

Distorting Neural Representations to Generate Highly Transferable Adversarial Examples

Muzammal Naseer[†], Salman H. Khan^{*‡}, Shafin Rahman^{†*}, Fatih Porikli[†]

[†]Australian National University, ^{*}Data61-CSIRO, [‡]Inception Institute of AI

muzammal.naseer@anu.edu.au

Abstract

Deep neural networks (DNN) can be easily fooled by adding human imperceptible perturbations to the images. These perturbed images are known as the ‘adversarial examples’ that pose a serious threat to security and safety critical systems. A litmus test for the strength of adversarial examples is their transferability across different DNN models in a black box setting (i.e. when target model’s architecture and parameters are not known to attacker).

Current attack algorithms that seek to enhance adversarial transferability work on the decision level i.e. generate perturbations that alter the network decisions. This leads to two key limitations: (a) An attack is dependent on the task-specific loss function (e.g. softmax cross-entropy for object recognition) and therefore does not generalize beyond its original task. (b) The adversarial examples are specific to the network architecture and demonstrate poor transferability to other network architectures.

We propose a novel approach to create adversarial examples that can broadly fool different networks on multiple tasks. Our approach is based on the following intuition: “Deep features are highly generalizable and show excellent performance across different tasks, therefore an ideal attack must create maximum distortions in the feature space to realize highly transferable examples”. Specifically, for an input image, we calculate perturbations that push its feature representations furthest away from the original image features. We report extensive experiments to show how adversarial examples generalize across multiple networks across classification, object detection and segmentation tasks.

1. Introduction

Transferability is a phenomenon where adversarial examples created for one network can fool the others. Transferability of adversarial examples makes it challenging to deploy deep neural networks in security critical environments. This is of high concern because it gives attackers

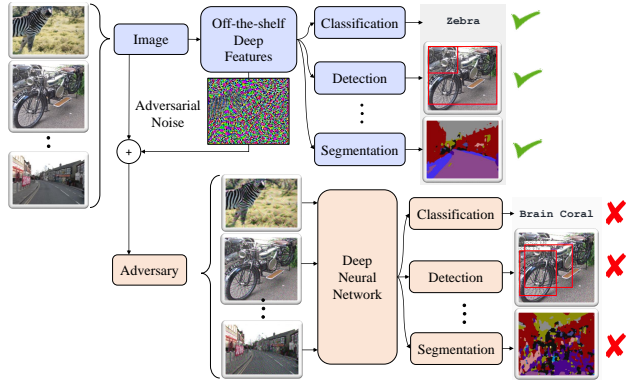


Figure 1: Similar to off-the-shelf deep features that are employed to boost the performance of different computer vision tasks, adversarial noise pertinent to deep feature spaces is transferable across different tasks. (Noise pattern is magnified for visualization)

the flexibility to train a local network and transfer its attack against an already deployed network without knowing its architecture or parameters (‘black-box attacks’). Current attack algorithms [8, 5] perform well when network architecture and parameters are known (‘white-box setting’), however, their strength significantly decreases in the black box setting as shown in [21]. Recent attempts on enhancing the transferability in black-box settings have been reported in [6, 26]. Nevertheless, their dependency on task-specific loss function makes them non-transferable across different tasks. For example, to fool classification models, the attacker starts from the softmax cross-entropy to find such a gradient direction that increases the model loss for a given sample. Examples found in this way are specific and do not generalize beyond their original task.

We propose a novel approach to generate high strength adversarial examples that are transferable across different network architectures and most importantly across different vision tasks (e.g., image segmentation, classification and object detection). Our approach is based on the intuition that neural networks trained on ImageNet [18] (or other

sufficiently large image datasets) learn generic internal representations that are transferable to new tasks and datasets [10, 19]. As a result, it is a common practice to use pre-trained classification networks as the basic building block (*network backbone*) for a variety of different tasks [13, 15]. We hypothesize that the adversarial examples found in the rich feature space of the networks trained with large image datasets are ideally suited to fool any deep network whether designed for classification, object detection, segmentation or other vision tasks. We present the first such algorithm that creates adversarial examples by distorting the deep neural activations. This not only generates high strength perturbations but also provides flexibility to work with any task as the proposed attack does not use any task dependent loss function.

To the best of our knowledge, the closest to our approach is a decision-boundary free attack (called as FFF) presented in [17]. The idea is to train a perturbation within a given metric norm to maximize activation response of network’s internal layers. After training, the perturbation is added to input images to make them adversarial. The problem with this approach is that it optimizes adversarial noise in a way that is independent of the data sample; hence noise severely overfits the network and has very low transferability. In contrast, we propose to maximize network internal representation distortion w.r.t the given original benign sample. One intriguing aspect of our approach is its simplicity and efficiency. For instance, we can only use features from a single layer (conv3.3) of VGG-16 [20] (instead of multiple layers in [17]) and calculate the mean squared difference between the original and adversarial examples to represent neural representation distortion (NRD). NRD is differentiable and minimizing it can help in image restoration problems [11]. Here, we propose to maximize the NRD to construct adversarial examples. Finding adversarial examples based on features representation makes the attack generalizable across different architectures of different tasks. Specifically, we show high inter-task and intra-task transferability of our approach for large-scale datasets including ImageNet [18], MS-COCO [14] and CAMVID [4].

