Catastrophic Child's Play: Easy to Perform, Hard to Defend Adversarial Attacks

Chih-Hui Ho[1]* Brandon Leung[1]* Erik Sandström[2] Yen Chang[1] Nuno Vasconcelos[1]
[1]University of California, San Diego [2]Lund University
{chh279,b7leung}@ucsd.edu tfy14esa@student.lu.se {yec084,nvasconcelos}@ucsd.edu

Abstract

The problem of adversarial CNN attacks is considered, with an emphasis on attacks that are trivial to perform but difficult to defend. A framework for the study of such attacks is proposed, using real world object manipulations. Unlike most works in the past, this framework supports the design of attacks based on both small and large image perturbations, implemented by camera shake and pose variation. A setup is proposed for the collection of such perturbations and determination of their perceptibility. It is argued that perceptibility depends on context, and a distinction is made between imperceptible and semantically imperceptible perturbations. While the former survives image comparisons, the latter are perceptible but have no impact on human object recognition. A procedure is proposed to determine the perceptibility of perturbations using Turk experiments, and a dataset of both perturbation classes which enables replicable studies of object manipulation attacks, is assembled. Experiments using defenses based on many datasets, CNN models, and algorithms from the literature elucidate the difficulty of defending these attacks – in fact, none of the existing defenses is found effective against them. Better results are achieved with real world data augmentation, but even this is not foolproof. These results confirm the hypothesis that current CNNs are vulnerable to attacks implementable even by a child, and that such attacks may prove difficult to defend.

1. Introduction

Convolutional neural networks (CNNs) trained on large corpora such as ImageNet [5] have enabled significant advances in computer vision in recent years. While initially popular for recognition, these models have shown to be remarkably easy to train and transfer across vision tasks, and are now almost universally used across computer vision. Recently, however, this robustness has been questioned by some puzzling findings derived from adversarial CNN attacks [14, 20, 18, 3, 17, 23]. Although CNNs have excellent, even superhuman [9], recognition performance on randomly internet-collected test images, it is quite easy to generate images where they fail dramatically [25, 8]. In fact, most images that a CNN classifies correctly can be transformed into images that it cannot classify, by the addition of a very small adversarial perturbation. Most interestingly, this perturbation can usually be made so small as to be imperceptible, i.e. impossible to detect, by a human. This suggests that the space of images correctly classified by most CNNs is, at most, a countable dense subset of the space of images recognizable by humans, e.g. similar to the relationship between rational and real numbers.

This problem is of great concern for many applications. For example, smart cars depend on CNNs to make decisions that could have life or death consequences, security and surveillance systems rely on CNNs for identity verification, etc. The existence of many images capable of fooling CNNs poses a significant challenge to such applications. This has spurred interest in adversarial attacks [7, 14] and a literature has emerged in the area, with many variants of the problem being proposed. In result, there are at least four fundamental dimensions along which adversarial algorithms can differ: they can be 1) "white" [25, 8, 14, 20, 18, 3, 17] or "black"-box [4, 30, 23, 6, 11], depending on whether knowledge of the CNN model to attack is required, 2) "targeted" or "non-targeted," depending on whether the goal is to induce the network to make specific errors [25, 20, 3] or to simply make an error [8, 14, 18, 17, 30], 3) "digital" or "real-world" depending on whether the examples used in the attack are produced by an algorithm [25, 8, 23] vs. object manipulation in the real world [14, 7, 1], and 4) "single model" or "universal" depending on whether they aim to fool a single network [14, 18, 3, 30] or many models [17]. Interestingly enough, the relative difficulty of these problems does not always correlate with what would be intuitively expected. For example, it appears that most of the attacks designed to fool a particular CNN, e.g. AlexNet [13], also fool most other CNNs [19, 26, 15, 16, 29], e.g. VGG [22], Inception [24],

*Equal contribution

or ResNet [10]. Similarly, some "black-box" attacks involving simple image transformations [6, 11] appear to be much more effective than "white-box" methods that require access to the CNN and optimization based on backpropagation style of algorithms.

In general however, it can be quite difficult to compare the merits of different algorithms. This is due to two main problems. First, most methods use perturbations that cannot be easily compared. While many authors rely on the standard of an image that is "perceptually indistinguishable from the original" to define a valid attack, it is not clear what the boundaries of "indistinguishable" are and no attempts have been made to define this concept. Instead, the standard is usually met by adoption of a very conservative attack strategy, e.g. the use of an "infinitesimally small step along some gradient direction". It is frequently unclear if the use of larger perturbations would enable the same algorithm to produce more successful attacks. Second, most adversarial works do not even attempt to compare performance with previous approaches. This is unlike most other areas of computer vision, where the ability to compare algorithms is considered critical to evaluate progress.

