Statistical Visual Computing Lab
UC San Diego

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

Chih-Hui Ho, Brandon Leung, Erik Sandström, Yen Chang, Nuno Vasconcelos

CVPR

Introduction

- Recently, the robustness of CNNs have been questioned by adversarial attacks -- imperceptible perturbations added to the original image, such that the CNN classifies incorrectly.
- Most attacks are imperceptible under some arbitrarily small perturbation (e.g. defined by an L_p norm). We introduce two natural perturbations reframed under an adversarial context, based on human perception, which allows study of large and small attacks.
- A new image dataset depicting objects under camera shake and pose change is presented. Collected with drones, it has large overlap with ImageNet classes to enable attacks on ImageNet trained CNNs.
- A dataset of image pairs deemed imperceptible under the proposed methodology is provided.
- Ultimately, current CNNs are vulnerable to attacks implementable even by a child, and such attacks may prove difficult to defend.

Previous Related Work

Some examples of popular adversarial attacks. While efforts are made for indistinguishability, human imperceptibility is not quantified unlike in the work presented here.

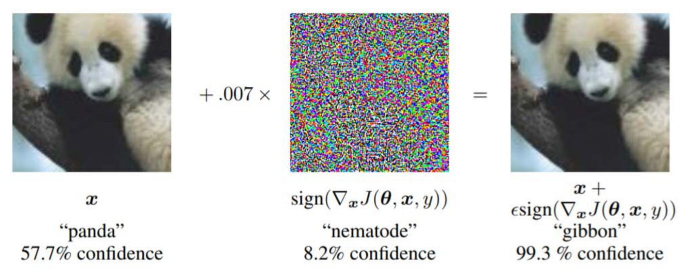

Figure 1. One of the first adversarial examples (Goodfellow Et al.[1]).

Figure 2. A physical adversarial attack, from Eykholt Et al.[2]

Figure 3. A targeted, real world adversarial attack; a 3D printed turtle designed to classify as "rifle" (from Athalye Et al.[3]).

Dataset Collection

Image Dataset Composition

- Pictures of 500 objects at 8 different angles, taken by drones. Each object has a predefined frontal angle.
- 30 images taken per angle, total of 120,000 images.
- Objects are evenly divided into 25 classes, such as "backpacks", "bottles", and "shoes".
- Each picture annotated with class, pose, blurriness level (0 to 2), and bounding box.

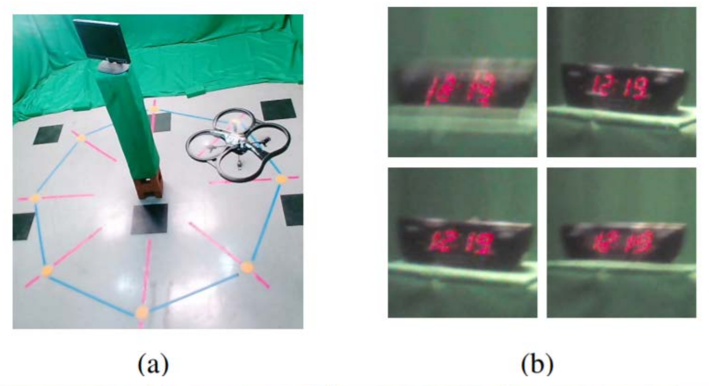

Figure 4.
(a) Drone capturing images during flight.
(b) Examples of varying levels of camera shake as the drone hovers.
(c) Example images collected per viewing angle.

Imperceptibility Annotation

- Turkers are presented pairs of images, and asked one of two questions.
  1) *"Are these images identical?"* If so, we have an "Imperceptible Perturbation"(IP): Pictures appear the same, down to the pixel level.
  2) *"Are the objects in these images the same?"* If so, we have an "Semantically Imperceptible Perturbation" (SIPs): Image pairs are clearly different, but the objects depicted are the same.
- A simple distraction task is presented between showing the two images to prevent trivial memorization.

