ADVERSARIAL MANIPULATION OF DEEP REPRESENTATIONS

Sara Sabour [*1], **Yanshuai Cao**[*1,2], **Fartash Faghri**[1,2] **& David J. Fleet**[1]
[1] Department of Computer Science, University of Toronto, Canada
[2] Architech Labs, Toronto, Canada
{saaraa,caoy,faghri,fleet}@cs.toronto.edu

ABSTRACT

We show that the image representations in a deep neural network (DNN) can be manipulated to mimic those of other natural images, with only minor, imperceptible perturbations to the original image. Previous methods for generating adversarial images focused on image perturbations designed to produce erroneous class labels. Here we instead concentrate on the internal layers of DNN representations, to produce a new class of adversarial images that differs qualitatively from others. While the adversary is perceptually similar to one image, its internal representation appears remarkably similar to a different image, from a different class and bearing little if any apparent similarity to the input. Further, they appear generic and consistent with the space of natural images. This phenomenon demonstrates the possibility to trick a DNN to confound almost any image with any other chosen image, and raises questions about DNN representations, as well as the properties of natural images themselves.

1 INTRODUCTION

Recent papers have shown that deep neural networks (DNNs) for image classification can be fooled, often using relatively simple methods to generate so-called *adversarial images* (Fawzi et al., 2015; Goodfellow et al., 2014; Gu & Rigazio, 2014; Nguyen et al., 2015; Szegedy et al., 2014; Tabacof & Valle, 2015). The existence of adversarial images is important, not just because they reveal weaknesses in learned representations and classifiers, but because 1) they provide opportunities to explore fundamental questions about the nature of DNNs, e.g., whether they are inherent in the network structure per se or in the learned models, and 2) such adversarial images might be harnessed to improve learning algorithms that yield better generalization and robustness (Goodfellow et al., 2014; Gu & Rigazio, 2014).

Research on adversarial images to date has focused mainly on disrupting classification, i.e., on algorithms that produce images classified with labels that are patently inconsistent with human perception. Given the large, potentially unbounded regions of feature space associated with a given class label, it may not be surprising that it is easy to disrupt classification. In this paper, in constrast to such *label adversaries*, we consider a new, somewhat more incidious class of adversarial images, called *feature adversaries*, which are confused with other images not just in the class label, but in their internal representations as well.

Given a source image, a target (guide) image, and a trained DNN, we find small perturbations to the source image that produce an internal representation that is remarkably similar to that of the guide image, and hence far from that of the source. With this new class of adversarial phenomena we demonstrate that it is possible to fool a DNN to confound almost any image with any other chosen image. We further show that the deep representations of such adversarial images are not outliers per se. Rather, they appear generic, indistinguishable from representations of natural images at multiple layers of a DNN. This phenomena raises questions about DNN representations, as well as the properties of natural images themselves.

[*]The first two authors contributed equally.

2 Related Work

Several methods for generating adversarial images have appeared in recent years. Nguyen et al. (2015) describe an evolutionary algorithm to generate images comprising 2D patterns that are classified by DNNs as common objects with high confidence (often 99%). While interesting, such adversarial images are quite different from the natural images used as training data. Because natural images only occupy a small volume of the space of all possible images, it is not surprising that discriminative DNNs trained on natural images have trouble coping with such out-of-sample data.

Szegedy et al. (2014) focused on adversarial images that appear natural. They used gradient-based optimization on the classification loss, with respect to the image perturbation, ϵ. The magnitude of the perturbation is penalized ensure that the perturbation is not perceptually salient. Given an image I, a DNN classifier f, and an erroneous label ℓ, they find the perturbation ϵ that minimizes $loss(f(I+\epsilon),\ell) + c\|\epsilon\|^2$. Here, c is chosen by line-search to find the smallest ϵ that achieves $f(I+\epsilon) = \ell$. The authors argue that the resulting adversarial images occupy low probability "pockets" in the manifold, acting like "blind spots" to the DNN. The adversarial construction in our paper extends the approach of Szegedy et al. (2014). In Sec. 3, we use gradient-based optimization to find small image perturbations. But instead of inducing misclassification, we induce dramatic changes in the internal DNN representation.

