTACK, CRAFT IN ILSVRC, ATTACK ON PLACES205

Model	Ours	UAP
CaffeNet	67.62	63.1
GoogleNet	55.87	54.0

[2] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 387–402.
[3] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ser. AISec '11, 2011.
[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.
[7] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.
[8] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[9] K. R. Mopuri, U. Garg, and R. V. Babu, "Fast feature fool: A data independent approach to universal adversarial perturbations," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.
[10] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," *arXiv preprint arXiv:1703.08603*, 2017.
[11] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *International Conference on Computer Vision (ICCV)*, 2017.
[12] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, Jan. 2009.
[13] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.
[14] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *arXiv preprint arXiv:1502.02590*, 2015.
[15] A. Fawzi, S. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: from adversarial to random noise," in *Advances in Neural Information Processing Systems NIPS*, 2016.
[16] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.
[17] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–32.
[18] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. L. Yuille, "Adversarial examples for semantic segmentation and object detection," *arXiv preprint arXiv:1703.08603*, 2017.
[19] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," *arXiv preprint arXiv:1701:04143*, 2017.
[20] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi, "Measuring neural net robustness with constraints," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.
[21] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017.
[22] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *CoRR*, vol. abs/1602.02697, 2016.
[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[26] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *CoRR*, vol. abs/1610.02055, 2016.
[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[28] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the BMVC*, 2014.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," vol. abs/1409.1556, 2014.
[30] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *arXiv preprint arXiv:1610.08401*, 2016.