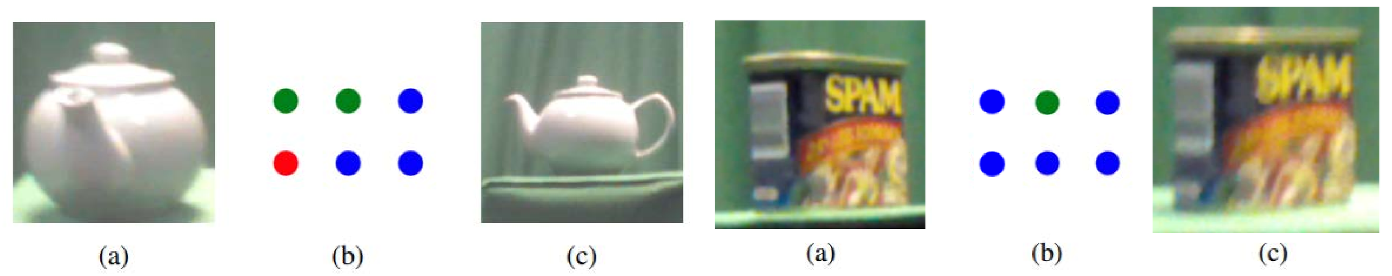

Figure 5. Two example pairs from the Turk experiment. (a) is shown for 750 ms and disappears afterward. (b) Distractor task: count dots of some color. (c) is then presented, along with the question.

Experiments & Findings

Attacks and Defenses

- We try to attempt real-world manipulations attacks on CNNs, with indistinguishable image pairs (table 1).
- Various current defense methods used (figure 6).
- Training is on either ImageNet, only frontal images of defense dataset, or the entire defense dataset.
- Pose variation is the most dangerous attack, and no current defense are completely effective. Instead, data collection seems to be most beneficial.
- Gradient defenses are less effective when camera shake and pose images are added. This supports the hypothesis that gradient defenses mostly push examples to the edge of the natural image space, which are useful in traditional attacks but not in the case of natural perturbations.

		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

Table 1. Recognition rates for camera shake and pose variation attacks, under several defense and training datasets. Averaged over AlexNet, ResNet34 and VGG16.

Transformation Defenses

- **Affine**: Random affine transformations with rotation less than 15 degrees.
- **Blur**: Gaussian blur kernel with random standard deviation in [0, 0.6].
- **Blur-Affine**: Affine and blur.
- **Worst-of**: The worst-of-K method of [4]. Ten affine transformations are randomly sampled and the one of highest loss is selected.
- **Color Jitter**: Image saturation and hue transformation according to [5].

Gradient Defenses

- **FGSM**: Fast gradient sign method [6].
- **ENS**: The ensemble adversarial training method of [7].
- **IFGSM**: The iterative fast gradient sign method of [8].

Figure 6. Methods used to defend against adversarial attacks.

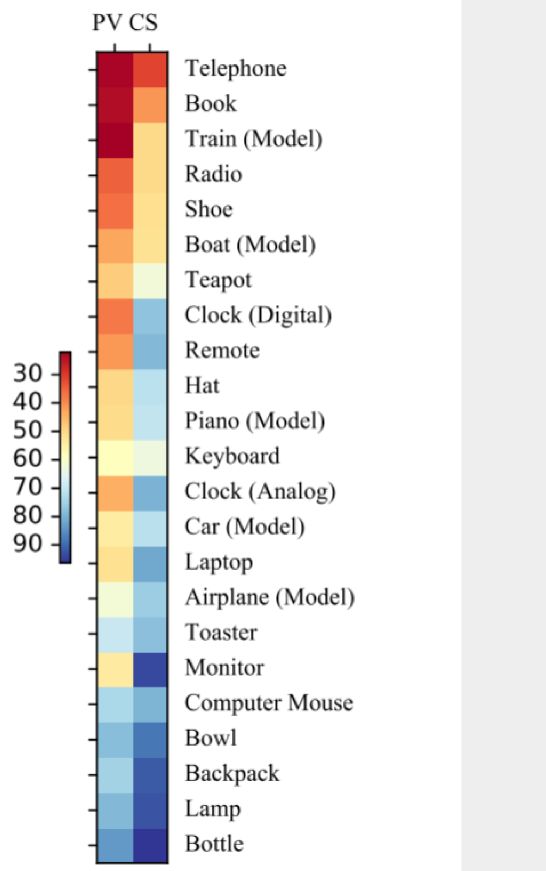

Figure 7. Class recognition rates, for PV (pose variation) and CS (camera shake) attacks.