Later work by Goodfellow et al. (2014) showed that adversarial images are more common, and can be found by taking steps in the direction of the gradient of $loss(f(I+\epsilon),\ell)$. Goodfellow et al. (2014) also show that adversarial examples exist for other models, including linear classifiers. They argue that the problem arises when models are "too linear". Fawzi et al. (2015) later propose a more general framework to explain adversarial images, formalizing the intuition that the problem occurs when DNNs and other models are not sufficiently "flexible" for the given classification task.

In Sec. 4, we show that our new category of adversarial images exhibits qualitatively different properties from those above. In particular, the DNN representations of our adversarial images are very similar to those of natural images. They do not appear unnatural in any obvious way, except for the fact that they remain inconsistent with human perception.

3 Adversarial Image Generation

Let I_s and I_g denote the *source* and *guide* images. Let ϕ_k be the mapping from an image to its internal DNN representation at layer k. Our goal is to find a new image, I_α, such that the Euclidian distance between $\phi_k(I_\alpha)$ and $\phi_k(I_g)$ is as small as possible, while I_α remains close to the source I_s. More precisely, I_α is defined to be the solution to a constrained optimization problem:

$$I_\alpha = \arg\min_I \| \phi_k(I) - \phi_k(I_g) \|_2^2 \tag{1}$$

$$\text{subject to } \|I - I_s\|_\infty < \delta \tag{2}$$

The constraint on the distance between I_α and I_s is formulated in terms of the L_∞ norm to limit the maximum deviation of any single pixel color to δ. The goal is to constrain the degree to which the perturbation is perceptible. While the L_∞ norm is not the best available measure of human visual discriminability (e.g., compared to SSIM (Wang et al., 2004)), it is superior to the L_2 norm often used by others.

Rather than optimizing δ for each image, we find that a fixed value of $\delta = 10$ (out of 255) produces compelling adversarial images with negligible perceptual distortion. Further, it works well with different intermediate layers, different networks and most images. We only set δ larger when optimizing lower layers, close to the input (e.g., see Fig. 6). As δ increases distortion becomes perceptible, but there is little or no perceptible trace of the guide image in the distortion. For numerical optimization, we use l-BFGS-b, with the inequality (2) expressed as a box constraint around I_s.

Figure 1 shows nine adversarial images generated in this way, all using the well-known BVLC Caffe Reference model (Caffenet) (Jia et al., 2014). Each row in Fig. 1 shows a source, a guide, and three adversarial images along with their differences from the corresponding source. The adversarial examples were optimized with different perturbation bounds (δ), and using different layers, namely FC7 (fully connected level 7), P5 (pooling layer 5), and C3 (convolution layer 3). Inspecting the adversarial images, one can see that larger values of δ allow more noticeable perturbations. That

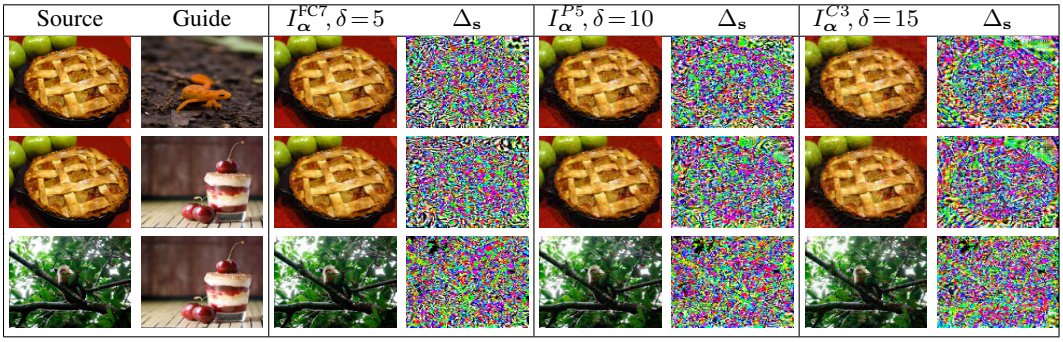


Figure 1: Each row shows examples of adversarial images, optimized using different layers of Caffenet (FC7, P5, and C3), and different values of $\delta = (5, 10, 15)$. Beside each adversarial image is the difference between its corresponding source image.

said, we have found no natural images in which the guide image is perceptible in the adversarial image. Nor is there a significant amount of salient structure readily visible in the difference images.