Recently, some works have started to address the second problem through a strategy that we denote as the "arms race". This exploits the fact that any attack procedure can be transformed into a defense, by 1) augmenting the training set, e.g. ImageNet, with examples produced by the procedure and 2) fine-tuning the network. While not guaranteeing full robustness against the attack [15, 26, 12], this defense strategy renders most attacks much less effective. Under the "arms race" paradigm, a new attack strategy is considered state of the art if it fools a network that implements defenses to previously known attacks [28]. The "arms race" captures the fact that, for practical applications, the only significant attacks are those for which no defenses are available. However, while knowing the attack procedure enables a defense, not all attacks are equally easy to defend. An important variable is the defense's cost. For example, attacks that require more computation to defend against are more costly than attacks than can be thwarted with little computation. Similarly, attacks have different costs. For example, white box attacks can be rendered impractical by the simple use of a proprietary CNN. Overall, the most concerning attacks are those easiest to execute and hardest to defend against.

In this work, we consider the design of such attacks. We argue that the most successful attacks are those that leverage the limitations of computer vision, namely those based on perturbations that are easily produced by people but cannot be replicated by computers. This exploits the large imbalance between the cost of attack and defense in terms of the number of required examples. While an attack requires a few well chosen examples, its defense requires augmenting the training set with an extensive number of examples. Hence, while attacks can be generated manually, those that cannot be defended with computer generated examples are impractical to defend against. We then consider a set of image perturbations based on *variation of object pose.* This is an operation that can be implemented by a child (simply by rotating an object) but is very hard to defend against, due to the well known difficulty of synthesizing objects under different poses [21, 27]. We consider attacks using both small and large perturbations, due to *camera shake* (CS) and *pose variations* (PV). However, the study of such attacks requires a definition of which perturbations are valid. After all, extreme poses can confuse even humans. Unfortunately, common definitions, such as "infinitesimal gradient steps" or imperceptibility on side-by-side image comparisons, are not suitable for *large* perturbations. We argue that these can only be declared imperceptible given an attack context and seek definitions of imperceptibility suited for the object recognition context. This suggests a distinction between imperceptible perturbations (IPs), which survive image comparisons, and semantically imperceptible perturbations (SIPs), which are perceptible on image comparisons but have no impact on human ability to recognize objects.

Overall, this work makes three contributions to the study of adversarial attacks on CNNs. The first is a dataset of images of multiple object classes under camera shake and pose variation. The object classes are a subset of ImageNet, to enable the attack of ImageNet trained CNNs, and each object is imaged with extensive coverage of both small (camera shake) and large (pose variation) view variability. The second contribution is a procedure to determine which perturbations are imperceptible to humans, using Amazon Turk experiments. The procedure is designed to support many attack contexts and could be used to characterize many other types of attacks. We consider two contexts, image and object retrieval, that enable the differentiation between imperceptible perturbations and semantically imperceptible perturbations for object recognition; these can be thought of as small vs. large perturbations. A dataset containing camera shake and pose variation perturbations of the two types is finally assembled, to support the study of recognition attacks. A final contribution is an extensive experimental study of camera shake and pose variation attacks' performance, against multiple CNN models, trained on multiple datasets, and augmented with multiple defenses from the literature. This shows that pose attacks are highly successful against existing CNNs, previous defenses are ineffective against them, and even data collection can have limited effectiveness. Thus, while easy to perform, pose attacks can be difficult to defend.

2. Prior work

There is now a significant literature on adversarial attacks. The most popular setting is a non-targeted white-box digital attack of a single model [8, 14, 18]. The attack is usually

an image perturbation based on an infinitesimal step along the gradient of the loss used to train the model, evaluated at the image [8, 14]. The simplest attacks reduce to one back-propagation iteration, computing derivatives with respect to the input image, and require a forward and backward pass through the network [8]. Many variants have been proposed, including different algorithms [20, 3, 16] or slight variations on the problem. For example, [17] proposed similar techniques for universal attacks, i.e. perturbations that fool many models, and [25, 20, 3] considered targeted attacks. These aim to induce specific errors, e.g. the classification of "apples" as "oranges", using a somewhat more sophisticated optimization. All these methods are digital and can, in principle, be defended against by using the attack algorithm to generate augmentation data to retrain the CNN.