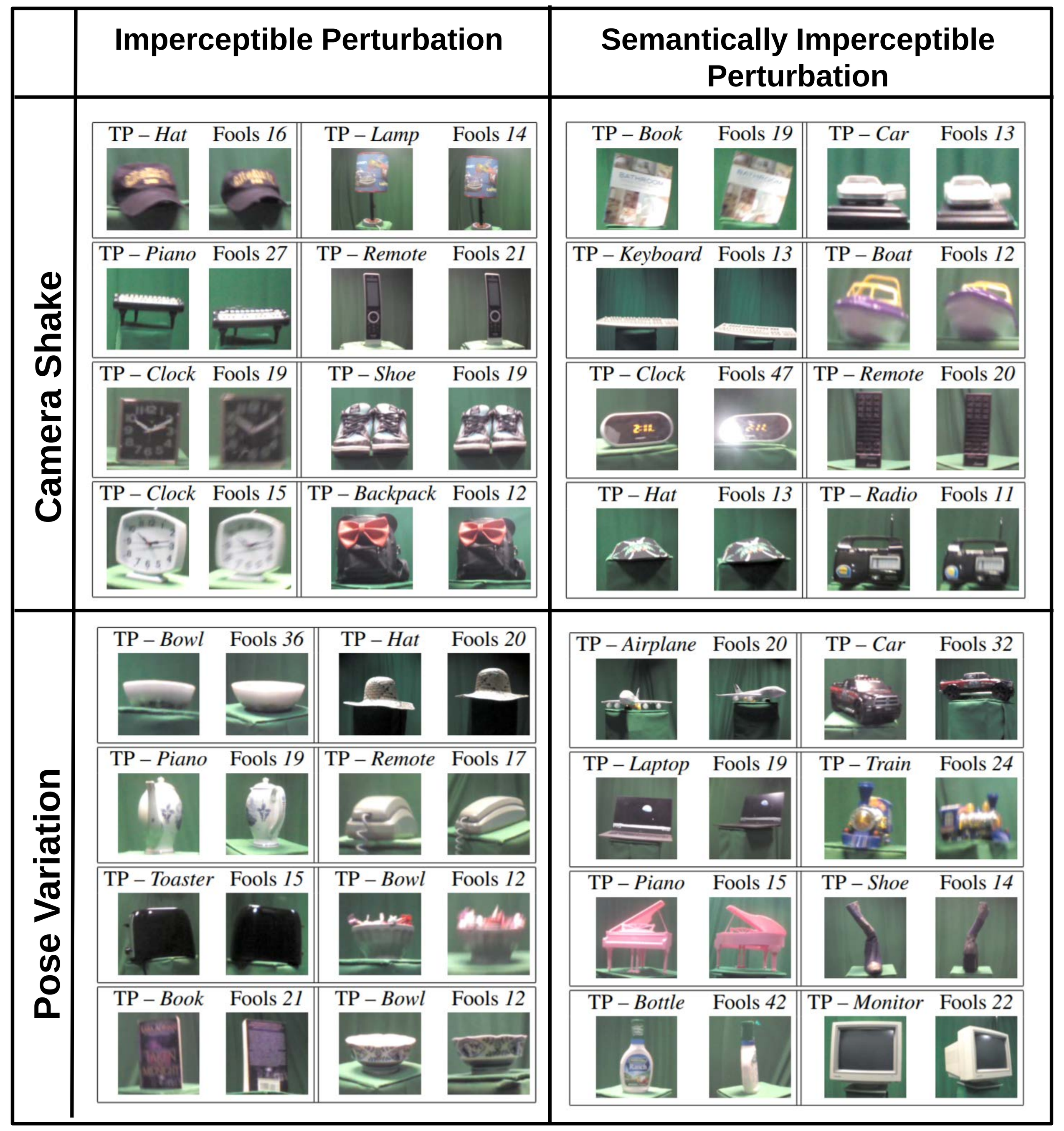

Figure 8. Examples of various adversarial attacks, fooling classifiers constructed with AlexNet, VGG, or ResNet with defenses in table 1. Perturbations of all sizes can fool a large number of models.

Conclusion

- A new dataset is used to study a class of human-based, *semantically* imperceptible attacks.
- Unlike previous works, we study both small and large perturbations based on camera shake and pose variation. A new framework is used to characterize imperceptibility.
- We show that these attacks proposed are easy to execute, but difficult to defend.
- The Amazon Turk based framework can be used to characterize many other types of future attacks.

References

[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.
[2] 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.
[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. CoRR, abs/1707.07397, 2017.
[4] 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.
[5] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. CoRR, abs/1804.00499, 2018.
[6] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. CoRR, abs/1611.01236, 2016.
[7] 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.
[8] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. CoRR, abs/1611.01236, 2016.

Dataset available at: svcl.ucsd.edu/projects/OOWL