Our method is not restricted to the original backbone models trained on a specific benchmark. Most backbone models are fine-tuned with additional training datasets to a specific task. As we elaborate in Sec. 6, our method successfully applies to any network that is pretrained on one benchmark then fine-tuned on another, e.g. RetinaNet [13] and SegNet [3].

Contributions: We study and highlight the importance of neural network’s internal representations (Fig. 1) in the context of adversarial attacks. Our major contributions are:

- We propose a generalizable, black-box, untargeted adversarial attack algorithm on neural network’s internal representation.

- We leverage on generic representations learned by models (e.g. VGG-16 [20]) trained on large image datasets (e.g. ImageNet [18]) to construct transferable adversarial examples.
- Our attack algorithm does not rely on a task-specific loss function or a specific set of input labels, therefore it demonstrates cross-network, cross-dataset, and cross-task transferability.

2. Related Work

Since the seminal work of Szegedy et al. [24] many adversarial attack algorithms [7, 8, 2, 6] are proposed to show vulnerability of neural networks against imperceptible changes to inputs. A single step attack, called fast gradient sign method (FGSM) was proposed by [7]. It is computationally efficient but not very robust and primary reason to propose such a single step attack was to use it in inner loop during training. This way of training when network is being trained on original as well as adversarial examples is called adversarial training [7]. In a follow-up work, Kurakin et al. [8] proposed a robust multiple steps attack, called iterative fast gradient sign methods (I-FGSM) that searches the loss surface of a network iteratively under a given metric norm. To improve transferability, a variant of I-FGSM, called momentum iterative fast gradient sign method (MI-FGSM) was introduced [6], which significantly enhances the transferability of untargeted attack on ImageNet [18] under perturbation budget of $l_\infty \leq 16$. Authors [6] associated the transferability of MI-FGSM with its ability to break local maxima as number of attack iterations increase.

Interestingly, NRD of I-FGSM decreases as number of attack iterations increase as compared to MI-FGSM as shown in Fig. 2. We generated adversarial examples for Inception-v3 (Inc-v3) on ImageNet [18] subset dataset provided by NIPS security challenge 2017 by running I-FGSM and MI-FGSM for iterations ranging from 2 to 10. These examples are then transferred to Inception-v4 (Inc-v4) [22]. Fig. 2 shows that as number of attack iteration increases, transferability and NRD of I-FGSM decreases while MI-FGSM maintain both transferability along with NRD. Recently, [26] proposed data augmentation technique to further boost the transferability of these attack methods.

3. Adversarial Attacks

In this section, we first provide our problem setting followed by a brief background to adversarial attacks. We explain how popular attack mechanisms such as FGSM [7], I-FGSM [8], MI-FGSM [6] differ from each other. This background will form the basis of our proposed attack in Sec. 4.

Problem Setting: In this paper, we specifically consider the transferability of untargeted attacks under the l_∞

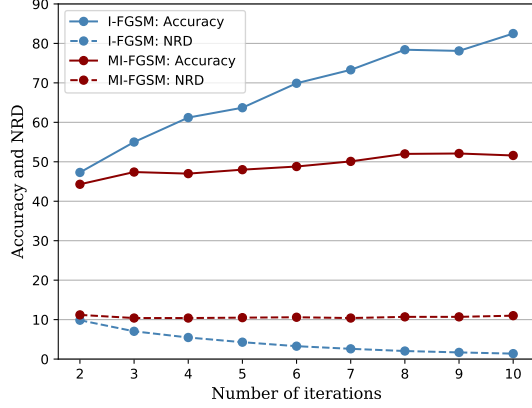


Figure 2: Accuracy of Inc-v4 and NRDM is shown for adversarial examples generated on Inc-v3 by I-FGSM and MI-FGSM. NRDM is averaged over all examples. As the number of iterations increases, accuracy of Inc-v4 on adversarial example found by I-FGSM increases, i.e., transferability of I-FGSM decreases along with its NRD.

norm constraint on perturbation strength. The untargeted attacks are considered because they have higher transferability compared to targeted attacks [6, 26]. Furthermore, to make sure that the benign and adversarial examples are close to each other, an attacker is constrained under a metric norm like $l_\infty \leq \epsilon$, i.e., in the case of images attacker can change each pixel intensity value by at maximum ϵ amount.

3.1. FGSM

Adversarial examples can be formulated as constrained optimization problem. Suppose we are given a classifier function \mathcal{F} that maps an input \mathbf{x} to its ground-truth class y , a cost function $J(\mathbf{x}, y)$ that is used to train the classifier and an allowed perturbation budget ' ϵ '. FGSM [7] finds an adversarial example \mathbf{x}' that satisfies $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon$ using the following formulation:

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}, y)), \quad (1)$$

where $\nabla_{\mathbf{x}} J(\mathbf{x}, y)$ represent gradient of cost function w.r.t input \mathbf{x} . A common choice for J is the cross-entropy loss. The problem with FGSM is that it is a single step attack which reduces the attack success rate due to underfitting the threat model. To overcome this difficulty, an iterative version of FGSM was proposed [8].