While the class label was not an explicit factor in the optimization, we find that class labels assigned to adversarial images by the DNN are almost always that of the guide. For example, we took 100 random source-guide pairs of images from Imagenet ILSVRC data (Deng et al., 2009), and applied optimization using layer FC7 of Caffenet, with $\delta = 10$. We found that class labels assigned to adversarial images were never equal to those of source images. Instead, in 95% of cases they matched the guide class. This remains true for source images from training, validation, and test ILSVRC data.

We found a similar pattern of behavior with other networks and datasets, including AlexNet (Krizhevsky et al., 2012), GoogleNet (Szegedy et al., 2015), and VGG CNN-S (Chatfield et al., 2014), all trained on the Imagenet ILSVRC dataset. We also used AlexNet trained on the Places205 dataset, and on a hybrid dataset comprising 205 scene classes and 977 classes from ImageNet (Zhou et al., 2014). In all cases, using 100 random source-guide pairs the class labels assigned to the adversarial images do not match the source. Rather, in 97% to 100% of all cases the predicted class label is that of the guide.

Like other approaches to generating adversarial images (e.g., Szegedy et al. (2014)), we find that those generated on one network are usually misclassified by other networks Using the same 100 source-guide pairs with each of the models above, we find that, on average, 54% of adversarial images obtained from one network are misclassified by other networks. That said, they are usually not consistently classified with the same label as the guide on different netowrks.

We next turn to consider internal representations – do they resemble those of the source, the guide, or some combination of the two? One way to probe the internal representations, following Mahendran & Vedaldi (2014), is to invert the mapping, thereby reconstructing images from internal representations at specific layers. The top panel in Fig. 2 shows reconstructed images for a source-guide pair. The *Input* row displays a source (left), a guide (right) and adervarisal images optimized to match representations at layers FC7, P5 and C3 of Caffenet (middle). Subsequent rows show reconstructions from the internal representations of these five images, again from layers C3, P5 and FC7. Note how lower layers bear more similarity to the source, while higher layers resemble the guide. When optimized using C3, the reconstructions from C3 shows a mixture of source and guide. In almost all cases we find that internal representations begin to mimic the guide at the layer targeted by the optimization. These reconstructions suggest that human perception and the DNN representations of these adversarial images are clearly at odds with one another.

The bottom panel of Fig. 2 depicts FC7 and P5 activation patterns for the source and guide images in Fig. 2, along with those for their corresponding adversarial images. We note that the adversarial activations are sparse and much more closely resemble the guide encoding than the source encoding. The supplementary material includes several more examples of adversarial images, their activation patterns, and reconstructions from intermediate layers.

