Adversarial Examples that Fool both Human and Computer Vision

Gamaleldin F. Elsayed [1 2] **Shreya Shankar** [2 3] **Brian Cheung** [2 4] **Nicolas Papernot** [2 5] **Alex Kurakin** [2]
Ian Goodfellow [2] **Jascha Sohl-Dickstein** [2]

Abstract

Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we create the first adversarial examples designed to fool humans, by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by modifying models to more closely match the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

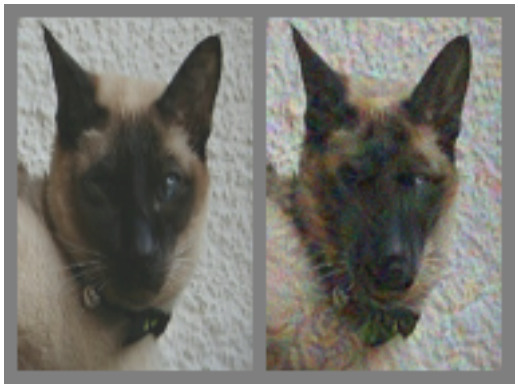

Figure 1. While in most cases our adversarial examples fool humans only after a brief exposure, the example depicted has a strong effect even for long viewing duration. On the left, we show an image of a cat. On the right, we show the same image after it has been adversarially perturbed to look like a dog. Although easily overlooked, note that cat-specific features can still be identified. For instance, the original boundary between the top of the cat head and the wall is still visible in the adversarial image, despite the top of the dog head seeming to be lower. Also, long white cat whiskers remain visible.

1. Introduction

Machine learning models are easily fooled by adversarial examples: inputs optimized by an adversary to produce an incorrect model classification (Szegedy et al., 2013; Biggio et al., 2013). In computer vision, an adversarial example is usually an image formed by making small perturbations to an example image from a dataset. Many popular algorithms for constructing adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014; Papernot et al., 2015; Kurakin et al., 2016; Madry et al., 2017) rely on access to both the architecture and the parameters of the model to perform gradient-based optimization on the input. Without similar access to the brain, these methods do not seem applicable to constructing adversarial examples for humans.

One interesting phenomenon is that adversarial examples often transfer from one model to another, making it possible to attack models that an attacker has no access to (Szegedy et al., 2013; Liu et al., 2016). This naturally raises the question of whether humans are susceptible to these adversarial examples. Clearly, humans are prone to many cognitive biases and optical illusions (Hillis et al., 2002), but these generally do not resemble small perturbations of natural images, nor are they currently generated by optimization of a machine learning loss function. Thus the current understanding in the field is that this class of transferable adversarial examples has no effect on human visual perception, yet no thorough empirical investigation has yet been performed.

A rigorous investigation of the above question creates an opportunity both for machine learning to gain knowledge from neuroscience, and for neuroscience to gain knowledge from machine learning. Neuroscience has often provided existence proofs for machine learning—before we had working object recognition algorithms, we hypothesized it should be possible to build them because the human brain can recognize objects. See Hassabis et al. (2017) for a review of the influence of neuroscience on artificial intelligence. If

[1]Work done as a member of the Google AI Residency program (g.co/airesidency). [2]Google Brain [3]Stanford University [4]UC Berkeley [5]Pennsylvania State University. Correspondence to: Gamaleldin F. Elsayed <gamaleldin@google.com>, Jascha Sohl-Dickstein <jaschasd@google.com>.

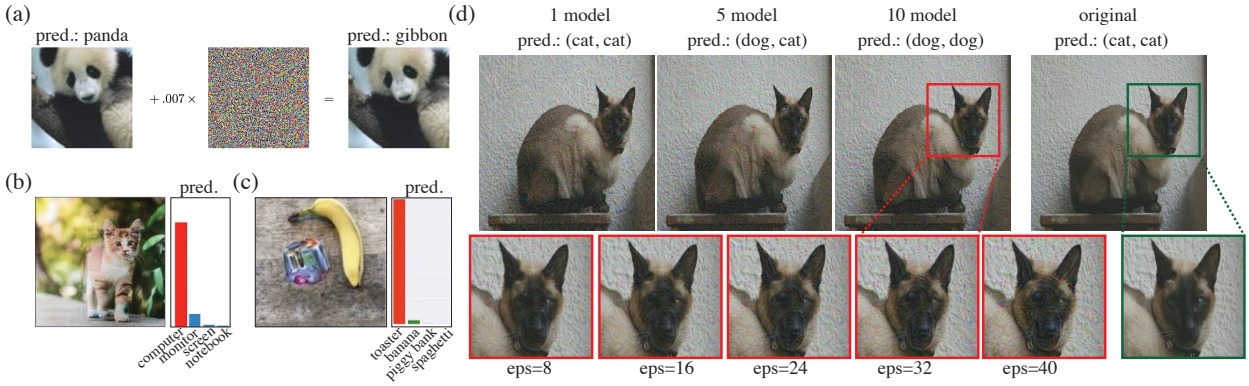

Figure 2. **Adversarial examples optimized on more models / viewpoints sometimes appear more meaningful to humans.** This observation is a clue that machine-to-human transfer may be possible. (a) A canonical example of an adversarial image reproduced from Goodfellow et al. (2014). This adversarial attack has moderate but limited ability to fool the model after geometric transformations or to fool models other than the model used to generate the image. (b) An adversarial attack causing a cat image to be labeled as a computer while being robust to geometric transformations, adopted from Athalye (2017). Unlike the attack in a, the image contains features that seem semantically computer-like to humans. (c) An adversarial patch that causes images to be labeled as a toaster, optimized to cause misclassification from multiple viewpoints, reproduced from Brown et al. (2017). Similar to b, the patch contains features that appear toaster-like to a human. (d) In our experiments, we find a similar effect when adversarial images are generated to fool multiple models, rather than to fool the same model from multiple viewpoints. The images presented here correspond to a sequence of adversarial attacks to classify a cat image as a dog. top: left to right, the attack is performed against a larger ensemble of models (original image on the right). The class predictions of two test models are included above each image. As the number of models targeted by the attack increases, the resulting image appears more dog-like to humans. Bottom: the attack magnitude ϵ (see Section 3.1.3) is varied for an attack against all 10 models. The image appears somewhat more dog-like to humans even for $\epsilon = 8$.

we knew conclusively that the human brain could resist a certain class of adversarial examples, this would provide an existence proof for a similar mechanism in machine learning security. If we knew conclusively that the brain can be fooled by adversarial examples, then machine learning security research should perhaps shift its focus from designing models that are robust to adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014; Papernot et al., 2016c; Xu et al., 2017; Tramèr et al., 2017; Madry et al., 2017; Kolter & Wong, 2017; Jacob Buckman, 2018) to designing systems that are secure despite including non-robust machine learning components. Likewise, if adversarial examples developed for computer vision affect the brain, this phenomenon discovered in the context of machine learning could lead to a better understanding of brain function.

In this work, we investigate the influence of adversarial examples that strongly transfer across computer vision models on human visual perception. We leverage three key ideas to test whether adversarial examples can cause an observable effect on the human visual system. First, we use the recent **black box** adversarial example construction techniques that create adversarial examples for a target model without access to the model's architecture or parameters. Second, we adapt machine learning models to mimic the initial visual processing of humans, making it more likely that adversarial examples will transfer from the model to a human observer. Third, we evaluate classification decisions of human observers in a time-limited setting, so that even subtle effects on human perception are detectable. In other words, when humans can achieve near-perfect accuracy on the classification task, small changes in their performance may not correspond to measurable changes in accuracy. By making image presentation sufficiently brief, humans are unable to achieve perfect accuracy even on clean images, and small changes in performance lead to more measurable changes in accuracy. Additionally, a brief image presentation limits the time in which the brain can utilize recurrent and top-down processing pathways (Potter et al., 2014), and is believed to make the processing in the brain more closely resemble that in a feedforward artificial neural network.

We find that adversarial examples that transfer across computer vision models *do* successfully influence the perception of human observers, thus uncovering a new class of illusions that are shared between computer vision models and the human brain.

2. Background and Related Work

2.1. Adversarial Examples

Goodfellow et al. (2017) define adversarial examples as

"inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake." In the context of visual object recognition, adversarial examples are images usually formed by applying a small perturbation to a naturally occurring image in a way that breaks the predictions made by a machine learning classifiers. See Figure 2a for a canonical example where adding a small perturbation to an image of a panda causes it to be misclassified as a gibbon. This perturbation is small enough to be imperceptible (i.e., it cannot be saved in a standard png file that uses 8 bits as the perturbation is smaller than $1/255$ of the pixel dynamic range). This perturbation is not noise—it relies on carefully chosen structure based on the parameters of the neural network—but when magnified to be perceptible, human observers cannot recognize any meaningful structure. Note that adversarial examples also exist in other domains like malware detection (Grosse et al., 2017), but we focus here on image classification tasks.

Two aspects of the definition of adversarial examples are particularly important for this work:

1. Adversarial examples are designed to cause a *mistake*. They are not (as is commonly misunderstood) defined to differ from human judgment. If adversarial examples were defined by deviation from human output, it would by definition be impossible to make adversarial examples for humans. On some tasks, like predicting whether input numbers are prime, there is a clear objectively correct answer, and we would like the model to get the correct answer, not the answer provided by humans (time-limited humans are probably not very good at guessing whether numbers are prime). It is challenging to define what constitutes a mistake for visual object recognition, since after adding a perturbation to an image it likely no longer corresponds to a photograph of a real physical scene, and it is philosophically difficult to define the real object class for an image that is not a picture of a real object. In this work, we assume that an adversarial image is misclassified if the output label differs from the human-provided label of the clean image that was used as the starting point for the adversarial examples. We make small adversarial perturbations and we assume that these small perturbations are insufficient to change the true class.

2. Adversarial examples are not (as is commonly misunderstood) defined to be imperceptible. If this were the case, it would be impossible by definition to make adversarial examples for humans, because changing the human's classification would constitute a change in what the human perceives. Moreover, in many domains, it is not possible to make imperceptible changes (e.g., in natural language processing, changing even one character of text is perceptible (Papernot et al., 2016b)). Computer vision algorithms can often be fooled by adversarial examples that are imperceptible to humans, but this property is not part of the general definition (also see Figure 2b,c).