3.2. I-FGSM

I-FGSM [8] iteratively applies FGSM with a small step size α for a given number of iterations T . The step size α can be calculated by dividing perturbation budget ϵ with number of iterations T i.e., $\alpha = \epsilon/T$. I-FGSM can be represented as follows for steps $t \in [1, T]$:

$$\mathbf{x}'_0 = \mathbf{x}, \quad \mathbf{x}'_{t+1} = \mathbf{x}'_t + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}'_t, y)). \quad (2)$$

The problem with I-FGSM is that it overfits the threat model reducing model accuracy to even 0% while producing small neural representation distortion (NRD) (See Fig. 2 for an empirical evidence). One side effect of having low NRD is the reduced transferability of adversarial examples. This is where Dong *et al.* [6] built upon and proposed an attack algorithm that finds adversarial examples iteratively while maintaining the transferability rate.

3.3. MI-FGSM

The work in [6] added momentum into the optimization objective of I-FGSM. It can be expressed as follows:

$$\begin{aligned} \mathbf{x}'_0 &= \mathbf{x}, \quad \mathbf{x}'_{t+1} = \mathbf{x}'_t + \alpha \cdot \text{sign}(\mathbf{g}_{t+1}), \quad t \in [1, T] \\ \mathbf{g}_{t+1} &= \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}} J(\mathbf{x}'_t, y)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}'_t, y)\|_1}. \end{aligned} \quad (3)$$

The strength of MI-FGSM can be described by two of its control parameters, number of iterations and momentum. Number of attack iterations allows it be strong in white-box settings (like I-FGSM) while momentum allows it to increase NRD enhancing the attack success rate in black-box settings.

Based on the above observations, we build our framework and propose to enhance the NRD directly to create strong adversarial examples for black-box attacks.

4. Neural Representation Distortion

The Problem: The strong white-box attack algorithms [8, 5] consider already known network parameters θ and perturb the input to create \mathbf{x}' such that the example is misclassified i.e., $\mathcal{F}(\mathbf{x}'; \theta) \neq y$. Since the perturbations are calculated using gradient directions that are specific to θ , the resulting perturbed images \mathbf{x}' do not generalize well to other networks [6, 21]. The attacks presented in [6, 26] show relatively better transferability, however, these attacks also perturb input images along gradient directions $\nabla_{\mathbf{x}} J$ that are dependent on the ground-truth label y and the definition of loss function J . This dependence limits the cross-network and cross-task transferability of these attacks.

Our Solution: We note that the transferability rate of an attack is correlated with the distortion in deep feature representations (Fig. 2). In this paper, we propose to directly maximize the representation loss in terms of deep feature activations by solving the following optimization problem:

$$\begin{aligned} \max_{\mathbf{x}'} \quad & \mathcal{F}(\mathbf{x}')|_k - \mathcal{F}(\mathbf{x})|_k \\ \text{subject to:} \quad & \|\mathbf{x} - \mathbf{x}'\|_\infty \leq \epsilon, \end{aligned} \quad (4)$$

where \mathcal{F} is DNN based classifier, k is the internal representation layer and ϵ is the allowed perturbation budget. We apply a transformation \mathcal{T} to input \mathbf{x} at first iteration (algorithm 1) to create neural representation difference of adversarial w.r.t benign example and then maximize the mean

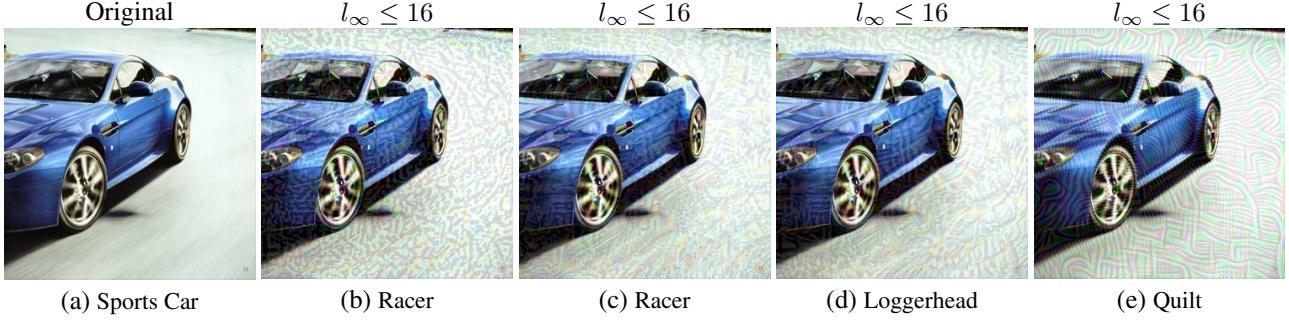


Figure 3: VGG-16 output is shown on example images. (a) represents benign example, while (b), (c), (d) and (e) show adversarial examples generated by FGSM, MI-FGSM, DIM and NRDM respectively against VGG-16. All adversarial examples have distance $l_\infty \leq 16$ from the original seed (a).

squared error of this difference with in a given perturbation budget. There can be different choices for \mathcal{T} but in this work \mathcal{T} simply adds random noise to input sample, i.e our algorithm takes a random step at first iteration. Random noise is convenient to attain a difference at the starting point of our algorithm and it is be more preferable to heuristic transformations that may cause methodical bias.

We use VGG-16 [20] conv3.3 feature map as neural representation distortion. This choice is based on observations reported in the recent study [21] that adversarial examples found in VGG space have high transferability. This is also evident in our experimentation (Table 4). Increasing representation loss at multiple network layers did not notably increase attack success yet added a significant computational overhead. We only report results by increasing the representation distortion at VGG-16 [20] conv3.3 layer.