More recently, there has been interest in attacks based on object manipulation in the real world [14, 7, 1]. Some of these address specific applications, such as recognition by smart cars. For example, [7] investigated attacks based on the addition of stickers to traffic signs. This is much less general than the attacks now proposed, which can be applied to any object. Others have investigated the manipulation of images in the world, or the fabrication of objects with certain properties. For example, [1] devised an interesting procedure to fabricate objects that can consistently fool CNNs irrespective of viewing angle. While having some similarities to the attacks now proposed, this setup is substantially more complex than the one presented, which does not require object fabrication. Fabrication raises the cost of attack, by requiring access to knowledge of object fabrication, and drastically reduces the cost of defense, since it relies on algorithms that can be leveraged to produce defenses digitally. For example, because the objects fabricated by [1] have digital textures, their images can be rendered by computer. This is unlike real objects and textures, which are well known to be difficult to capture and render accurately under pose variation [27].

Perhaps most related to this work are previous efforts based on image transformations. For example, [6] has shown that black box attacks by simple image rotation can fool CNNs more effectively than white-box attacks based on gradient optimization. A recent extension of this idea uses spatial transformer networks to synthesize image transformation attacks more general than rotations [28]. This work again showed that image transformations are successful even on networks that implement defenses against gradient attacks. However, all these methods implement digital attacks, using algorithms that can in turn be exploited to defend against them. We propose a setting that generalizes these procedures, relying on real world image manipulation. This is much harder to defend against.

3. Using pose to attack recognition networks

There are several challenges to the study of adversarial attacks. A meaningful attack requires two images: a *true positive* x, i.e. a successfully classified image, and a perturbation x'. A first difficulty is that x' should be, in some sense, "identical" to x. Otherwise, it is illogical to ask the classifier to assign it to the same class, and the attack is ill-defined. We refer to this as the problem of *attack validity.* Consider the popular framework of attacks based on additive perturbations, $x' = x + \eta\delta$, where δ is a function of the gradient of the classification loss with respect to x [15]. In the absence of a criterion to test whether x and x' are "identical", validity is sought by constraining η to be very small, so as to make x' *visually indistinguishable* from x. However, this is not a full guarantee of validity, since a person with infinite time can frequently identify the perturbed image. There can also be moiré-like interference patterns that easily give the perturbation away. Some methods attempt to address the problem by thresholding the gradient, but this can produce salt-and-pepper artifacts. In general, it is difficult to guarantee that x and x' are indistinguishable.

For these methods, the validity problem follows from the lack of realism in the perturbations used for the attack. We refer to this as the *realism problem.* The difficulty is that δ *is not* a natural image. Hence, the methods above simply produce images at the "edge" of the space of natural images. While overly large steps along δ produce completely unrealistic images, a small enough η guarantees they are acceptable. Yet, because the perturbed images do not occur in the real world, the perturbations must be very small for the attack to remain valid. This leads to a third problem, which is the *small perturbation problem,* i.e. exclusion of attacks that are not immediate neighbors of the true positive. For most applications, such attacks are a much stronger concern than infinitesimal steps towards the edge of image space. For example, the shake and pose attacks proposed in this work can occur *naturally* during the operation of a vision system. This also implies that they are much easier to perform and thus much more likely to be executed – imagine a world where any child can hack a robot simply by showing it familiar objects in strange poses.

In summary, because there is lack of realism, validity can only be guaranteed by small perturbations. This has motivated a recent emergence of perturbations $x' = f(x)$ where f is no longer additive. Various functions have been proposed, from affine transformations [6] to affixing stickers on images [7, 2], to building 3D objects [1]. Because they are more realistic, the perturbations can be larger. On the other hand, large realistic perturbations exacerbate the difficulty of the validity problem since it is even harder to define an "indistinguishable" transformation. For example, a simple in-plane rotation can turn a '6' into a '9'. Similarly, if one is allowed to affix fur to a traffic sign, or repaint it, it will

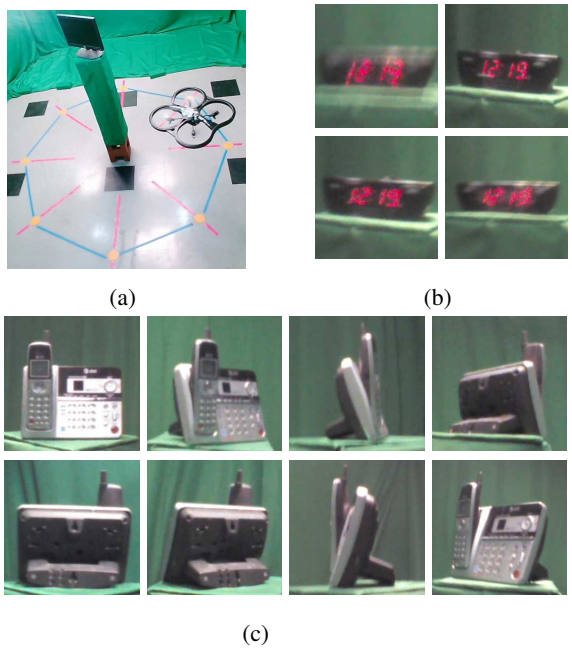


Figure 1: (a) Drone capturing images during flight. (b) Examples of varying levels of camera shake as the drone hovers. (c) Example images collected per viewing angle.