2.1.1. Adversarial Examples Transfer

Adversarial examples that fool one model often fool another model with a different architecture (Szegedy et al., 2013), another model that was trained on a different training set (Szegedy et al., 2013), or even trained with a different algorithm (Papernot et al., 2016a) (e.g., adversarial examples designed to fool a convolution neural network may also fool a decision tree). The transfer effect makes it possible to perform black box attacks, where adversarial examples fool models that an attacker does not have access to (Szegedy et al., 2013; Papernot et al., 2017). Kurakin et al. (2016) found that adversarial examples transfer from the digital to the physical world, despite many transformations such as lighting and camera effects that modify their appearance when they are photographed in the physical world. Liu et al. (2016) showed that the transferability of an adversarial example can be greatly improved by optimizing it to fool *many* machine learning models rather than an individual machine learning model: an adversarial example that fools five models used in the optimization process is extremely likely to fool an arbitrary sixth model.

2.1.2. Clues that Transfer to Humans is Possible

Some observations give clues that transfer to humans may be possible for adversarial examples that transfer across viewpoints or models. Adversarial perturbations generated for an individual machine learning model or from a single viewpoint typically do not appear meaningful to humans. However, recent studies on adversarial examples that transfer across multiple settings have sometimes produced adversarial examples that appear more meaningful to human observers. For instance, a cat adversarially perturbed to resemble a computer (Athalye & Sutskever, 2017) while transfering across geometric transformations develops features that appear computer-like (Figure 2b), and the 'adversarial toaster' from Brown et al. (2017) possesses features that seem toaster-like (Figure 2c). We observe a similar effect in our own experiments when an adversarial example is forced to transfer across an ensemble of models, rather than across geometric transformations (Figure 2d).

This development of human-meaningful features is consistent with the adversarial example coming closer to fooling humans. However, this should be interpreted cautiously, as it is also consistent with the adversarial example simply coming closer to being a real example from the adversarial target class (where the definition of 'real' further requires

addressing the difficult question of how to assign ground truth class labels to images which were not generated by photographing a real object).

2.2. Biological and Artificial Vision

2.2.1. Similarities

Recent research has found similarities in representation and behavior between deep convolutional neural networks (CNNs) and the primate visual system (Cadieu et al., 2014). This further motivates the possibility that adversarial examples may transfer from computer vision models to humans. Cadieu et al. (2014); Yamins & DiCarlo (2016) observed that activity in deeper CNN layers is predictive of activity recorded in the visual pathway of primates. Riesenhuber & Poggio (1999) developed a model of object recognition in cortex that closely resembles many aspects of modern CNNs. Kümmerer et al. (2014; 2017) showed that CNNs are predictive of human gaze fixation. Style transfer (Gatys et al., 2015) demonstrated that intermediate layers of a CNN capture notions of artistic style which are meaningful to humans. Freeman & Simoncelli (2011) used representations in a CNN-like model to develop psychophysical metamers, which are indistinguishable to humans when viewed briefly and with carefully controlled fixation. Geirhos et al. (2017); Rajalingham et al. (2018) performed psychophysics experiments comparing the pattern of errors made by humans, to that made by neural network classifiers.

2.2.2. Notable Differences

Differences between machine and human vision occur in the early sensory apparatus. Images are typically presented to CNNs as a static rectangular pixel grid with constant spatial resolution. The primate eye on the other hand has an eccentricity dependent spatial resolution. Resolution is high in the fovea, or central $\sim 5^\circ$ of the visual field, but falls off linearly with increasing eccentricity (Van Essen & Anderson, 1995). A high spatial frequency perturbation in the periphery of an image, as might occur as part of an adversarial example, would be undetectable by the eye, and thus would have no impact on human perception. Further differences include the sensitivity of the eye to temporal as well as spatial features, as well as non-uniform color sensitivity (Land & Nilsson, 2012). Modeling the early visual system continues to be an area of active study (Olshausen, 2013; McIntosh et al., 2016). As we describe in section 3.1.2, we mitigate some of these differences by using a biologically-inspired image input layer.

Beyond early visual processing, there are more major computational differences between CNNs and the human brain. All the CNNs we consider are fully feedforward architectures, while the visual cortex has many times more feedback than feedforward connections, as well as extensive recurrent dynamics (Olshausen, 2013). Possibly due to these differences in architecture, humans have been found experimentally to make classification mistakes that are qualitatively different than those made by deep networks (Eckstein et al., 2017). Additionally, the brain does not treat a scene as a single static image, but actively explores it with saccades (Ibbotson & Krekelberg, 2011). As is common in psychophysics experiments (KovAcs et al., 1995), we mitigate these differences in processing by limiting both the way in which the image is presented, and the time which the subject has to process it, as described in section 3.2.

3. Methods

Section 3.1 details our machine learning vision pipeline. Section 3.2 describes our psychophysics experiment to evaluate the impact of adversarial images on human subjects.

3.1. The Machine Learning Vision Pipeline

3.1.1. Dataset

In our experiment, we used images from ImageNet (Deng et al., 2009). ImageNet contains 1,000 highly specific classes that typical people may not be able to identify, such as "Chesapeake Bay retriever." Thus, we combined some of these fine classes to form six coarse classes we were confident would be familiar to our experiment subjects ($\{$dog, cat, broccoli, cabbage, spider, snake$\}$). We then grouped these six classes into the following groups: (i) **Pets** group (dog and cat images); (ii) **Hazard** group (spider and snake images); (iii) **Vegetables** group (broccoli and cabbage images).

3.1.2. Ensemble of Models

We constructed an ensemble of k CNN models trained on ImageNet ($k = 10$). Each model is an instance of one of these architectures: Inception V3, Inception V4, Inception ResNet V2, ResNet V2 50, ResNet V2 101, and ResNet V2 152 (Szegedy et al., 2015; 2016; He et al., 2016). To better match the initial processing of human visual system, we prepend each model input with a retinal layer, which incorporates some of the transformations performed by the human eye. In that layer, we perform an eccentricity dependent blurring of the image to approximate the input which is received by the visual cortex of human subjects through their retinal lattice. The details of this model are described in Appendix B. We use eccentricity-dependent spatial resolution measurements from Van Essen & Anderson (1995) (based on the macaque visual system), along with the known geometry of the viewer and the screen, to determine the degree of spatial blurring at each image location to limit the CNN to information which is also available to the human visual system. The layer is fully differentiable, allowing gra-

dients to backpropagate through the network when running adversarial attacks. Further details of the models and their classification performance are provided in Appendix D.

3.1.3. Generating Adversarial Images

For a given image group, we wish to generate targeted adversarial examples that strongly transfer across models. This means that for a class pair (A, B) (e.g., A: cats and B: dogs), we generate adversarial perturbations such that models will classify perturbed images from A as B; similarly, we perturbed images from B to be classified as A. A different perturbation is constructed for each image; however, the ℓ_∞ norm of the perturbations are constrained to some ϵ.

Formally: given a classifier which assigns probability $F(y \mid X)$ to each coarse class y given an input image X, a specified target class y_{target} and a maximum perturbation ϵ, we want to find the image X_{adv} that minimizes $-\log(F(y_{\text{target}} \mid X_{adv}))$ with the constraint that $||X_{adv} - X||_\infty \leq \epsilon$. See Appendix C for details on computing the coarse class probabilities $F(y \mid X)$. With the classifier's parameters, we can perform iterated gradient descent on X in order to generate our X_{adv}.

In the pipeline, an image is drawn from the source coarse class and perturbed to be classified as an image from the target coarse class. The attack method we use, the iterative targeted attack (Kurakin et al., 2016), is performed as

$$\tilde{X}^n_{adv} = X^{n-1}_{adv} - \alpha * \text{sign}(\nabla_{X^n}(J(X^n|y_{\text{target}}))),$$
$$X^n_{adv} = \text{clip}\left(\tilde{X}^n_{adv}, [X - \epsilon, X + \epsilon]\right), \quad (1)$$

where J is the cost function as described below, y_{target} is the label of the target class, α is the step size, $X^0_{adv} = X$ is the original clean image, and $X_{adv} = X^N_{adv}$ is the final adversarial image. We set $\alpha = 2$, and ϵ is given per-condition in Section 3.2.2. After optimization, any perturbation whose ℓ_∞-norm was less than ϵ was scaled to have ℓ_∞-norm of ϵ, to guarantee a consistent norm for all perturbations.

Our goal was to create adversarial examples that transferred across many machine learning models before assessing their transferability to humans. To accomplish this, we created an ensemble from the geometric mean of several image classifiers, and performed the iterative attack on the ensemble loss (Liu et al., 2016)

$$J\left(X|y_{target}\right) = -\log\left[P_{\text{ens}}\left(y_{\text{target}}|X\right)\right], \quad (2)$$
$$P_{\text{ens}}\left(y|X\right) \propto \exp\left(\mathbb{E}_k[\log F_k\left(y|X\right)]\right), \quad (3)$$

where $F_k\left(y|X\right)$ is the coarse class probabilities from model k, and $P_{\text{ens}}\left(y|X\right)$ is the probability from the ensemble.

To encourage a high transfer rate, we retained only adversarial examples that were successful against all 10 models for the `adv` condition and at least $7/10$ models for the `false` condition (see Section 3.2.2 for condition definitions).

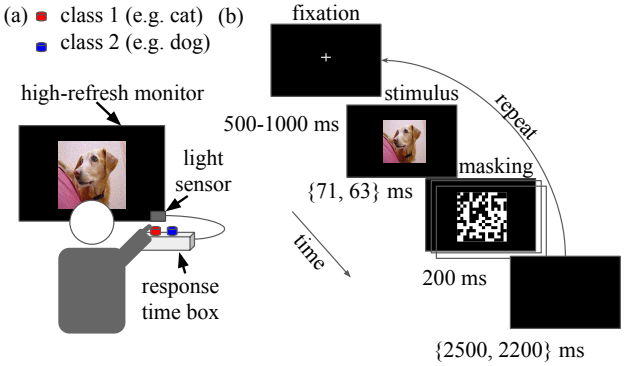

Figure 3. **Experiment setup and task.** (a) Experiment setup and recording apparatus. (b) Task structure and timings. The subject is asked to repeatedly identify which of two classes (e.g. dog vs. cat) a briefly presented image belongs to. The image is either adversarial, or belongs to one of several control conditions. See Section 3.2 for details.