Our attack algorithm does not rely on the cross-entropy loss and the input labels. This makes it a generic algorithm and can be used to attack any system that uses off the shelf features in their pipeline. This makes several popular computer vision tasks vulnerable to adversarial attacks that are based on ImageNet [18] trained backbones e.g., object detection and segmentation. Furthermore, our proposed approach is complementary to recent best-performing attack methods such as MI-FGSM [6] and DIM [26]. Therefore, we demonstrates that it can be used alongside them, which further boosts the strength of the NRDM approach. Our proposed method to maximize NRD for a given input sample is summarized in Algorithm 1 and Fig. 4.

5. Experiments

5.1. Evaluation Protocol

In this section, we describe the datasets used for evaluation, networks architectures under attack, and the parameter settings for each attack algorithms.

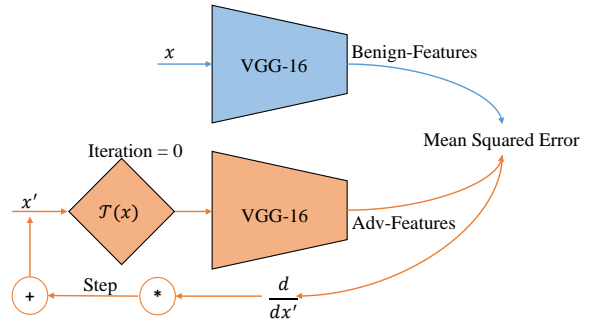


Figure 4: Our NRDM (Algorithm 1) transforms the benign sample at iteration zero and then optimizes for an adversary response based on the feature representation loss.

Algorithm 1 Neural Representation Distortion Method

Input: A classifier \mathcal{F} , input sample x , input transformation \mathcal{T} , internal network layer k , perturbation budget ϵ and number of iterations T .

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

- 1: $g_0 = 0$; $x' = x$;
- 2: **for** $t = 0$ to $T - 1$ **do**
- 3: **if** $t = 0$ **then**
- 4: $x' = \mathcal{T}(x)$
- 5: **end if**
- 6: Forward pass x'_t to \mathcal{F} and compute \mathcal{L} as follows;

$$\mathcal{L} = \|\mathcal{F}(x')|_k - \mathcal{F}(x)|_k\|_2; \quad (5)$$

- 7: Compute gradients $g_t = \nabla_x \mathcal{L}(x'_t, x)$;
- 8: Apply the following equation;
- 9: Project adversary into the vicinity of x

$$x'_{t+1} = x'_t + \epsilon \cdot \text{sign}(g_t); \quad (6)$$

$$x'_{t+1} = \text{clip}(x'_{t+1}, x - \epsilon, x + \epsilon); \quad (7)$$

- 10: **end for**
 - 11: **return** $x' = x'_T$.
-

5.1.1 Datasets

We have used MNIST, CIFAR10 test sets and ImageNet [18] subset provided by NIPS security challenge 2017 (ImageNet-NIPS) to validate the effectiveness of the proposed attack against classification models. MNIST and CIFAR10 test set contain 10k samples each, while ImageNet-NIPS contains 1k image samples. For object detection, We used MS-COCO [14] validation set that contains 40.5k images. This is a multi-task dataset popular for image segmentation, object detection and image captioning tasks. We report adversarial attack performance against object detection, however adversarial examples found on this dataset can be used to fool other related tasks e.g., visual question answering. For segmentation, we used CAMVID [4] test set to measure segmentation robustness against adversarial examples generated by our method (Algorithm 1). This dataset contains 233 image samples extracted from video sequences of driving scenes.

model-m	model-c
conv2d(32, 3x3)	2*{conv2d(96, 3x3)}
maxpool(2x2)	conv2d(96, 3x3, s=2)
conv2d(64, 3x3)	2*{conv2d(192, 3x3)}
maxpool(2x2)	conv2d(96, 3x3, s=2)
conv2d(64, 3x3)	2* conv2d(192, 3x3)
fc(64)	conv2d(10, 3x3)
softmax(10)	avg-pool
	softmax(10)

Table 1: Architectures of naturally trained convolutional networks for MNIST (model-m) and CIFAR10 (model-c). ‘*’ indicates number of times a layer is repeated. ‘s’ represent stride. Each convolution layer is followed by ReLU activation. Batch-norm is used after each convolutional layer in model-c. Layers whose output is used by NRDM are highlighted in bold.

5.1.2 Network Architectures

Classification: We study eight models trained on ImageNet dataset [18]. These models can be grouped into two categories. (a) *Naturally trained:* Five of these models are only trained on benign examples. These include Inceptionv3 (Inc-v3) [23], Inceptionv4 (Inc-v4), Inception Resnet v2 (IncRes-v2) [22] and Resnet v2-152 (Res-152) [9] and VGG-19 [20]. (b) *Adversarially trained:* The other three models including Adv-v3 [12], Inc-v3_{ens3} and IncRes-v2_{ens} [25] are adversarially trained and made publicly available. The specific details about these models can be found in [12, 25]. Attacks are created for naturally trained models while tested against all of them.