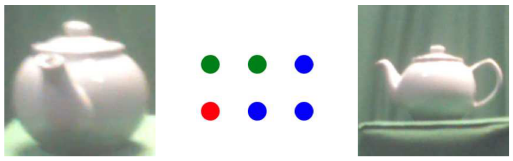
eventually stop being a traffic sign. While most works make an effort to select perturbations indistinguishable from the true positive in some form, this is never quantified. Beyond potentially compromising the significance of these studies, this makes it difficult to compare attacks.

In this work, we avoid these problems by introducing a new attack strategy based entirely on real-world object manipulations. This automatically eliminates the realism problem, since all attacks are based on natural images. We then propose a protocol to *guarantee* the validity of all attacks, by verifying that all perturbations are imperceptible to humans. Finally, we consider a domain (view transformations) that enables the characterization of the *size* of a perturbation. This enables the study of *both* small and large perturbations. We next discuss these contributions in detail.

3.1. Camera shake and pose manipulations

The ultimate goal of this work is to explore the space of attacks that are easy to perform (sometimes even arising naturally from real-world vision systems) but difficult to defend. The idea is to exploit image transformations that can be easily performed in the real world but are hard to replicate by computer. This leverages the fact that while an attack may be performed with a single example, most attacks can only be defended by training the classifier with many examples. When example collecting is costly, the defense becomes impractical. In this context, digital attacks which use algorithms to produce examples are easier to defend than real world attacks involving image manipulations not replicable by computer. Despite significant advances in photo-realistic rendering, it is still not possible to synthesize truly realistic examples from most object classes, at least without a significant investment in a sophisticated computer graphics infrastructure, rendering experts, etc. Hence, attacks with examples of objects under novel views or novel imaging conditions are difficult to defend. An additional benefit of these attacks is that they make it relatively easy to manipulate perturbation size, which correlates with the degree of view change. We illustrate this by introducing a family of attacks ranging from small transformations due to "camera shake" (CS – small variations of camera position) to larger transformations due to "pose variation" (PV – changes in viewing angle). These attacks are also particularly important because they are trivial to perform. For example, a child can shake a camera or rotate an object. In fact, they are inevitable in certain domains of computer vision, such as robotics, where objects can appear in many 3D orientations and the vision system is subject to nuisances such as shaking due to robot movement.

The first contribution of this work is a dataset to enable replicable studies of camera shake and pose variation attacks. For this, we relied on a drone-based imaging setup. A drone was flown around an object, as illustrated in Figure 1a, using markings on the ground to define picture taking stops at regularly spaced intervals. By collecting images at these stops under alignment with the markings on the ground, the drone assembled a set of views of the object corresponding to different orientations of the object in 3D. We refer to these as "object poses". Examples of multiple poses of an object are shown in Figure 1c. Within each stop, the drone was allowed to hover and collect several images of the object, as shown in Figure 1b. Due to the small hovering motion, many of these images are indistinguishable to the inattentive eye. They show the same pose of the same object, varying by very small translations of the camera and some amount of motion blur. We refer to this as "camera shake". The procedure was repeated for 20 objects per class from 23 different object classes. To facilitate attacks on existing object recognizers, these are classes represented in the ImageNet dataset, where the recognizers are trained. Overall, the dataset contains 30 camera shake images per pose and 8 poses for 460 objects, totaling 110,400 images. It is split into a *defense dataset* containing 16 objects per class and an *attack dataset* containing 4 objects per class. The defense dataset can be used to learn defenses against the proposed attacks. Each object is furthermore assigned a "frontal" pose by visual inspection, e.g. the frontal pose of the telephone in Figure 1c is that in the upper left corner. It should be noted that this setup is only necessary to enable *replicable studies* of the proposed attacks and to collect data for *defense purposes.* The attacks themselves can be performed by simply rotating objects.



(a) (b) (c)

Figure 2: Turk experiment. (a) True positive (TP) x, shown for 750 ms. (b) Distractor task: count dots of some color. (c) After correctly counting, x' shown for 750 ms. Then, turkers asked if the image (object) has changed.