3.2. Human Psychophysics Experiment

A total of 38 subjects with normal or corrected vision participated in the experiment. Subjects gave informed consent to participate in the study, and were awarded a reasonable compensation for their time and effort[1].

3.2.1. Experimental Setup

Subjects sat on a fixed chair 61cm away from a high refresh-rate computer screen (ViewSonic XG2530) in a room with dimmed light (Figure 3a). Subjects were asked to classify images that appeared on the screen to one of two classes (two alternative forced choice) by pressing buttons on a response time box (LOBES v5/6:USTC) using two fingers on their right hand. The assignment of classes to buttons was randomized for each experiment session. Each trial started with a fixation cross displayed in the middle of the screen for $500 - 1000$ ms, instructing subjects to direct their gaze to the fixation cross (Figure 3b). After the fixation period, an image of size 15.24 cm $\times$ 15.24 cm ($14.2°$ visual angle) was presented briefly at the center of the screen for a period of 63 ms (71 ms for some sessions). The image was followed by a sequence of ten high contrast binary random masks, each displayed for 20 ms (see example in Figure 3b). Subjects were asked to classify the object in the image (e.g., cat vs. dog) by pressing one of two buttons starting at the image presentation time and lasting until 2200 ms (or 2500 ms for some sessions) after the mask was turned off. The wait to start the next trial was the same whether subjects responded quickly or slowly. Realized exposure durations were $\pm$4ms from the times reported above, as measured by

[1]The study was granted an Institutional Review Board (IRB) exemption by an external, independent, ethics board (Quorum review ID 33016).

a photodiode and oscilloscope in a separate test experiment. Each subject's response time was recorded by the response time box relative to the image presentation time (monitored by a photodiode). In the case where a subject pressed more than one button in a trial, only the class corresponding to their first choice was considered. Each subject completed between 140 and 950 trials.

3.2.2. Experiment Conditions

Each experimental session included only one of the image groups (Pets, Vegetables or Hazard). For each group, images were presented in one of four conditions as follows:

- `image`: images from the ImageNet training set (rescaled to the $[40, 255-40]$ range to avoid clipping when adversarial perturbations are added).
- `adv`: `image` with an added adversarial perturbation δ_{adv}, crafted to cause model misclassification of `adv` as the opposite class in the group (e.g., if `image` was originally cat class, we perturbed the image to be classified as a dog). We used a perturbation size large enough to be noticeable by humans on the computer screen but still small with respect to the image intensity scale ($\epsilon = 32$).
- `flip`: similar to `adv`, but the adversarial perturbation (δ_{adv}) was flipped vertically before adding it to `image`. This is a control condition, to test whether adversarial perturbations change human perceptions more than non-adversarial perturbations with nearly identical statistics.
- `false`: in this condition, the subject is forced to make a mistake. We include this condition because if adversarial perturbations *reduce the accuracy* of human observers, this could just be because the perturbations degrade the image quality. To show that adversarial perturbations *actually control the chosen class* we include this condition where neither of the two options available to the subject is correct so their accuracy is always zero, and test whether adversarial perturbations can influence which of the two wrong choices they make. We show a random image from an ImageNet class other than the two classes in the group, and adversarially perturb it toward one of the two classes in the group. The subject must then choose from these two classes. For example, we might show an airplane adversarially perturbed toward the dog class, while a subject is in a session classifying images as cats or dogs. We used a slightly larger perturbation in this condition ($\epsilon = 40$).

The conditions (`image`, `adv`, `flip`) are ensured to have balanced number of trials within a session by either uniformly sampling the condition type in some of the sessions or randomly shuffling a sequence with identical trial counts for each condition in other sessions. The number of trials for each class in the group was also constrained to be equal. Similarly for the `false` condition the number of trials adversarially perturbed towards class 1 and class 2 were balanced for each session. To reduce subjects using strategies based on overall color or brightness distinctions between classes, we pre-filtered the dataset to remove images that showed an obvious effect of this nature. Most significantly, in the pets group we excluded images that included large green lawns or fields, since in almost all cases these were photographs of dogs. The list of images we included in the experiment for each coarse class is given in Appendix E. For examples images for each condition, see Figures 5, and Supp.3 through Supp.6.

4. Results

4.1. Adversarial Examples Transfer to Computer Vision Models

We first assess the transferability of our constructed images to two test models that were not included in the ensemble used to generate adversarial examples. Specifically, our test models are: an adversarially trained Inception V3 model (Kurakin et al., 2016), denoted as Inception V3**, and a ResNet V2 50 model. We measure the *attack success* on these models, defined as the fraction of images that are misclassified by the model as coming from the target classes. We also report classification accuracy in the Table Supp.4.

For the `image` condition, the accuracy of the ten models in the ensemble from the three groups (pets, hazard and vegetables) was 100% as we pre-filtered images to ensure that clean images are correctly classified by all models in the ensemble. For the two test models that we used, the accuracy on `image` from pets, hazard, and vegetables groups was very high (Resnet V2 50: 99%, 98%, 96%; Inception V3**: 99%, 99%, 100%, respectively; see accuracy of all models in other conditions in Table Supp.4 and Table Supp.3). We measured the success of our targeted attacks. All models had very low attack success rate on `flip` images (see Table 1 and Table Supp.5). For the `adv` images, the attack success was at 100% for the model ensemble as designed by our adversarial examples generation pipeline (Table Supp.5). More importantly, the attack success on the two test models was also generally high (see Table 1). Consistent with previous work (Liu et al., 2016), these results demonstrate that iterative fast gradient sign method on a large ensemble can generate very strong black-box adversarial attacks.

Table 1. **Attack success on test models.** ** model trained on both clean and adversarial images. Numbers triplet is error on pets, hazard, and vegetables groups, respectively.

Model	`adv` (%)	`flip` (%)
Resnet V2 50	87, 85, 57	1.3, 0.0, 0.0
Inception V3**	89, 87, 74	1.5, 0.5, 0.0

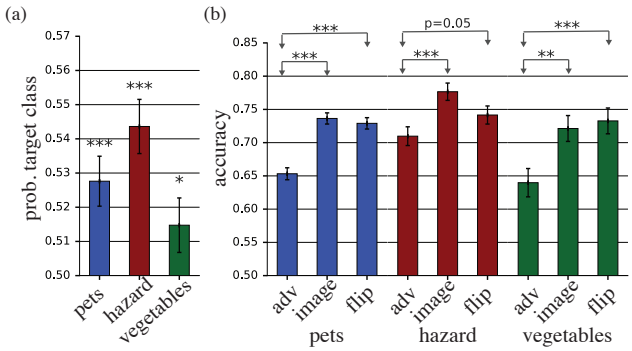

Figure 4. **Adversarial images transfer to humans.** (a) By adding adversarial perturbations to an image, we are able to bias which of two incorrect choices subjects make. Plot shows probability of choosing the adversarially targeted class when the true image class is not one of the choices that subjects can report (`false` condition), estimated by averaging the responses of all subjects (two-tailed t-test relative to chance level 0.5). (b) Adversarial images cause more mistakes than either clean images or images with the adversarial perturbation flipped vertically before being applied. Plot shows probability of choosing the true image class, when the true class is one of the choices that subjects can report, estimated by averaging across all subjects. Accuracy is significantly less than 1 even for clean images due to the brief image presentation time. (error bars $\pm$ SE; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

4.2. Adversarial Perturbations Bias Human Perception

As described in Section 3.2.2, we used the `false` condition to test whether adversarial perturbations can influence which of two incorrect classes a subject chooses (see example images in Figure Supp.3). We do this because if we measured only our ability to cause the subject to choose an incorrect class rather than the correct class, the result could be due to degrading the image quality / discarding information, rather than influencing the specific decision of the subject.

We measured our effectiveness at changing the perception of subjects by measuring the rate that subjects reported the adversarially targeted class. If the adversarial perturbation were completely ineffective we would expect the choice of targeted class to be uncorrelated with the subject's reported class. The average rate at which the subject chooses the target class metric would be 0.5 as each `false` image can be perturbed to class 1 or class 2 in the group with equal probability. Figure 4a shows the probability of choosing the target class averaged across all subjects for all the three experiment groups. In all cases, the probability was significantly above the chance level of 0.5. This demonstrates that the adversarial perturbations generated using CNNs biased human perception towards the targeted class. This effect was stronger for the the hazard, then pets, then vegetables group. This difference in probability among the class groups was significant ($p < 0.05$; one-way ANOVA test).

We also observed a significant difference in the mean response time between the class groups ($p < 0.001$; one-way ANOVA test; see Figure Supp.1a). Interestingly, the response time pattern across image groups (Figure Supp.1a)) was inversely correlated to the perceptual bias pattern (Figure 4a)) (Pearson correlation $= -1$, $p < 0.01$; two-tailed Pearson correlation test). In other words, subjects made quicker decisions for the hazard group, then pets group, and then vegetables group. This is consistent with subjects being more confident in their decision when the adversarial perturbation was more successful in biasing subjects perception. This inverse correlation between attack success and response time was observed within group, as well as between groups (Figure Supp.2).

4.3. Adversarial Examples Transfer to Humans

We demonstrated that we are able to bias human perception to a target class when the true class of the image is not one of the options that subjects can choose. Now we show that adversarial perturbations can be used to cause the subject to choose an incorrect class even though the correct class is an available response. As described in Section 3.2.2, we presented `image`, `flip`, and `adv`.

Most subjects had lower accuracy in `adv` than `image` (Table Supp.1). This is also reflected on the average accuracy across all subjects, which was significantly lower for the `adv` than `image` (Figure 4b).