For classification on smaller datasets, we studied three models each for MNIST and CIFAR10. Among these

res-m	res-c
conv2d(16, 3x3)	conv2d(16, 3x3)
1*rb { conv2d(16, 3x3) conv2d(16, 3x3)	3*rb { conv2d(16, 3x3) conv2d(16, 3x3)
1*rb { conv2d(32, 3x3) conv2d(32, 3x3, s=2)	3*rb { conv2d(32, 3x3) conv2d(32, 3x3, s=2)
1*rb { conv2d(64, 3x3) conv2d(64, 3x3, s=2)	3*rb { conv2d(64, 3x3) conv2d(64, 3x3, s=2)
softmax(10)	avg-pool(8x8) softmax(10)

Table 2: Architectures of naturally trained residual networks for MNIST (res-m) and CIFAR10 (res-c). ‘*’ indicates number of times a layer is repeated. ‘s’ and ‘rb’ represent stride and residual block respectively. Each convolution layer is followed by ReLU activation. Batch-norm is used after each convolutional layer in res-c.

models, two are naturally trained and one is adversarially trained using saddle point optimization [16]. Adversarial examples are created for naturally trained models named model-m and model-c for MNIST and CIFAR10 respectively (see Table 1). These examples are subsequently transferred to adversarially trained Madry’s models [16] and naturally trained ResNet models named res-m and res-c for MNIST and CIFAR10 respectively (see Table 2).

Object Detection: To show cross-task and cross-dataset transferability, we study naturally trained RetinaNet [13] performance against adversarial examples found by NRDM approach (Algorithm 1) on MS-COCO validation set.

Segmentation: We evaluated robustness of naturally trained SegNet-basic [3] against adversarial examples generated by NRDM approach (Algorithm 1) on CAMVID [4] test set.

5.1.3 Attack Parameters

FGSM is a single step attack. Its step size was set to 16. In the case of R-FGSM, we took a step of size $\alpha=16/3$ in a random direction and then gradient step of size $16-\alpha$ to maximize model loss. The attack methods, I-FGSM, MI-FGSM and DIM, were ran for ten iterations. Step size for these attacks was set to 1.6 as per standard practice. The momentum decay factor for MI-FGSM was set to one. This means that attack accumulates all the previous gradient information to perform the current update and is shown to have best success rate [6]. For DIM, the transformation probability is set to 0.7. In the case of FFF [17], we train adversarial attack for 10K number of iterations to maximize response at activation layers of VGG-16 [20]. For the NRDM algorithm 1, We used VGG-16 [20] conv3-3 feature map as representation loss. Since NRDM maximizes loss w.r.t benign example, so it does not suffer from over-fitting

Accuracy	Naturally Trained					Adv. Trained		
	Inc-v3	Inc-v4	Res-152	IncRes-v2	VGG-19	Adv-v3	Inc-v3 _{ens3}	IncRes-v2 _{ens}
T-1	95.3	97.7	96.1	100.0	85.5	94.3	90.2	96.9
T-5	99.8	99.8	99.9	100.0	96.7	99.4	95.5	99.8

Table 3: Model accuracies are reported on original data set ImageNet-NIPS containing benign examples only. T-1: top-1 and T-5: top-5 accuracies. Best and second best performances are colored.

	Attack	Naturally Trained										Adv. Trained					
		Inc-v3		Inc-v4		Res-152		IncRes-v2		VGG-19		Adv-v3		Inc-v3 _{ens3}		IncRes-v2 _{ens}	
		T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Inc-v3	FGSM	22.0*	45.7*	62.5	84.7	64.6	85.8	65.9	85.9	49.9	75.7	69.1	88.1	77.2	90.9	90.8	98.3
	R-FGSM	16.7*	38.0*	65.8	86.0	69.5	89.7	68.8	88.7	61.4	83.8	76.4	90.9	77.9	91.2	88.8	97.6
	I-FGSM	0.0*	1.7*	82.0	97.6	86.5	98.6	90.6	99.1	76.7	95.0	88.5	98.7	84.9	94.4	94.6	99.5
	MI-FGSM	0.0*	1.5*	47.1	78.8	47.1	84.5	52.5	81.9	47.3	76.7	71.6	89.8	73.8	90.7	88.3	98.0
	DIM	0.2*	1.3*	27.8	63.1	42.1	75.2	34.6	65.4	40.2	71.4	65.2	87.9	68.3	89.6	86.3	97.5
Res-152	FGSM	54.1	79.3	61.2	84.2	16.5*	41.0*	62.5	85.6	46.0	72.7	67.3	87.4	74.0	89.4	88.4	97.7
	R-FGSM	58.5	83.4	64.9	86.6	12.9*	35.2*	69.1	88.5	56.1	80.8	74.5	90.6	75.5	90.4	86.5	96.5
	I-FGSM	80.0	96.6	84.1	98.4	0.9*	6.2*	92.5	99.1	75.7	94.9	87.4	99.0	85.5	94.8	93.4	99.3
	MI-FGSM	43.5	76.8	49.9	79.2	0.9*	5.1*	54.8	82.4	46.8	76.0	72.6	90.7	71.1	90.1	86.0	97.5
	DIM	20.1	51.2	22.0	54.6	0.6*	4.2*	24.6	57.3	33.3	62.6	53.5	82.5	55.2	83.1	74.4	94.1
IncRes-v2	FGSM	61.7	83.8	69.6	87.6	68.4	89.6	50.1*	73.9*	52.3	76.5	72.0	89.6	79.0	91.6	90.0	97.7
	R-FGSM	66.6	87.0	71.8	89.4	73.5	91.5	46.1*	71.3*	62.9	84.1	75.5	91.2	79.3	91.5	87.4	97.3
	I-FGSM	62.8	88.4	68.3	91.9	77.2	94.8	1.1*	2.6*	71.4	91.7	85.6	97.5	83.8	95.6	89.8	98.4
	MI-FGSM	36.0	67.5	42.4	73.2	49.3	82.2	1.0*	2.4*	51.3	76.8	70.0	90.1	71.5	92.2	81.8	96.3
	DIM	21.4	49.8	23.5	53.4	32.3	64.3	4.8*	13.7*	39.7	69.2	54.9	81.4	57.5	85.9	73.5	94.4
VGG16	FGSM	30.1	56.0	34.0	58.0	36.6	65.2	42.2	66.1	9.1	27.9	48.8	72.6	53.5	79.5	72.8	91.1
	R-FGSM	41.5	67.9	45.1	72.5	49.2	78.4	54.9	77.7	12.9	35.8	63.9	86.2	63.5	85.3	77.1	93.0
	I-FGSM	69.4	93.0	75.3	94.5	79.5	95.7	87.2	97.9	18.3	56.1	82.2	97.5	80.9	93.7	91.5	99.1
	MI-FGSM	16.9	42.0	18.7	40.1	24.9	51.6	26.1	52.5	2.0	14.4	38.8	68.1	42.5	72.4	64.2	87.7
	DIM	12.9	35.5	15.2	35.8	20.6	45.7	19.7	43.8	0.6	8.8	31.6	59.0	32.1	61.0	56.3	81.0
	FFF	61.7	80.7	60.8	78.7	72.8	90.1	76.1	90.1	44.0	68.0	79.6	93.1	83.1	93.1	92.8	98.5
	NRDM	5.1	10.2	6.2	12.4	15.6	27.6	13.6	23.0	4.5	14.2	27.7	46.8	54.2	75.4	75.3	89.8
	NRDM-DIM	4.9	10.5	5.7	12.0	16.0	28.6	12.7	22.6	5.0	14.0	28.7	45.7	52.9	73.8	74.0	89.5