3.2. Characterizing indistinguishable perturbations

A difficulty of real world attacks, especially those involving larger perturbations, is to guarantee their validity. After all, under extreme viewing angles, objects can be hard to recognize even for humans. The second contribution of this work is a replicable procedure, based on Amazon Turk experiments, to characterize *indistinguishable perturbations* (IP). We start by proposing that perturbations can only be declared indistinguishable within a certain application context. For object recognition, we identify two contexts of interest. The first is *image retrieval* (IR). This addresses the question of whether a person can distinguish two images. However, we do not pursue the forensic definition of distinguishability commonly used in the literature. Instead, we limit the resources available to the person, by asking them to compare the perturbation to an image they have committed to memory. This is more closely related to object recognition than forensic comparisons.

The setup is illustrated in Figure 2. A turker is shown the true positive x for 750 ms (see supplementary materials for details) and asked to memorize it. The object disappears and the turker is asked to count the number of dots of some color in a 2x3 grid. This is a distractor task to prevent a purely iconic matching of image details. A second image x' is finally shown for 750 ms, and the turker is asked to indicate if x' was the *image* seen earlier. The second image can be of four types: the *true positive* x (15% of the time), a perturbation of x due to camera shake (35%), a perturbation of x due to a pose variation (35%), or an image of a *different object* (15%). All images used in this experiment are from the attack dataset of the previous section. From 30 frontal pose images, examples randomly sampled (with replacement) are used as the *true positive* per object. Camera shake perturbations are also "frontal" poses. Given a true positive, an *image pair* is created by sampling one of the 29 remaining frontal poses of the object. This procedure was used to produce 70 camera shake pairs per object. Pose variation perturbations use images from all object views. The set of perturbed images was created by randomly sampling 10 images from each pose, excluding frontal. Finally, for *different object,* examples were sampled randomly from frontal poses of other objects which belong to different classes. A probability of error was recorded per type of x'. These are denoted p^{TP}, p^{CS}, p^{PV}, and p^{DO}, respectively. The values of p^{CS} and p^{PV} are used as *indistinguishable perturbation rates* (IPR) for camera shake and pose variation perturbations. A high indistinguishable perturbation rate implies that turkers are not able to tell apart the perturbed x' from the true positive x. p^{TP} and p^{DO} are upper and lower bounds for the indistinguishable perturbation rate, respectively. For increased accuracy, each image-perturbation pair was evaluated by 3 turkers. A final quality control threshold was imposed per turker: those who scored above 10% for p^{DO} and those who did not score at least 90% for p^{TP} were excluded, as evaluation is trivial in these cases. For the remaining image pairs, those with less than two identical evaluations were eliminated. The indistinguishable perturbation rate was finally determined from the remaining evaluations, by majority vote.

So far, the experiments test if turkers can distinguish perturbed images. This is informative for image retrieval, but ultimately not the goal of object recognition. For example, the rotation of a digit by 30^o is a very perceptible image transformation. While most humans can easily tell the image has been rotated, this makes little difference for recognition. An equally large percentage of the population will be able to effortlessly recognize the rotated digit. In other words, recognition is invariant to the perturbation. We argue that, for recognition, it is also important to define the notion of *semantically indistinguishable perturbations* (SIPs). These are perturbations that may be noticeable but do not alter the image semantics. Semantically indistinguishable perturbations differ from indistinguishable perturbations in that they are tied to the semantics of interest for an application. For example, while a smart car only cares about the presence/absence of pedestrians on the road, a face recognizer aims to determine the person's identity. Hence, replacing the true positive by an image of another person is a semantically indistinguishable perturbation for pedestrian detection but not for face recognition.

An interesting property of the experimental setup above is that it can be extended to semantically indistinguishable perturbations by simply modifying the *context* of the experiment. This is done by changing the *question* asked to the turkers. Rather than asking them if x' is the same *image* as x, they can be asked if it is an image of the same *person, object, animal, scene,* or whatever the semantics of interest are for the application. In this work, we consider the generic *object recognition* (OR) context, asking the turkers if the two images are of the same object. The probabilities p^{CS} and p^{PV} then become *semantically indistinguishable perturbation rates* (SIPR) for camera shake and pose variation perturbations. They capture the degree to which the perturbations are imperceptible for OR. Note that a large pose transformation,

clearly perceptible for image retrieval can easily be imperceptible for object recognition. This difference is captured by the two questions (*is this the same image?* vs. *is this the same object?*) that set different contexts for the experiment.