The above result may simply imply that the signal to noise ratio in the adversarial images is lower than that of clean images. We already partially addressed this objection with the `false` experiment results in Section 4.2. Additionally, we tested accuracy on `flip` images. This control case uses perturbations with identical statistics to `adv` up to a flip of the vertical axis. However, this control breaks the pixel-to-pixel correpsondence between the adversarial perturbation and the image. The majority of subjects had lower accuracy in the `adv` condition than in the `flip` condition (Table Supp.1). When averaging across all trials, this effect was very significant for the pets and vegetables group ($p < 0.001$), and less significant for the hazard group ($p = 0.05$) (Figure 4b). These results suggest that the direction of the adversarial image perturbation, in combination with a

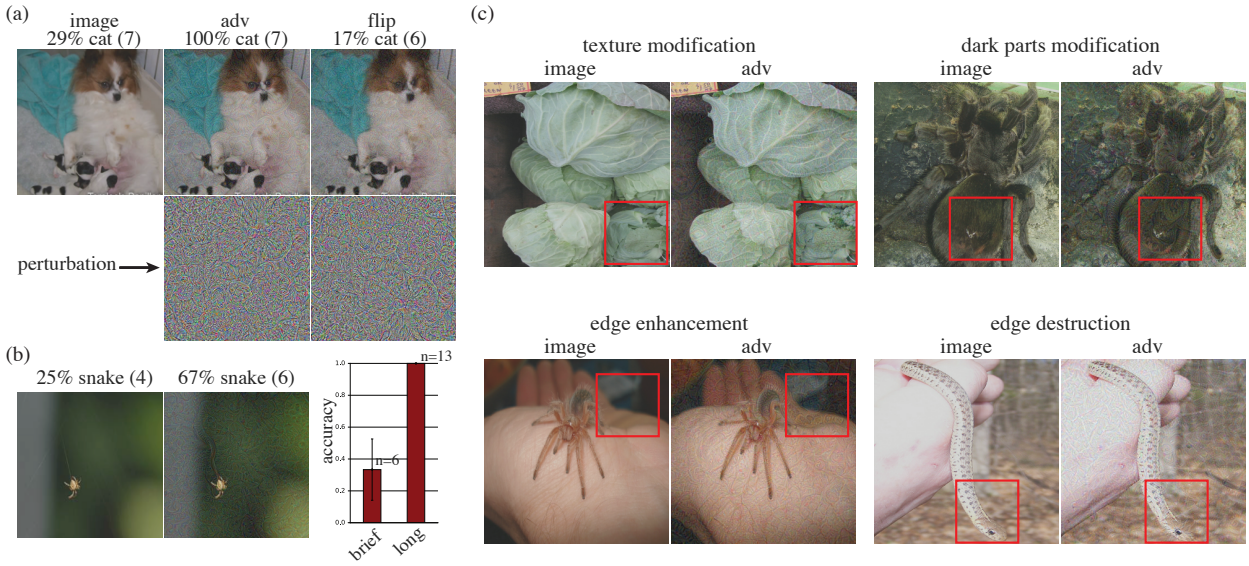

Figure 5. **Examples of adversarial images.** (a) A dog image that time-limited humans frequently perceived as a cat (top parentheses: number of subjects tested on this image). (b) Similar to a, a spider image is perceived as snake. right: accuracy on this adversarial image when presented briefly compared to when presented for long time (long presentation is based on a post-experiment email survey of 13 participants). (c) Examples of types of manipulations performed by the adversarial attack. See Figures Supp.4 through Supp.6 for additional examples of adversarial images. Also see Figure Supp.3 for adversarial examples from the `false` condition.

specific image, is perceptually relevant to features that the human visual system uses to classify objects. These findings thus give evidence that strong black box adversarial attacks can transfer from CNNs to humans, and show remarkable similarities between failure cases of CNNs and the human visual system.

The average response time was in all cases longer for the `adv` cases relative to the other conditions (Figure Supp.1b), though this result was only statistically significant for two comparisons. If this trend remains predictive, it would seem to contradict the case when we presented `false` images (Figure Supp.1a). One interpretation is that in the `false` case, the transfer of adversarial features to humans was accompanied by more confidence, whereas here the transfer was accompanied by less confidence, possibly due to competition between the adversarial and true class features in the `adv` condition.

5. Discussion & Conclusion

5.1. Commonly Observed Attack Attributes

Our adversarial examples are designed to fool human perception, so we should be careful using subjective human perception to understand how they work. With that caveat, we observed a few categories of recurring modification types, illustrated in Figure 5: disrupting object edges, especially by mid-frequency modulations perpendicular to the edge; enhancing edges both by increasing contrast and creating texture boundaries; modifying texture; and taking advantage of dark regions in the image, where the perceptual magnitude of small ϵ perturbations can be larger.

5.2. Short versus Long Exposure

The adversarial images in this paper transferred to humans when presented for extremely short exposure times, followed by a masking stimulus. In this configuration, there is little time for more than a single feedforward pass through the human visual pathway. Upon longer consideration, the true class of images remained obvious in most of the cases. A likely explanation for the greater resistance humans show upon longer consideration is that additional time for top-down and recurrent effects, and for the use of higher level cognitive mechanisms, improves classification accuracy and robustness (Di Lollo, 2012). This suggests that classification models, such as Sabour et al. (2017); Tang et al. (2017), that incorporate feedback and recurrent dynamics may prove more robust to adversarial examples in the same way as the human brain. Another possibility is that more brain-like models may instead lead to stronger adversarial examples that transfer to humans even after longer consideration.

5.3. Risks

The development of machine learning models which can generate fake images, audio, and video which appears realistic is already a source of acute concern (Farell & Perlstein, 2018). Adversarial examples provide one more way in which machine learning might plausibly be used to subtly manipulate humans. For instance, an ensemble of deep models might be trained on human ratings of face trustworthiness. It might then be possible to generate adversarial perturbations which enhance or reduce human impressions of trustworthiness, and those perturbed images might be used in news reports or political advertising.

More speculative risks involve the possibility of crafting sensory stimuli that hack the brain in a more diverse set of ways, and with larger effect (Stephenson, 1992). As one example, many animals have been observed to be susceptible to supernormal stimuli. For instance, cuckoo chicks generate begging calls and an associated visual display that causes birds of other species to prefer to feed the cuckoo chick over their own offspring (Hama, 2011). Adversarial examples can be seen as a form of supernormal stimuli for neural networks. A worrying possibility is that supernormal stimuli designed to influence human behavior or emotions, rather than merely the perceived class label of an image, might also transfer from machines to humans.

5.4. Positive Applications

If it is possible to generate adversarial examples with broader impact than in our work, then this would have promising as well as worrying applications. For instance, perhaps image perturbations could be designed to improve saliency, or attentiveness, when performing tasks like air traffic control or examination of radiology images, which are potentially tedious, but where the consequences of inattention are dire. User interface designers could use image perturbations to create more naturally intuitive designs.

5.5. Interpretation

Our study raises fundamental questions how adversarial examples work, how CNN models work, and how the brain works. Do adversarial attacks transfer from CNNs to humans because the semantic representation in a CNN is similar to that in the human brain? Do they instead transfer because both the representation in the CNN and the human brain are similar to some inherent semantic representation which naturally corresponds to reality? Some of these interpretation issues are clouded by the nature of the task we studied: visual object recognition, where it is difficult to define objectively correct answers. Is Figure 1 *objectively a dog* or is it objectively a cat but fools people into thinking it is a dog? Future work could clarify these interpretations by studying human performance on tasks with objectively correct answers. For example, one could study adversarial examples for both human and machine question answering systems applied to math questions. Future research that explores what properties of an adversarial example cause it to transfer to humans, and how those properties relate to properties of the physical world, will be of great interest, and perhaps can provide better understanding of both the brain and deep neural network models.

Acknowledgements

We are grateful to Ari Morcos, Bruno Olshausen, David Sussillo, Hanlin Tang, Santani Teng, and Daniel Yamins for useful discussions. We also thank Dan Abolafia Simon Kornblith, Katherine Lee, Niru Maheswaranathan, Catherine Olsson, David Sussillo, and Santani Teng, for helpful feedback on the manuscript. We thank Google Brain residents for useful feedback on the work. We also thank Deanna Chen, Leslie Philips, Sally Jesmonth, Phing Turner, Melissa Strader, Lily Peng, and Ricardo Prada for assistance with IRB and experiment setup.

References

Athalye, Anish. Robust adversarial examples, 2017. URL https://blog.openai.com/robust-adversarial-inputs.

Athalye, Anish and Sutskever, Ilya. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Biggio, Battista, Corona, Igino, Maiorca, Davide, Nelson, Blaine, Srndic, Nedim, Laskov, Pavel, Giacinto, Giorgio, and Roli, Fabio. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, pp. 387–402, 2013. doi: 10.1007/978-3-642-40994-3_25.

Brown, Tom B, Mané, Dandelion, Roy, Aurko, Abadi, Martín, and Gilmer, Justin. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Cadieu, Charles F, Hong, Ha, Yamins, Daniel LK, Pinto, Nicolas, Ardila, Diego, Solomon, Ethan A, Majaj, Najib J, and DiCarlo, James J. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12): e1003963, 2014.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical

image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

Di Lollo, Vincent. The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences*, 16(6): 317–321, 2012.

Eckstein, Miguel P, Koehler, Kathryn, Welbourne, Lauren E, and Akbas, Emre. Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27 (18):2827–2832, 2017.

Farell, Henry J. and Perlstein, Rick. Our hackable political future. https://www.nytimes.com/2018/02/04/opinion/hacking-politics-future.html, November 2018. Accessed: 2018-02-07.

Freeman, Jeremy and Simoncelli, Eero P. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195, 2011.

Gatys, Leon A, Ecker, Alexander S, and Bethge, Matthias. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Geirhos, Robert, Janssen, David HJ, Schütt, Heiko H, Rauber, Jonas, Bethge, Matthias, and Wichmann, Felix A. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

Goodfellow, Ian, Papernot, Nicolas, Huang, Sandy, Duan, Yan, Abbeel, Pieter, and Clark, Jack. Attacking machine learning with adversarial examples, 2017. URL https://blog.openai.com/adversarial-example-research/.

Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Grosse, Kathrin, Papernot, Nicolas, Manoharan, Praveen, Backes, Michael, and McDaniel, Patrick D. Adversarial examples for malware detection. In *ESORICS 2017*, pp. 62–79, 2017. doi: 10.1007/978-3-319-66399-9_4. URL https://doi.org/10.1007/978-3-319-66399-9_4.