Table 4: Model accuracies are reported under untargeted l_∞ adversarial attacks on ImageNet-NIPS with perturbation budget $l_\infty \leq 16$ for pixel space [0-255]. T-1 and T-2 represent top-1 and top-5 accuracies, respectively. NRDM shows higher or competitive success rates for black-box models than FGSM [7], I-FGSM [8], MI-FGSM [6], DIM [26] and FFF [17]. NRDM-DIM combines input diversity as well as momentum with NRDM. “*” indicates the white-box attacks. Best and second best black-box attacks are colored.

problem. We ran NRDM for the maximum number of 100 iterations. Transferability of different attacks is compared against number of iterations in Fig. 5. MI-FGSM and DIM quickly reach to their full potential within ten number of iterations. I-FGSM strength decreases while NRDM strength increases with the number of attack iterations.

Dataset ↓	Metric ↓	Naturally Trained		Adv. Trained
		model-m	res-m	Madry-M
MNIST	Accuracy	99.30	98.88	98.40
		model-c	res-c	Madry-C
CIFAR10	Accuracy	85.44	80.56	87.62

Table 5: Model accuracies on original test data sets for MNIST and CIFAR10 containing benign examples only. Best and second best performances are colored.

5.2. Input Transformations

Different input transformations have been proposed to mitigate the adversarial effect but they can be easily broken in a *white-box* scenario. This is because an attacker can be adaptive and incorporate transformations into adversary generation process. Even non-differentiable transformations can be by-passed by approximating them with an identity function [1]. However in a *black-box* scenario, the attacker does not have any knowledge of transformation function along with network architecture and its parameters. We have tested strength of our adversarial attack against well studied transformations including:

- *JPEG*: This transformation reduces adversarial effect by removing high frequency components in the input image.
- *Total Variation Minimization (TVM)*: TVM measures small variations thus it can be effective against relatively smaller adversarial perturbations.

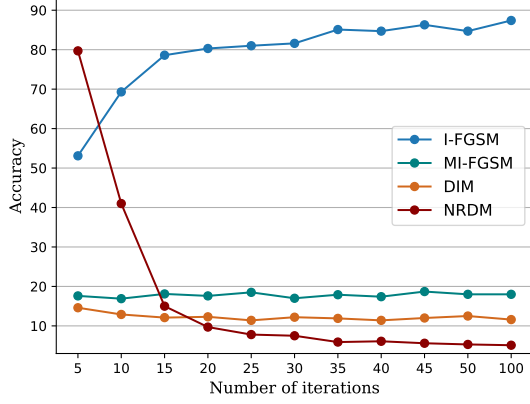


Figure 5: Accuracy of Inc-v3 for adversarial examples generated on VGG-16 by I-FGSM and MI-FGSM, DIM and NRDM. NRDM strength increases with number of iterations in comparison to MI-FGSM and DIM.