In summary, the probabilities p^{CS} and p^{PV} can be indistinguishable perturbation rates (IPR) or semantically indistinguishable perturbation rates (SIPR), depending on the context (image retrieval or object recognition respectively). Table 1 summarizes the rates observed on the Turk experiments. Several conclusions can be taken from the table. First, turkers' scores were excellent when spotting replicas of the true positive (IPR $> 99\%$) or rejecting images from different objects (SIPR $\leq 1\%$). This suggests that the experimental protocol is robust. Second, all rates were lower for pose variation than for camera shake. This was expected, because pose variation induces larger image variations. These results confirm the hypothesis that camera shake is a *small* perturbation, while pose variation is a *larger* perturbation. Note that only 7% of the pose variation perturbations were indistinguishable perturbations, while this held for 72% of the camera shake perturbations. Finally, it is clear that indistinguishability depends on context. While only 72% of the camera shake perturbations were indistinguishable perturbations, 92% were semantically indistinguishable perturbations. Similarly, while only 8% of the pose variation perturbations were indistinguishable perturbations, 82% were semantically indistinguishable perturbations.

3.3. Attacks and defenses

The third contribution of this work is a study of the difficulty of defending attacks based on real-world object manipulations. This is based on the image pairs declared as indistinguishable by the Turk experiment [1]. While experiments were performed for both indistinguishable perturbations and semantically indistinguishable perturbations, we report semantically indistinguishable perturbations only, since these are the most relevant perturbations for object recognition. Indistinguishable perturbation results are discussed in supplementary material. Three datasets were used to implement all defenses: 1) a subset of ImageNet containing all object classes used for attacks, denoted "ImageNet," 2) a subset of the defense dataset of Section 3.1 containing only frontal pose images, denoted "Frontal," and 3) the entire defense dataset, denoted "All". Every attack was performed on AlexNet [13], ResNet34 [10], and VGG16 [22].

To evaluate the impact of different object manipulations, the attacks were implemented with both camera shake and pose variation semantically imperceptible perturbations. For each true positive x, the associated perturbation x' was fed to the classifier and recognition rates (RR) r^{CS} and r^{PV} are recorded. Defenses were evaluated under the "arms race" strategy, by synthesizing examples with different attack

[1] All data collected in this work will be made available publicly.

	IPR (%) (Image Retrieval)	SIPR (%) (Object Recognition)
p^{TP}	99.7	99.8
p^{CS}	72.4	91.6
p^{PV}	7.5	81.5
p^{DO}	0.2	1.0

Table 1: Turker imperceptibility rates for true positive (TP), camera shake (CS), pose variation (PV), and different object (DO) pairs. For image retrieval task, *indistinguishable perturbation rates* (IPR) is considered, while for object recognition task, *semantic IPR* (SIPR) is used.

methods and retraining the network on a dataset augmented with these examples. We considered methods from the two broad categories discussed above: 1) transformation based and 2) gradient based.

1. **Transformation based**
   - **Affine:** Random affine transformations with rotation less than 15 degrees.
   - **Blur:** Gaussian blur kernel with random standard deviation in $[0, 0.6]$.
   - **Blur-Affine:** Combination of affine and blur.
   - **Worst-of:** The worst-of-K method of [6]. Ten affine transformations are randomly sampled and the one of highest loss is selected.
   - **Color Jitter:** Image saturation and hue transformation according to [11].
2. **Gradient based**
   - **FGSM:** The fast gradient sign method of [15].
   - **ENS:** The ensemble adversarial training method of [26].
   - **IFGSM:** The iterative fast gradient sign method of [15].
3. **Standard** training is also experimented as baseline for comparison. The standard training method included random cropping and random horizontal flipping. The learning rate was set to 0.001.

4. Experiments

4.1. Implementation

All experiments were conducted with Pytorch. For training process, we found that at most 20 epochs were enough for the classifier to converge, and the maximum number of epochs was set to this value. Vanilla SGD was used as the optimizer and momentum was set to 0.9 for all classifiers.

Table 2: Recognition rates for camera shake and pose variation semantically indistinguishable perturbation attacks, under different defense methods and datasets. Recognition rates are averaged over AlexNet, ResNet34 and VGG16.

		Attack							
		CS	PV	CS	PV	CS	PV	CS	PV
Defense		ImageNet		Frontal		All		Avg	
	None	73.7	47.2	82.0	63.7	87.1	**79.1**	80.9	**63.3**
Transformation	Affine	71.8	45.1	83.4	58.8	85.2	76.5	80.1	60.1
	Blur	74.2	45.2	84.8	**64.1**	86.9	78.3	82.0	62.5
	Blur-Affine	75.4	47.5	83.5	60.0	**88.0**	76.6	82.3	61.3
	Worst-of	73.0	47.1	83.8	63.0	86.4	76.1	81.0	62.0
	Color Jitter	74.5	45.5	**86.4**	61.6	87.1	**79.1**	**82.7**	62.0
	Avg	73.8	46.1	84.4	61.5	86.7	77.3	81.6	61.6
Gradient	FGSM	72.9	**49.2**	84.7	61.1	83.2	74.3	80.3	61.5
	ENS	**75.7**	46.3	83.6	58.1	81.9	72.8	80.4	59.0
	IFGSM	71.8	47.0	82.8	55.5	83.3	70.0	79.3	57.5
	Avg	73.5	47.5	83.7	58.2	82.8	72.3	80.0	59.3

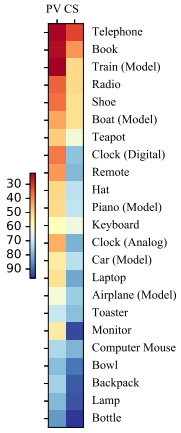


Figure 3: Per class RR for CS/PV SIP attacks.