Hama, Aldric. Supernormal stimuli: How primal urges overran their evolutionary purpose. *Mankind Quarterly*, 51(3):356, 2011.

Hassabis, Demis, Kumaran, Dharshan, Summerfield, Christopher, and Botvinick, Matthew. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Identity Mappings in Deep Residual Networks. *ArXiv e-prints*, March 2016.

Hillis, James M, Ernst, Marc O, Banks, Martin S, and Landy, Michael S. Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298 (5598):1627–1630, 2002.

Ibbotson, Michael and Krekelberg, Bart. Visual perception and saccadic eye movements. *Current opinion in neurobiology*, 21(4):553–558, 2011.

Jacob Buckman, Aurko Roy, Colin Raffel Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S18Su--CW. accepted as poster.

Kolter, J Zico and Wong, Eric. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

KovAcs, GYULA, Vogels, Rufin, and Orban, Guy A. Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences*, 92(12):5587–5591, 1995.

Kümmerer, Matthias, Theis, Lucas, and Bethge, Matthias. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

Kümmerer, Matthias, Wallis, Tom, and Bethge, Matthias. Deepgaze ii: Predicting fixations from deep features over time and tasks. *Journal of Vision*, 17(10):1147–1147, 2017.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial Machine Learning at Scale. *ArXiv e-prints*, November 2016.

Kurakin, Alexey, Goodfellow, Ian, and Bengio, Samy. Adversarial examples in the physical world. In *ICLR'2017 Workshop*, 2016. URL https://arxiv.org/abs/1607.02533.

Land, Michael F and Nilsson, Dan-Eric. *Animal eyes*. Oxford University Press, 2012.

Liu, Yanpei, Chen, Xinyun, Liu, Chang, and Song, Dawn. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Madry, Aleksander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, and Vladu, Adrian. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

McIntosh, Lane, Maheswaranathan, Niru, Nayebi, Aran, Ganguli, Surya, and Baccus, Stephen. Deep learning models of the retinal response to natural scenes. In *Advances*

in neural information processing systems, pp. 1369–1377, 2016.

Olshausen, Bruno A. 20 years of learning about vision: Questions answered, questions unanswered, and questions not yet asked. In *20 Years of Computational Neuroscience*, pp. 243–270. Springer, 2013.

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

Papernot, Nicolas, McDaniel, Patrick, and Goodfellow, Ian. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

Papernot, Nicolas, McDaniel, Patrick, Swami, Ananthram, and Harang, Richard. Crafting adversarial input sequences for recurrent neural networks. In *Military Communications Conference, MILCOM 2016-2016 IEEE*, pp. 49–54. IEEE, 2016b.

Papernot, Nicolas, McDaniel, Patrick, Wu, Xi, Jha, Somesh, and Swami, Ananthram. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pp. 582–597. IEEE, 2016c.

Papernot, Nicolas, McDaniel, Patrick, Goodfellow, Ian, Jha, Somesh, Celik, Z Berkay, and Swami, Ananthram. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Potter, Mary C, Wyble, Brad, Hagmann, Carl Erick, and McCourt, Emily S. Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76 (2):270–279, 2014.

Rajalingham, Rishi, Issa, Elias B., Bashivan, Pouya, Kar, Kohitij, Schmidt, Kailyn, and DiCarlo, James J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 2018. doi: 10.1101/240614. URL https://www.biorxiv.org/content/early/2018/01/23/240614.1.

Riesenhuber, Maximilian and Poggio, Tomaso. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999.

Sabour, Sara, Frosst, Nicholas, and Hinton, Geoffrey E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pp. 3859–3869, 2017.

Stephenson, N. *Snow Crash*. Bantam Books, 1992.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *ArXiv e-prints*, December 2015.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *ArXiv e-prints*, February 2016.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, and Fergus, Rob. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tang, Hanlin, Lotter, Bill, Schrimpf, Martin, Paredes, Ana, Caro, Josue Ortega, Hardesty, Walter, Cox, David, and Kreiman, Gabriel. Recurrent computations for visual pattern completion. *arXiv preprint arXiv:1706.02240*, 2017.

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., and McDaniel, P. Ensemble Adversarial Training: Attacks and Defenses. *ArXiv e-prints*, May 2017.

Van Essen, D. C. and Anderson, C. H. Information processing strategies and pathways in the primate visual system. In F., Zornetzer S., L., Davis J., C., Lau, and T., McKenna (eds.), *An introduction to neural and electronic networks*, pp. 4576, San Diego, CA, 1995. Academic Press.

Xu, Weilin, Evans, David, and Qi, Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Yamins, Daniel L. K. and DiCarlo, James J. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 2016.

Supplemental material

A. Supplementary Figures and Tables

Table Supp.1. **Adversarial examples transfer to humans.** Number of subjects that reported the correct class of images in the `adv` condition with lower mean accuracy compared to their mean accuracy in the `image` and `flip` conditions.

Group	`adv` < `image`	`adv` < `flip`	**total**
pets	29	22	35
hazard	19	16	24
vegetables	21	23	32

Table Supp.2. **Accuracy of models on ImageNet validation set.** * models trained on ImageNet with retina layer pre-pended and with train data augmented with rescaled images in the range of $[40, 255-40]$; ** model trained with adversarial examples augmented data. First ten models are models used in the adversarial training ensemble. Last two models are models used to test the transferability of adversarial examples.

Model	**Top-1 accuracy**
Resnet V2 101	0.77
Resnet V2 101*	0.7205
Inception V4	0.802
Inception V4*	0.7518
Inception Resnet V2	0.804
Inception Resnet V2*	0.7662
Inception V3	0.78
Inception V3*	0.7448
Resnet V2 152	0.778
Resnet V2 50*	0.708
Resnet V2 50 (test)	0.756
Inception V3** (test)	0.776

Table Supp.3. **Accuracy of ensemble used to generate adversarial examples on images at different conditions.** * models trained on ImageNet with retina layer appended and with train data augmented with rescaled images in the range of $[40, 255-40]$; Numbers triplet reflects accuracy on images from pets, hazard, and vegetables groups, respectively.

Train Model	`adv` (%)	`flip` (%)
Resnet V2 101	0.0, 0.0, 0.0	95, 92, 91
Resnet V2 101*	0.0, 0.0, 0.0	87, 87, 77
Inception V4	0.0, 0.0, 0.0	96, 95, 86
Inception V4*	0.0, 0.0, 0.0	87, 87, 73
Inception Resnet V2	0.0, 0.0, 0.0	97, 95, 95
Inception Resnet V2*	0.0, 0.0, 0.0	87, 83, 73
Inception V3	0.0, 0.0, 0.0	97, 94, 89
Inception V3*	0.0, 0.0, 0.0	83, 86, 74
Resnet V2 152	0.0, 0.0, 0.0	96, 95, 91
Resnet V2 50*	0.0, 0.0, 0.0	82, 85, 81

Table Supp.4. **Accuracy of test models on images at different conditions.** ** model trained on both clean and adversarial images. Numbers triplet is accuracy on pets, hazard, and vegetables groups, respectively.

Model	`adv` (%)	`flip` (%)
Resnet V2 50	8.7, 9.4, 13	93, 91, 85
Inception V3**	6.0, 6.9, 17	95, 92, 94

Table Supp.5. **Attack success on model ensemble.** Same convention as Table Supp.3

Model	`adv` (%)	`flip` (%)
Resnet V2 101	100, 100, 100	2, 0, 0
Resnet V2 101*	100, 100, 100	3, 0, 0
Inception V4	100, 100, 100	1, 0, 1
Inception V4*	100, 100, 100	4, 1, 0
Inception Resnet V2	100, 100, 100	1, 0, 1
Inception Resnet V2*	100, 100, 100	5, 2, 0
Inception V3	100, 100, 100	1, 0, 0
Inception V3*	100, 100, 100	5, 1, 1
Resnet V2 152	100, 100, 100	1, 0, 0
Resnet V2 50*	100, 100, 100	3, 1, 0

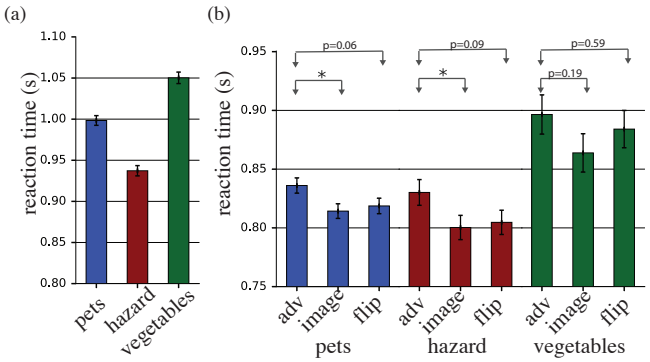

Figure Supp.1. **Reaction time.** (a) subjects average response time to `false` images. (b) response time of subjects for the `adv`, `image`, and `flip` conditions (error bars $\pm$ SE; * reflects $p < 0.05$; two sample two-tailed t-test).

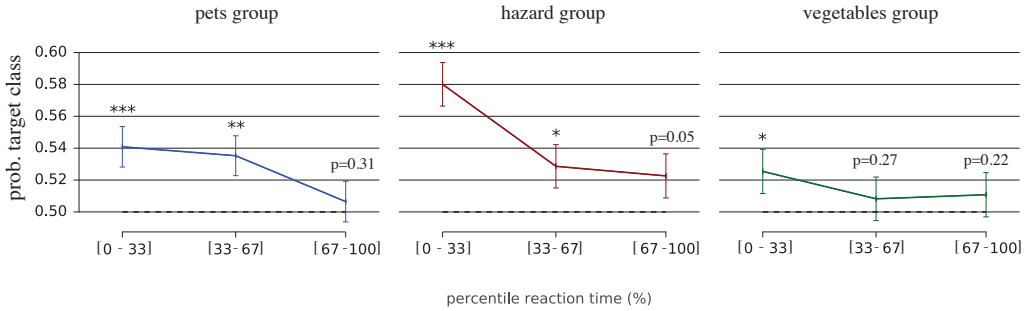

Figure Supp.2. **Adversarial Perturbation Bias of Human perception when Confident** probability of choosing the adversarially targeted class, when the true image class is not one of the choices that subjects can report, estimated by averaging the responses of all subjects (two-tailed t-test relative to chance level 0.5; error bars $\pm$ SE; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$). The probability of choosing the targeted label is computed by binning trials within percentile reaction time ranges (0-33 percentile, 33-67 percentile, and 67-100 percentile). The bias relative to chance level of 0.5 is significant when people reported their decision quickly (i.e., when they are more confident), but not significant when they reported their decision more slowly.