Datasets ↓	Attack ↓	Naturally Trained		Adv. Trained Madry-M
		model-m	res-m	
MNIST	FGSM	42.28*	53.15	95.96
	I-FGSM	40.66*	51.04	96.64
	MI-FGSM	40.66*	48.19	95.96
	NRDM	4.39*	23.54*	97.56
CIFAR10		model-c	res-c	Madry-C
	FGSM	5.47*	24.19	85.54
	I-FGSM	2.52*	36.81	87.00
	MI-FGSM	2.52*	16.56	85.71
	NRDM	11.92*	23.98	86.99

Table 6: Model accuracies under untargeted l_∞ adversarial attacks on MNIST and CIFAR10 with perturbation budget $l_\infty \leq 76.5$ and $l_\infty \leq 8$ respectively for pixel space [0-255], as per standard practice [16]. NRDM shows higher or competitive success rates for black-box models than FGSM, I-FGSM and MI-FGSM. ‘*’ indicates the white-box attacks. Best and second best attacks are colorized.

- **Median Filtering:** This transformation filters out the input image by replacing each pixel with the median of its neighboring ones.

We report our experimental results using the above mentioned network architectures and input transformations in the following section.

6. Results

Classification: We report the performance of our attack against a number of CNN architectures on ImageNet-NIPS dataset in Table 4. The following insights can be drawn from our results. (1) In comparison to other state of the art attacks, our approach consistently demonstrates a much higher transferability rate for naturally trained images. Specifically, NRDM attack have much higher transferability on naturally trained models bringing down top-1

Method	No Attack	NRDM	
		$l_\infty \leq 8$	$l_\infty \leq 16$
No Defense	79.70	52.48	32.59
JPEG (quality=75)	77.25	51.76	32.44
JPEG (quality=50)	75.27	52.45	33.16
JPEG (quality=20)	68.82	53.08	35.54
TVM (weights=30)	73.70	55.54	34.21
TVM (weights=10)	70.38	59.52	34.57
MF (window=3)	75.65	49.18	30.52

Table 7: Segnet-Basic accuracies on CAMVID test set with and without input transformations against NRDM. Best and second best performances are colorized.

Method	No Attack	NRDM	
		$l_\infty \leq 8$	$l_\infty \leq 16$
No Defense	53.78	22.75	5.16
JPEG (quality=75)	49.57	20.73	4.7
JPEG (quality=50)	46.36	19.89	4.33
JPEG (quality=20)	40.04	19.13	4.58
TVM (weights=30)	47.06	27.63	6.36
TVM (weights=10)	42.79	32.21	9.56
MF (window=3)	43.48	19.59	5.05

Table 8: mAP (with IoU = 0.5) of RetinaNet is reported on MS-COCO validation set with and without input transformations against NRDM. Best and second best performances are colorized.

accuracy of IncRes-v2 [22] from 100.0% (see Table 3) to 12.7% (see Table 4). (2) In comparison, MI-FGSM [6] and DIM [26] perform slightly better on adversarially trained ensemble models [25] with NRDM showing competitive success rate. This is because the MI-FGSM and DIM methods use decision boundary information while NRDM is agnostic to decision-level information about the classifier. (3) We also test with adversarial examples found using different network architectures (i.e., Inc-v3, Res-152, IncRes-v2, VGG16). Overall, we conclude that the adversarial examples found in VGG-16 [20] space have very high transferability.

On small datasets (MNIST and CIFAR10), similar to other attacks, the NRDM becomes ineffective against adversarially trained Madry models [16] (see Tables 5 and 6) in black-box settings. This shows that finding better methods for adversarial training is a way forward to defend against these attacks. Input transformations can somewhat help to mitigate the adversarial effect in black-box settings (see Table 9). TVM is the most effective against all the attacks while median filtering perform better against DIM [26]. JPEG is the least effective against untargeted adversarial attacks.

Segmentation: NRDM attack created on CAMVID [4] in VGG-16 feature space able to bring down per pixel accuracy of Segnet-Basic by 47.11% within $l_\infty \leq 16$ (see Table 7). JPEG and TVM transformations are slightly effective

	No Attack		FGSM		R-FGSM		I-FGSM		MI-FGSM		DIM		NRDM	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
No Defense	95.3	99.8	30.1	56.0	41.5	67.9	69.4	93.0	16.9	42.0	12.9	35.5	5.1	10.2
JPEG (quality=75)	93.9	99.5	30.4	55.4	41.8	67.0	69.7	92.3	18.4	42.0	13.5	33.3	5.4	12.6
JPEG (quality=50)	91.3	99.3	31.0	55.4	40.5	65.3	68.7	91.8	18.1	42.1	13.1	34.4	6.5	12.8
JPEG (quality=20)	86.0	97.6	29.9	53.9	38.0	64.6	69.8	90.9	18.4	42.1	14.1	34.2	8.4	18.7
TVM (weights=30)	93.1	99.4	30.6	56.2	41.7	67.7	73.7	94.5	17.2	42.1	14.9	33.5	9.8	18.5
TVM (weights=10)	88.8	97.6	32.1	57.3	43.6	69.4	73.9	93.4	19.8	45.7	15.8	37.1	24.0	40.5
MF (window=3)	93.2	99.1	24.3	45.5	36.1	62.3	62.8	89.9	16.2	36.8	18.8	42.1	9.9	17.9

Table 9: Inc-v3 accuracy is reported with and with-out input transformations. Adversarial examples are generated for VGG-16 in white-box setting by FGSM, R-FGSM, I-FGSM, MI-FGSM, DIM and NRDM under perturbation budget $l_\infty \leq 16$ and then transferred to Inc-v3. T-1 and T-2 represent top-1 and top-5 accuracies, respectively. Best and second best performances are colored.