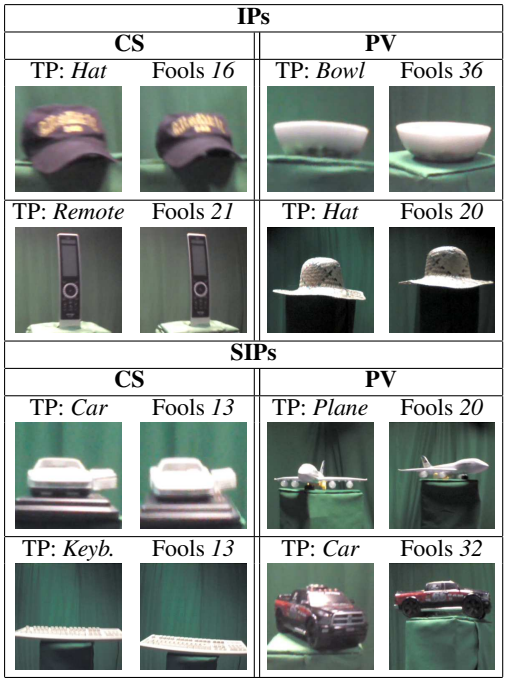
4.2. Qualitative results

Attack and defense efficiency: A preliminary observation was that the attacks had similar effect on the three networks. While some models have higher accuracy than others, the relative drop in accuracy due to the attacks were nearly identical. Hence, for brevity, we only discuss average accuracy of the three models here. More detailed, per-model, results are given in the supplement. Table 2 presents the recognition rates of camera shake and pose variation manipulation attacks, for networks with various defenses. Each defense was implemented on the three defense datasets and recognition rates are presented per defense method and dataset. Since all perturbations have been declared semantically indistinguishable perturbations by turkers, the human recognition rate is 1 on these experiments (under the assumption that turkers could correctly classify the true positive).

Various conclusions can be drawn. First, as expected, *pose variation is the more dangerous attack.* For standard ImageNet classifiers the recognition rate drops to almost half (from 70s to 40s), independently of the defense implemented. Second, *no defense method stands out.* While gradient methods achieve best performance for ImageNet training, transformations have superior performance for Frontal and All training. Within each category, relative performance varies with dataset and perturbation type. On average (as seen in the last column of the table), Color Jitter is the top defense against camera shake. Third, none of the defense algorithms improves significantly on no defense. In fact, *the absence of defense is the best defense against pose variation,* and close to the best (80.9 vs. 82.7) against camera shake, on average. Fourth, *data collection is a much more effective defense than algorithms.* Independently of the algorithm, recognition rates increase significantly from ImageNet to Frontal (10+ points) and increase further from Frontal to All (2 points). However, even the collection of data with camera shake and pose variation perturbations fails to fully defend against real-world object manipulations. The best performance against pose variation (none) has a recognition rate of only 63.3%. For camera shake the top recognition rate is 82.7%. All these observations support the hypothesis that real-world manipulations are a very effective tool to attack CNNs. Besides being trivial to perform, they can be very hard to defend. Since the collection of real data fails to produce a foolproof defense, it is questionable that digital defenses could fully neutralize these attacks. Clearly, simple digital defenses such as Affine or Blur transformations are ineffective.

In-depth comparisons of the table also challenge some common notions in the adversarial literature. One striking effect is the reversal of performance between gradient and transformation based methods with the defense dataset. Gradient methods work better on ImageNet, but are not effective when camera shake and pose variation perturbations are added to the defense set. This supports the hypothesis that they mostly push examples to the edge of the natural image space. Better coverage of these regions, by camera shake and more camera views, eliminates these methods' gains. For example, the average recognition rate of the gradient methods on the All defense dataset is 4 to 7 points weaker than using no defense algorithm at all. Transformation based methods perform significantly better on this dataset. In the adversarial literature, IFGSM and ENS have also been claimed to outperform FGSM. This is because IFGSM generates stronger adversarial examples and ENS decouples the adversarial example generator for the defender (by adding adversarial examples from a third party to the training set). However, this is nearly the opposite of the results on Table 2. On average, FGSM outperforms IFGSM and ENS. Again, this is likely due to the real world nature of the attacks. The fact that IFGSM and ENS are better defenses against digital attacks, seems to translate into no benefits for real world attacks.