Image removed due to file size constraints. See http://goo.gl/SJ8jpq for full Supplemental Material with all images.

Figure Supp.3. **Adversarial Examples for false condition** (a) pets group. (b) hazard group. (c) vegetables group.

Image removed due to file size constraints. See http://goo.gl/SJ8jpq for full Supplemental Material with all images.

Figure Supp.4. **Adversarial Examples** pets group

Image removed due to file size constraints. See http://goo.gl/SJ8jpq for full Supplemental Material with all images.

Figure Supp.5. **Adversarial Examples** hazard group

Image removed due to file size constraints. See http://goo.gl/SJ8jpq for full Supplemental Material with all images.

Figure Supp.6. **Adversarial Examples** vegetables group

B. Details of retinal blurring layer

B.1. Computing the primate eccentricity map

Let d_{viewer} be the distance (in meters) of the viewer from the display and d_{hw} be the height and width of a square image (in meters). For every spatial position (in meters) $c = (x, y) \in R^2$ in the image we compute the retinal eccentricity (in radians) as follows:

$$\theta(c) = \tan^{-1}\left(\frac{||c||_2}{d_{viewer}}\right) \tag{4}$$

and turn this into a target resolution in units of radians

$$r_{rad}(c) = \min(\alpha\theta(c), \beta). \tag{5}$$

We then turn this target resolution into a target spatial resolution in the plane of the screen,

$$r_m(c) = r_{rad}(c)\left(1 + \tan^2(\theta(c))\right), \tag{6}$$

$$r_{pixel}(c) = r_m(c) \cdot [\text{pixels per meter}]. \tag{7}$$

This spatial resolution for two point discrimination is then converted into a corresponding low-pass cutoff frequency, in units of cycles per pixel,

$$f(c) = \frac{\pi}{r_{pixel}}, \tag{8}$$

where the numerator is π rather than 2π since the two point discrimination distance r_{pixel} is half the wavelength.

Finally, this target low-pass spatial frequency $f(c)$ for each pixel is used to linearly interpolate each pixel value from the corresponding pixel in a set of low pass filtered images, as described in the following algorithm (all operations on matrices are assumed to be performed elementwise), We additionally cropped $X_{retinal}$ to 90% width before use, to remove artifacts

Algorithm 1 Applying retinal blur to an image

$X_{img} \leftarrow$ input image
$F \leftarrow$ image containing corresponding target lowpass frequency for each pixel, computed from $f(c)$
$\tilde{X} \leftarrow \text{FFT}(X_{img})$
$G \leftarrow$ norm of spatial frequency at each position in Y
CUTOFF_FREQS $\leftarrow$ list of frequencies to use as cutoffs for low-pass filtering
for f' in CUTOFF_FREQS **do**
    $\tilde{Y}_{f'} \leftarrow \tilde{X} \odot \exp\left(-\frac{G^2}{f^2}\right)$
    $Y_f \leftarrow \text{InverseFFT}(\tilde{Y}_{f'})$
end for
$w(c) \leftarrow$ linear interpolation coefficients for $F(c)$ into CUTOFF_FREQS $\quad \forall c$
$X_{retinal}(c) \leftarrow \sum_{f'} w_{f'}(c) Y_{f'}(c) \quad \forall c$

from the image edge.

Note that because the per-pixel blurring is performed using linear interpolation into images that were low-pass filtered in Fourier space, this transformation is both fast to compute and fully differentiable.

C. Calculating probability of coarse class

To calculate the probability a model assigns to a coarse class, we summed probabilities assigned to the individual classes within the coarse class. Let S_{target} be the set of all individual labels in the target coarse class. Let S_{other} be all other individual labels not in the target coarse class. $|S_{\text{target}}| + |S_{\text{other}}| = 1000$, since there are 1000 labels in ImageNet. Let Y be the coarse class variable and y_{target} be our target coarse class. We can compute the probability a model k assigns to a

coarse class given image X as

$$P_k(Y = y_{\text{target}}|X) = \sigma\left(\log \frac{\sum_{i \in S_{\text{target}}} \tilde{F}_k\left(i|X\right)}{\sum_{i \in S_{\text{other}}} \tilde{F}_k\left(i|X\right)}\right) \tag{9}$$

where $\tilde{F}_k\left(i|X\right)$ is the unnormalized probability assigned to fine class i.

D. Convolutional Neural Network Models

Some of the models in our ensemble are from a publicly available pretrained checkpoints[2], and others are our own instances of the architectures, specifically trained for this experiment on ImageNet with the retinal layer prepended. To encourage invariance to image intensity scaling, we augmented each training batch with another batch with the same images but rescaled in the range of $[40, 255 - 40]$, instead of $[0, 255]$. Supplementary Table Supp.2 identifies all ten models used in the ensemble, and shows their top-1 accuracies, along with two holdout models that we used for evaluation.

E. Image List from Imagenet

The specific imagenet images used from each class in the experiments in this paper are as follows:

dog:

'n02106382_564.JPEG', 'n02110958_598.JPEG', 'n02101556_13462.JPEG', 'n02113624_7358.JPEG',
'n02113799_2538.JPEG', 'n02091635_11576.JPEG', 'n02106382_2781.JPEG', 'n02112706_105.JPEG',
'n02095570_10951.JPEG', 'n02093859_5274.JPEG', 'n02109525_10825.JPEG', 'n02096294_1400.JPEG',
'n02086646_241.JPEG', 'n02098286_5642.JPEG', 'n02106382_9015.JPEG', 'n02090379_9754.JPEG',
'n02102318_10390.JPEG', 'n02086646_4202.JPEG', 'n02086910_5053.JPEG', 'n02113978_3051.JPEG',
'n02093859_3809.JPEG', 'n02105251_2485.JPEG', 'n02109525_35418.JPEG', 'n02108915_7834.JPEG',
'n02113624_430.JPEG', 'n02093256_7467.JPEG', 'n02087046_2701.JPEG', 'n02090379_8849.JPEG',
'n02093754_717.JPEG', 'n02086079_15905.JPEG', 'n02102480_4466.JPEG', 'n02107683_5333.JPEG',
'n02102318_8228.JPEG', 'n02099712_867.JPEG', 'n02094258_1958.JPEG', 'n02109047_25075.JPEG',
'n02113624_4304.JPEG', 'n02097474_10985.JPEG', 'n02091032_3832.JPEG', 'n02085620_859.JPEG',
'n02110806_582.JPEG', 'n02085782_8327.JPEG', 'n02094258_5318.JPEG', 'n02087046_5721.JPEG',
'n02095570_746.JPEG', 'n02099601_3771.JPEG', 'n02102480_41.JPEG', 'n02086910_1048.JPEG',
'n02094114_7299.JPEG', 'n02108551_13160.JPEG', 'n02110185_9847.JPEG', 'n02097298_13025.JPEG',
'n02097298_16751.JPEG', 'n02091467_555.JPEG', 'n02113799_2504.JPEG', 'n02085782_14116.JPEG',
'n02097474_13885.JPEG', 'n02105251_8108.JPEG', 'n02113799_3415.JPEG', 'n02095570_8170.JPEG',
'n02088238_1543.JPEG', 'n02097047_6.JPEG', 'n02104029_5268.JPEG', 'n02100583_11473.JPEG',
'n02113978_6888.JPEG', 'n02104365_1737.JPEG', 'n02096177_4779.JPEG', 'n02107683_5303.JPEG',
'n02108915_11155.JPEG', 'n02086910_1872.JPEG', 'n02106550_8383.JPEG', 'n02088094_2191.JPEG',
'n02085620_11897.JPEG', 'n02096051_4802.JPEG', 'n02100735_3641.JPEG', 'n02091032_1389.JPEG',
'n02106382_4671.JPEG', 'n02097298_9059.JPEG', 'n02107312_280.JPEG', 'n02111889_86.JPEG',
'n02113978_5397.JPEG', 'n02097209_3461.JPEG', 'n02089867_1115.JPEG', 'n02097658_4987.JPEG',
'n02094114_4125.JPEG', 'n02100583_130.JPEG', 'n02112137_5859.JPEG', 'n02113799_19636.JPEG',
'n02088094_5488.JPEG', 'n02089078_393.JPEG', 'n02098413_1794.JPEG', 'n02113799_1970.JPEG',
'n02091032_3655.JPEG', 'n02105855_11127.JPEG', 'n02096294_3025.JPEG', 'n02094114_4831.JPEG',
'n02111889_10472.JPEG', 'n02113624_9125.JPEG', 'n02097474_9719.JPEG', 'n02094433_2451.JPEG',
'n02095889_6464.JPEG', 'n02093256_458.JPEG', 'n02091134_2732.JPEG', 'n02091244_2622.JPEG',
'n02094114_2169.JPEG', 'n02090622_2337.JPEG', 'n02101556_6764.JPEG', 'n02096051_1459.JPEG',
'n02087046_9056.JPEG', 'n02098105_8405.JPEG', 'n02112137_5696.JPEG', 'n02110806_7949.JPEG',
'n02097298_2420.JPEG', 'n02085620_6814.JPEG', 'n02108915_1703.JPEG', 'n02100877_19273.JPEG',
'n02106550_3765.JPEG', 'n02107312_3524.JPEG', 'n02111889_2963.JPEG', 'n02113624_9129.JPEG',
'n02097047_3200.JPEG', 'n02093256_8365.JPEG', 'n02093991_9420.JPEG', 'n02112137_1635.JPEG',
'n02111129_3530.JPEG', 'n02101006_8123.JPEG', 'n02102040_5033.JPEG', 'n02113624_437.JPEG',
'n02090622_5866.JPEG', 'n02110806_3711.JPEG', 'n02112137_14788.JPEG', 'n02105162_7406.JPEG',