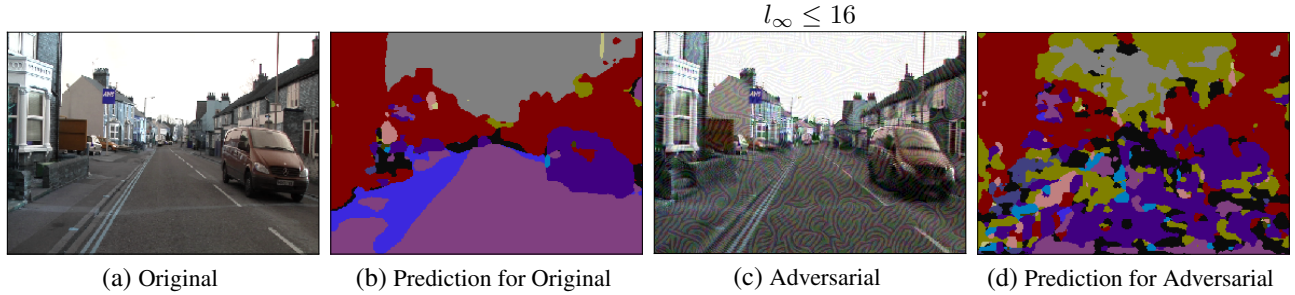


Figure 6: Segnet-Basic output is shown for different images. (a) is the original image, while (b) shows predictions for the original image. (c) is the adversary found by NRDM algorithm 1, while (d) shows predictions for the adversarial image. Perturbation budget is written on the top of adversarial image.

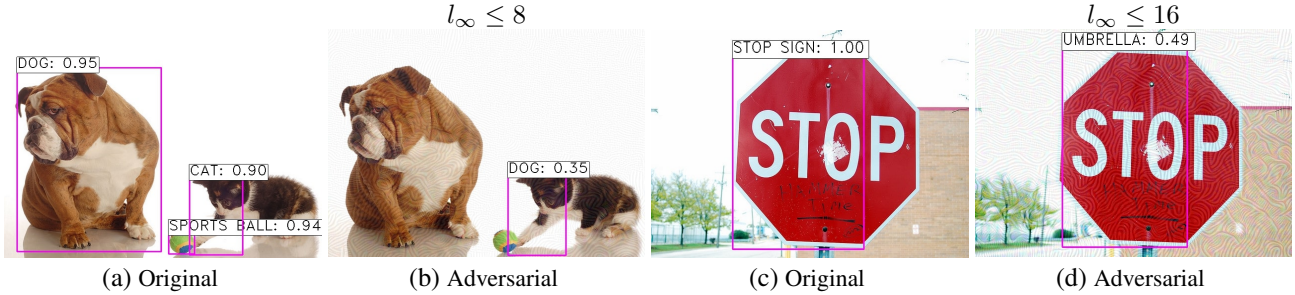


Figure 7: RetinaNet detection results are shown for different images. (a) and (c) show detection for the original images, while (b) and (d) show detection for adversaries found using NRDM algorithm 1. Perturbation budget is written on the top of each adversarial image.

but only at the cost of accuracy on benign examples.

Object Detection: RetinaNet [13] collapses in the presence of adversaries found by NRDM on MS-COCO validation set using the VGG-16 [20] feature space. Its mean average precision (mAP) with 0.5 intersection over union (IOU) drops from 53.78% to 5.16% under perturbation budget $l_\infty \leq 16$ (see Table 8). TVM is relatively more effective compared to other transforms against the NRDM attack.

7. Conclusion

In this paper, we improve transferability of adversarial examples under a constrained attack scenario. We highlight

the importance of neural representation distortion (NRD) in the context of adversarial attacks and propose to directly maximize feature distortion with respect to benign examples. We propose an attack algorithm to demonstrate how to benefit from generic internal neural representations of pre-trained models (like VGG-16) on ImageNet dataset to exhibit cross-architecture, cross-dataset and cross-task transferability. Generalizability of our attack algorithm makes it suitable to make any system robust against adversaries via adversarial training that benefits from generic off-the-shelf feature representations of pre-trained classification models.

References

- [1] A. Athalye, N. Carlini, and D. A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 6
- [2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017. 2
- [3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2017. 2, 5
- [4] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 2, 5, 7
- [5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1, 3
- [6] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 4, 5, 6, 7
- [7] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICRL)*, 2015. 2, 3, 6
- [8] I. J. Goodfellow, J. Shlens, and C. Szegedy. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICRL)*, 2017. 1, 2, 3, 6
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5
- [10] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018. 2
- [11] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [12] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 5
- [13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2, 5, 8
- [14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5
- [15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 5, 7
- [17] 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. 2, 5, 6
- [18] 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)*, 115(3):211–252, 2015. 1, 2, 4, 5
- [19] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 2
- [20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2, 4, 5, 7, 8
- [21] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy? - a comprehensive study on the robustness of 18 deep image classification models. *CoRR*, abs/1808.01688, 2018. 1, 3, 4
- [22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017. 2, 5, 7
- [23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 5
- [24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICRL)*, 2014. 2
- [25] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICRL)*, 2018. 5, 7
- [26] C. Xie, Z. Zhang, J. Wang, Y. Zhou, Z. Ren, and A. Yuille. Improving transferability of adversarial examples with input diversity. *arXiv preprint arXiv:1803.06978*, 2018. 1, 2, 3, 4, 6, 7