Table 3: Examples of IPs and SIPs, for CS and PV perturbations, that fool many classifiers. In all cases, TP is left, perturbation right. Also shown is ground truth class and # of classifiers fooled (out of 81, see supplementary material).



Objects: It is also pertinent to ask which types of objects lead to more successful attacks. Figure 3 shows the recognition rate of camera shake and pose variation perturbations per object class. While camera shake leads to higher recognition rates for all objects, the recognition rate trend is similar for camera shake and pose variation. This suggests that attack efficiency is indeed determined by object properties. Finally, a "lack of symmetry" seems to be the object property most predictive of successful attacks – symmetric objects, such as bottles, lamps, and bowls are less effective attackers than less symmetric objects like telephones, radios, and trains.

Universal attacks: A final question is which attacks fool a large number of classifiers. Table 3 shows some examples of the most successful perturbations from this point of view (more in supplementary material). Some interesting observations can be made. First, perturbations that are clearly noticeable under a forensic comparison (side-by-side images, infinite time) can become indistinguishable under the memory recall paradigm of Figure 2. Take the "bowl" and "hat" examples for instance, which were deemed indistinguishable perturbations by the turkers; the fact that these perturbations were deemed the same *image* as the true positive shows that the standard practice of determining attack validity by forensic comparisons is poorly suited for object recognition. Second, it appears that perturbations of all sizes can fool a large number of models. Note that the perturbations shown range from "insignificant" (almost imperceptible even on a forensic comparison, e.g. "remote") to "large" (significant pose variations, e.g. "car" on the bottom right). Overall, it appears that even very elementary natural perturbations can fool state-of-the-art classifiers.

Table 4: Classifier accuracy for crafted vs random attack examples.

Dataset	ImageNet		Frontal	
Attack	CS	PV	CS	PV
Random	73.7	47.2	82.0	63.7
Crafted	51.3	33.0	66.2	49.3

Crafted attacks: Since the proposed attack happens naturally in the real world, one might criticize that this is different from L_p norm based attacks, which can be designed and crafted. Inspired by [6], which generates attacks by rotating the image and proposes a worst-of-K method to pick the most adversarial transformation , we implemented this for our attacks with $K = 5$ similar to the set up in [6]. Table 4 presents the results for no defense algorithm on random and crafted attack examples. These CS/PV attacks are intentionally crafted, by picking the CS/PV instance most likely to fool the network. They are more effective than the random attacks as expected.

5. Conclusion

This work makes several contributions to the study of adversarial attacks that are easy to execute but difficult to defend, using a new setup based on real-world object manipulations. Unlike the standard practice in the literature, we considered both small and large perturbations, generated by camera shake and pose variation, and introduced a procedure for systematic collection of such perturbations. This was complemented by a replicable procedure to measure the imperceptibility of perturbations, using Turk experiments. These contributions enabled the creation of a dataset of small and large perturbations, imperceptible under two contexts of interest for object recognition. Experimental results comparing defenses based on many datasets, CNN models, and algorithms from the literature elucidated the difficulty of defending these attacks. None of the existing defenses were effective against them, and while better results were achieved with real world data augmentation, this is costly and not foolproof. These results suggest that more research is needed on defenses against "easy to perform" attacks and that the data now assembled can play an important role in this regard.

Acknowledgments This work was partially funded by NSF awards IIS-1546305 and IIS-1637941, a gift from Northrop Grumman, and NVIDIA GPU donations. We thank David Orozco and Amir Persekian for their contributions to the drone and Mark Milam for inspiring this work.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *CoRR*, abs/1707.07397, 2017.

[2] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017.

[3] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016.

[4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[6] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *CoRR*, abs/1712.02779, 2017.

[7] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *CoRR*, abs/1707.08945, 2017.

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. *CoRR*, abs/1804.00499, 2018.

[12] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *CoRR*, abs/1511.03034, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016.

[15] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016.

[16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *CoRR*, abs/1611.02770, 2016.

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *CoRR*, abs/1511.04599, 2015.

[19] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016.

[20] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *CoRR*, abs/1511.07528, 2015.

[21] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. *CoRR*, abs/1703.02921, 2017.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864, 2017.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[26] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[27] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

[28] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *CoRR*, abs/1801.02612, 2018.

[29] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *CoRR*, abs/1712.07107, 2017.

[30] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *CoRR*, abs/1710.11342, 2017.