[2]https://github.com/tensorflow/models/tree/master/research/slim

'n02097047_5061.JPEG', 'n02108422_11587.JPEG', 'n02091467_4265.JPEG', 'n02091467_12683.JPEG',
'n02104365_3628.JPEG', 'n02086646_3314.JPEG', 'n02099849_736.JPEG', 'n02100735_8112.JPEG',
'n02112018_12764.JPEG', 'n02093428_11175.JPEG', 'n02110627_9822.JPEG', 'n02107142_24318.JPEG',
'n02105162_5489.JPEG', 'n02093754_5904.JPEG', 'n02110958_215.JPEG', 'n02095314_4027.JPEG',
'n02109961_3250.JPEG', 'n02108551_7343.JPEG', 'n02110627_10272.JPEG', 'n02088364_3099.JPEG',
'n02110806_2721.JPEG', 'n02095314_2261.JPEG', 'n02106550_9870.JPEG', 'n02107574_3991.JPEG',
'n02095570_3288.JPEG', 'n02086079_39042.JPEG', 'n02096294_9416.JPEG', 'n02110806_6528.JPEG',
'n02088466_11397.JPEG', 'n02092002_996.JPEG', 'n02098413_8605.JPEG', 'n02085620_712.JPEG',
'n02100236_3011.JPEG', 'n02086646_7788.JPEG', 'n02085620_4661.JPEG', 'n02098105_1746.JPEG',
'n02113624_8608.JPEG', 'n02097474_1168.JPEG', 'n02107683_1496.JPEG', 'n02110185_12849.JPEG',
'n02085620_11946.JPEG', 'n02087394_16385.JPEG', 'n02110806_22671.JPEG', 'n02113624_526.JPEG',
'n02096294_12642.JPEG', 'n02113023_7510.JPEG', 'n02088364_13285.JPEG', 'n02095889_2977.JPEG',
'n02105056_9215.JPEG', 'n02102318_9744.JPEG', 'n02097298_11834.JPEG', 'n02111277_16201.JPEG',
'n02085782_8518.JPEG', 'n02113978_11280.JPEG', 'n02106382_10700.JPEG'.

cat:

'n02123394_661.JPEG', 'n02123045_11954.JPEG', 'n02123394_3695.JPEG', 'n02123394_2692.JPEG',
'n02123597_12166.JPEG', 'n02123045_7014.JPEG', 'n02123159_2777.JPEG', 'n02123394_684.JPEG',
'n02124075_543.JPEG', 'n02123597_7557.JPEG', 'n02124075_7857.JPEG', 'n02123597_3770.JPEG',
'n02124075_4986.JPEG', 'n02123045_568.JPEG', 'n02123394_1541.JPEG', 'n02123597_3498.JPEG',
'n02123597_10304.JPEG', 'n02123394_2084.JPEG', 'n02123597_5283.JPEG', 'n02123597_13807.JPEG',
'n02124075_12282.JPEG', 'n02123597_8575.JPEG', 'n02123045_11787.JPEG', 'n02123394_888.JPEG',
'n02123045_1815.JPEG', 'n02123394_7614.JPEG', 'n02123597_27865.JPEG', 'n02124075_1279.JPEG',
'n02123394_4775.JPEG', 'n02123394_976.JPEG', 'n02123394_8385.JPEG', 'n02123597_14791.JPEG',
'n02123045_10424.JPEG', 'n02123597_7698.JPEG', 'n02124075_8140.JPEG', 'n02123045_3754.JPEG',
'n02123597_1819.JPEG', 'n02123597_395.JPEG', 'n02123394_415.JPEG', 'n02124075_9747.JPEG',
'n02123045_9467.JPEG', 'n02123159_6842.JPEG', 'n02123394_9611.JPEG', 'n02123597_7283.JPEG',
'n02123597_11799.JPEG', 'n02123597_660.JPEG', 'n02123045_7511.JPEG', 'n02123597_10723.JPEG',
'n02123159_7836.JPEG', 'n02123597_14530.JPEG', 'n02123597_28555.JPEG', 'n02123394_6079.JPEG',
'n02123394_6792.JPEG', 'n02123597_11564.JPEG', 'n02123597_8916.JPEG', 'n02124075_123.JPEG',
'n02123045_5150.JPEG', 'n02124075_353.JPEG', 'n02123597_12941.JPEG', 'n02123045_10095.JPEG',
'n02123597_6533.JPEG', 'n02123045_4611.JPEG', 'n02123597_754.JPEG', 'n02123394_8561.JPEG',
'n02123597_6409.JPEG', 'n02123159_4909.JPEG', 'n02123597_564.JPEG', 'n02123394_1633.JPEG',
'n02123394_1196.JPEG', 'n02123394_2787.JPEG', 'n02124075_10542.JPEG', 'n02123597_6242.JPEG',
'n02123597_3063.JPEG', 'n02123597_13164.JPEG', 'n02123045_7449.JPEG', 'n02123045_13299.JPEG',
'n02123394_8165.JPEG', 'n02123394_1852.JPEG', 'n02123597_8771.JPEG', 'n02123159_6581.JPEG',
'n02123394_5906.JPEG', 'n02124075_2747.JPEG', 'n02124075_11383.JPEG', 'n02123597_3919.JPEG',
'n02123394_2514.JPEG', 'n02124075_7423.JPEG', 'n02123394_6968.JPEG', 'n02123045_4850.JPEG',
'n02123045_10689.JPEG', 'n02124075_13539.JPEG', 'n02123597_13378.JPEG', 'n02123159_4847.JPEG',
'n02123394_1798.JPEG', 'n02123597_27951.JPEG', 'n02123159_587.JPEG', 'n02123597_1825.JPEG',
'n02123159_2200.JPEG', 'n02123597_12.JPEG', 'n02123597_6778.JPEG', 'n02123597_6693.JPEG',
'n02123045_11782.JPEG', 'n02123597_13706.JPEG', 'n02123394_9032.JPEG', 'n02124075_4459.JPEG',
'n02123597_13752.JPEG', 'n02123394_2285.JPEG', 'n02123597_1410.JPEG', 'n02123159_6134.JPEG',
'n02123597_11290.JPEG', 'n02123597_6347.JPEG', 'n02123394_1789.JPEG', 'n02123045_11255.JPEG',
'n02123394_6096.JPEG', 'n02123394_4081.JPEG', 'n02123394_5679.JPEG', 'n02123394_2471.JPEG',
'n02123159_5797.JPEG', 'n02123597_13894.JPEG', 'n02124075_10854.JPEG', 'n02123394_8605.JPEG',
'n02124075_8281.JPEG', 'n02123597_11724.JPEG', 'n02123394_8242.JPEG', 'n02123394_3569.JPEG',
'n02123597_10639.JPEG', 'n02123045_3818.JPEG', 'n02124075_6459.JPEG', 'n02123394_185.JPEG',
'n02123597_8961.JPEG', 'n02124075_9743.JPEG', 'n02123394_1627.JPEG', 'n02123597_13175.JPEG',
'n02123045_2694.JPEG', 'n02123597_4537.JPEG', 'n02123597_6400.JPEG', 'n02123045_7423.JPEG',
'n02123597_3004.JPEG', 'n02123394_2988.JPEG', 'n02124075_9512.JPEG', 'n02123394_6318.JPEG',
'n02123597_1843.JPEG', 'n02124075_2053.JPEG', 'n02123597_3828.JPEG', 'n02123394_14.JPEG',
'n02123394_8141.JPEG', 'n02124075_1624.JPEG', 'n02123597_459.JPEG', 'n02124075_6405.JPEG',
'n02123045_8595.JPEG', 'n02123159_3226.JPEG', 'n02124075_9141.JPEG', 'n02123597_2031.JPEG',

'n02123045_2354.JPEG', 'n02123597_6710.JPEG', 'n02123597_6613.JPEG', 'n02123159_1895.JPEG',
'n02123394_2953.JPEG', 'n02123394_5846.JPEG', 'n02123394_513.JPEG', 'n02123045_16637.JPEG',
'n02123394_7848.JPEG', 'n02123394_3229.JPEG', 'n02123045_8881.JPEG', 'n02123394_8250.JPEG',
'n02124075_7651.JPEG', 'n02123394_200.JPEG', 'n02123394_2814.JPEG', 'n02123045_6445.JPEG',
'n02123394_2467.JPEG', 'n02123045_3317.JPEG', 'n02123597_1422.JPEG', 'n02123597_13442.JPEG',
'n02123394_8225.JPEG', 'n02123597_9337.JPEG', 'n02123394_32.JPEG', 'n02123394_2193.JPEG',
'n02123394_1625.JPEG', 'n02123597_8799.JPEG', 'n02123597_13241.JPEG', 'n02123597_7681.JPEG',
'n02123597_4550.JPEG', 'n02123597_3896.JPEG', 'n02123394_9554.JPEG', 'n02124075_13600.JPEG',
'n02123394_571.JPEG', 'n02123597_10886.JPEG', 'n02123045_6741.JPEG', 'n02123045_10438.JPEG',
'n02123045_9954.JPEG'.

spider:

'n01775062_517.JPEG', 'n01774750_18017.JPEG', 'n01774384_13186.JPEG', 'n01774750_3115.JPEG',
'n01775062_5075.JPEG', 'n01773549_1541.JPEG', 'n01775062_4867.JPEG', 'n01775062_8156.JPEG',
'n01774750_7128.JPEG', 'n01775062_4632.JPEG', 'n01773549_8734.JPEG', 'n01773549_2274.JPEG',
'n01773549_10298.JPEG', 'n01774384_1811.JPEG', 'n01774750_7498.JPEG', 'n01774750_10265.JPEG',
'n01773549_1964.JPEG', 'n01774750_3268.JPEG', 'n01773549_6095.JPEG', 'n01775062_8812.JPEG',
'n01774750_10919.JPEG', 'n01775062_1180.JPEG', 'n01773549_7275.JPEG', 'n01773549_9346.JPEG',
'n01773549_8243.JPEG', 'n01775062_3127.JPEG', 'n01773549_10608.JPEG', 'n01773549_3442.JPEG',
'n01773157_1487.JPEG', 'n01774750_7775.JPEG', 'n01775062_419.JPEG', 'n01774750_7638.JPEG',
'n01775062_847.JPEG', 'n01774750_3154.JPEG', 'n01773549_1534.JPEG', 'n01773157_1039.JPEG',
'n01775062_5644.JPEG', 'n01775062_8525.JPEG', 'n01773797_216.JPEG', 'n01775062_900.JPEG',
'n01774750_8513.JPEG', 'n01774750_3424.JPEG', 'n01774750_3085.JPEG', 'n01775062_3662.JPEG',
'n01774384_15681.JPEG', 'n01774750_326.JPEG', 'n01773157_9503.JPEG', 'n01774750_3332.JPEG',
'n01774750_2799.JPEG', 'n01773157_10606.JPEG', 'n01773157_1905.JPEG', 'n01773549_379.JPEG',
'n01773797_597.JPEG', 'n01773157_3226.JPEG', 'n01774750_7875.JPEG', 'n01774384_16102.JPEG',
'n01773549_2832.JPEG', 'n01775062_5072.JPEG', 'n01773549_4278.JPEG', 'n01773549_5854.JPEG',
'n01774384_1998.JPEG', 'n01774750_13875.JPEG', 'n01775062_8270.JPEG', 'n01773549_2941.JPEG',
'n01774750_5235.JPEG', 'n01773549_4150.JPEG', 'n01774750_6217.JPEG', 'n01775062_3137.JPEG',
'n01774750_5480.JPEG', 'n01774384_11955.JPEG', 'n01775062_8376.JPEG', 'n01773157_2688.JPEG',
'n01773549_6825.JPEG', 'n01774750_10422.JPEG', 'n01774384_20786.JPEG', 'n01773549_398.JPEG',
'n01773549_4965.JPEG', 'n01774750_7470.JPEG', 'n01775062_1379.JPEG', 'n01774384_2399.JPEG',
'n01773549_9799.JPEG', 'n01775062_305.JPEG', 'n01774384_15519.JPEG', 'n01774750_3333.JPEG',
'n01774750_2604.JPEG', 'n01774750_3134.JPEG', 'n01774750_4646.JPEG', 'n01775062_5009.JPEG',
'n01774750_10200.JPEG', 'n01775062_7964.JPEG', 'n01774384_2458.JPEG', 'n01773797_3333.JPEG',
'n01774750_9987.JPEG', 'n01773549_5790.JPEG', 'n01773549_854.JPEG', 'n01774750_11370.JPEG',
'n01774750_10698.JPEG', 'n01774750_9287.JPEG', 'n01773797_6703.JPEG', 'n01773797_931.JPEG',
'n01773549_5280.JPEG', 'n01773797_5385.JPEG', 'n01773797_1098.JPEG', 'n01774750_436.JPEG',
'n01774384_13770.JPEG', 'n01774750_9780.JPEG', 'n01774750_8640.JPEG', 'n01774750_653.JPEG',
'n01774384_12554.JPEG', 'n01774750_9716.JPEG'

snake:

'n01737021_7081.JPEG', 'n01728572_16119.JPEG', 'n01735189_10620.JPEG', 'n01751748_3573.JPEG',
'n01729322_6690.JPEG', 'n01735189_20703.JPEG', 'n01734418_4792.JPEG', 'n01749939_2784.JPEG',
'n01729977_4113.JPEG', 'n01756291_6505.JPEG', 'n01742172_3003.JPEG', 'n01728572_19317.JPEG',
'n01739381_5838.JPEG', 'n01737021_1381.JPEG', 'n01749939_4704.JPEG', 'n01755581_10792.JPEG',
'n01729977_9474.JPEG', 'n01744401_11909.JPEG', 'n01739381_10303.JPEG', 'n01749939_820.JPEG',
'n01728572_27743.JPEG', 'n01734418_12057.JPEG', 'n01742172_8636.JPEG', 'n01729977_14112.JPEG',
'n01739381_6286.JPEG', 'n01734418_761.JPEG', 'n01740131_13437.JPEG', 'n01728920_9571.JPEG',
'n01753488_4234.JPEG', 'n01749939_5712.JPEG', 'n01739381_6072.JPEG', 'n01739381_7683.JPEG',
'n01729322_9202.JPEG', 'n01751748_13413.JPEG', 'n01756291_4626.JPEG', 'n01742172_9733.JPEG',
'n01737021_12610.JPEG', 'n01739381_87.JPEG', 'n01729977_1134.JPEG', 'n01753488_637.JPEG',
'n01748264_18478.JPEG', 'n01728572_22360.JPEG', 'n01737021_3386.JPEG', 'n01751748_560.JPEG',
'n01751748_18223.JPEG', 'n01749939_5750.JPEG', 'n01748264_7044.JPEG', 'n01739381_1163.JPEG',

'n01751748_311.JPEG', 'n01756291_9028.JPEG', 'n01739381_10473.JPEG', 'n01728572_1415.JPEG', 'n01729322_10918.JPEG', 'n01748264_653.JPEG', 'n01753488_10957.JPEG', 'n01756291_3990.JPEG', 'n01756291_11915.JPEG', 'n01756291_6776.JPEG', 'n01740131_11661.JPEG', 'n01729977_5715.JPEG', 'n01737021_16733.JPEG', 'n01753488_15197.JPEG', 'n01744401_7248.JPEG', 'n01728572_7661.JPEG', 'n01740131_13680.JPEG', 'n01729322_5446.JPEG', 'n01749939_6508.JPEG', 'n01748264_2140.JPEG', 'n01729977_16782.JPEG', 'n01748264_7602.JPEG', 'n01756291_17857.JPEG', 'n01729977_461.JPEG', 'n01742172_20552.JPEG', 'n01735189_3258.JPEG', 'n01728920_9265.JPEG', 'n01748264_18133.JPEG', 'n01748264_16699.JPEG', 'n01739381_1006.JPEG', 'n01753488_10555.JPEG', 'n01751748_3202.JPEG', 'n01734418_3929.JPEG', 'n01751748_5908.JPEG', 'n01751748_8470.JPEG', 'n01739381_3598.JPEG', 'n01739381_255.JPEG', 'n01729977_15657.JPEG', 'n01748264_21477.JPEG', 'n01751748_2912.JPEG', 'n01728920_9154.JPEG', 'n01728572_17552.JPEG', 'n01740131_14560.JPEG', 'n01729322_5947.JPEG'.

Broccoli:

'n07714990_8640.JPEG', 'n07714990_5643.JPEG', 'n07714990_7777.JPEG', 'n07714990_888.JPEG', 'n07714990_3398.JPEG', 'n07714990_4576.JPEG', 'n07714990_8554.JPEG', 'n07714990_1957.JPEG', 'n07714990_4201.JPEG', 'n07714990_3130.JPEG', 'n07714990_4115.JPEG', 'n07714990_524.JPEG', 'n07714990_6504.JPEG', 'n07714990_3125.JPEG', 'n07714990_5838.JPEG', 'n07714990_1779.JPEG', 'n07714990_6393.JPEG', 'n07714990_1409.JPEG', 'n07714990_4962.JPEG', 'n07714990_7282.JPEG', 'n07714990_7314.JPEG', 'n07714990_11933.JPEG', 'n07714990_1202.JPEG', 'n07714990_3626.JPEG', 'n07714990_7873.JPEG', 'n07714990_3325.JPEG', 'n07714990_3635.JPEG', 'n07714990_12524.JPEG', 'n07714990_14952.JPEG', 'n07714990_7048.JPEG', 'n07714990_500.JPEG', 'n07714990_7950.JPEG', 'n07714990_2445.JPEG', 'n07714990_1294.JPEG', 'n07714990_7336.JPEG', 'n07714990_14743.JPEG', 'n07714990_1423.JPEG', 'n07714990_2185.JPEG', 'n07714990_6566.JPEG', 'n07714990_567.JPEG', 'n07714990_1532.JPEG', 'n07714990_5212.JPEG', 'n07714990_8971.JPEG', 'n07714990_6116.JPEG', 'n07714990_5462.JPEG', 'n07714990_7644.JPEG', 'n07714990_8596.JPEG', 'n07714990_1138.JPEG', 'n07714990_15078.JPEG', 'n07714990_1602.JPEG', 'n07714990_2460.JPEG', 'n07714990_159.JPEG', 'n07714990_9445.JPEG', 'n07714990_471.JPEG', 'n07714990_1777.JPEG', 'n07714990_9760.JPEG', 'n07714990_1528.JPEG', 'n07714990_12338.JPEG', 'n07714990_2201.JPEG', 'n07714990_6850.JPEG', 'n07714990_4492.JPEG', 'n07714990_7791.JPEG', 'n07714990_9752.JPEG', 'n07714990_1702.JPEG', 'n07714990_3682.JPEG', 'n07714990_14342.JPEG', 'n07714990_2661.JPEG', 'n07714990_5467.JPEG'.

Cabbage:

'n07714571_14784.JPEG', 'n07714571_4795.JPEG', 'n07714571_11969.JPEG', 'n07714571_1394.JPEG', 'n07714571_4155.JPEG', 'n07714571_3624.JPEG', 'n07714571_13753.JPEG', 'n07714571_7351.JPEG', 'n07714571_10316.JPEG', 'n07714571_7235.JPEG', 'n07714571_17716.JPEG', 'n07714571_1639.JPEG', 'n07714571_5107.JPEG', 'n07714571_4109.JPEG', 'n07714571_11878.JPEG', 'n07714571_15910.JPEG', 'n07714571_14401.JPEG', 'n07714571_2741.JPEG', 'n07714571_8576.JPEG', 'n07714571_1624.JPEG', 'n07714571_13479.JPEG', 'n07714571_2715.JPEG', 'n07714571_3676.JPEG', 'n07714571_12371.JPEG', 'n07714571_4829.JPEG', 'n07714571_3922.JPEG', 'n07714571_10377.JPEG', 'n07714571_8040.JPEG', 'n07714571_8147.JPEG', 'n07714571_10377.JPEG', 'n07714571_8040.JPEG', 'n07714571_5730.JPEG', 'n07714571_16460.JPEG', 'n07714571_8198.JPEG', 'n07714571_1095.JPEG', 'n07714571_3922.JPEG', 'n07714571_7745.JPEG', 'n07714571_6301.JPEG'.