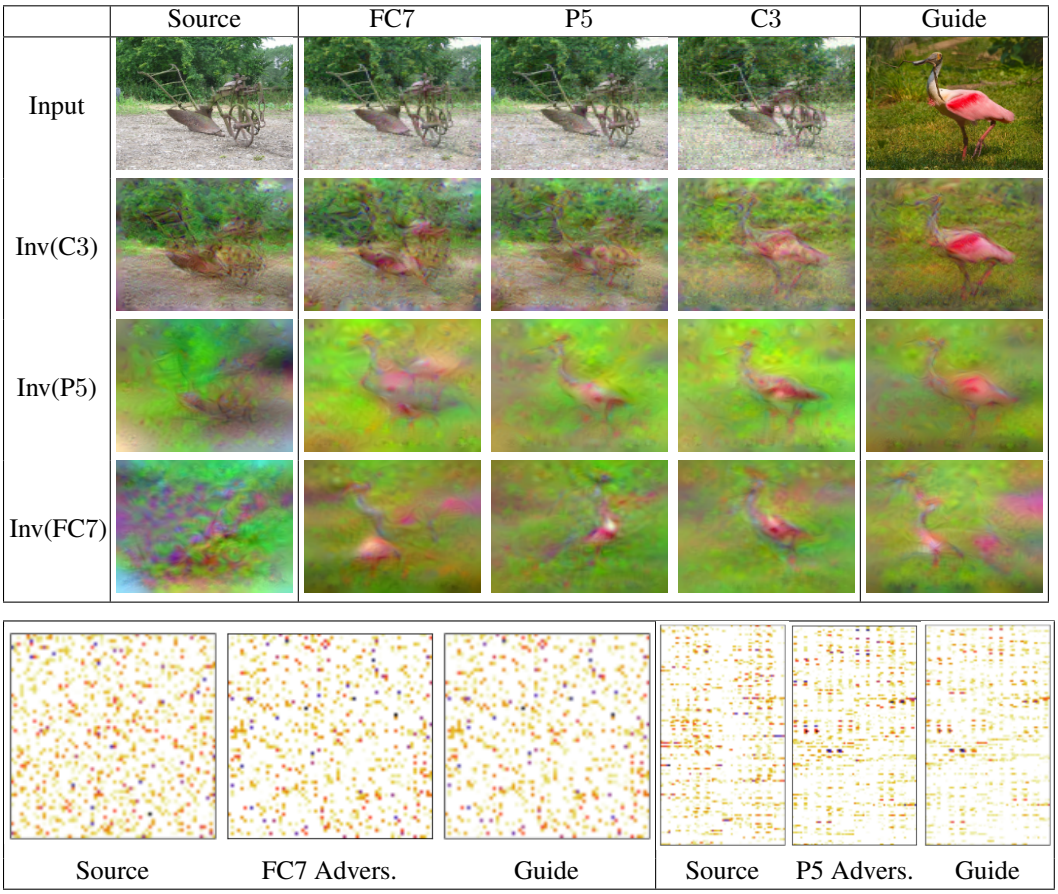


Figure 2: (Top Panel) The top row shows a source (left), a guide (right), and three adversarial images (middle), optimized using layers FC7, P5, and C3 of Caffenet. The next three rows show images obtained by inverting the DNN mapping, from layers C3, P5, and FC7 respectively (Mahendran & Vedaldi, 2014). (Lower Panel) Activation patterns are shown at layer FC7 for the source, guide and FC7 adversarial above, and at layer P5 for the source, guide and P5 adversarial image above.

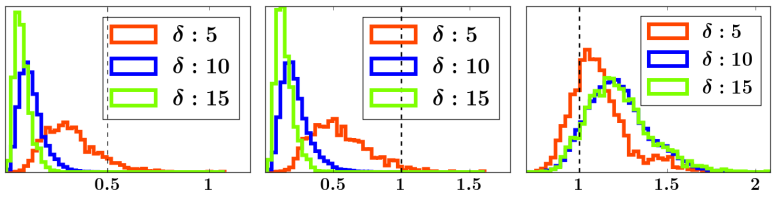
4 Experimental Evaluation

We investigate further properties of adversarial images by asking two questions. To what extent do internal representations of adversarial images resemble those of the respective guides, and are the representations unnatural in any obvious way? To answer these questions we focus mainly on Caffenet, with random pairs of source-guide images drawn from the ImageNet ILSVRC datasets.

4.1 Similarity to the Guide Representation

We first report quantitative measures of proximity between the source, guide, and adversarial image encodings at intermediate layers. Surprisingly, despite the constraint that forces adversarial and source images to remain perceptually indistinguishable, the intermediate representations of the adversarial images are much closer to guides than source images. More interestingly, the adversarial representations are often nearest neighbors of their respective guides. We find this is true for a remarkably wide range of natural images.

For optimizations at layer FC7, we test on a dataset comprising over 20,000 source-guide pairs, sampled from training, test and validation sets of ILSVRC, plus some images from Wikipedia to increase diversity. For layers with higher dimensionality (e.g., P5), for computational expedience, we use a smaller set of 2,000 pairs. Additional details about how images are sampled can be found in the supplementary material. To simplify the exposition in what follows, we use $\mathbf{s}$, $\mathbf{g}$ and $\boldsymbol{\alpha}$ to denote



(a) $d(\boldsymbol{\alpha},\mathbf{g})/d(\mathbf{s},\mathbf{g})$ (b) $d(\boldsymbol{\alpha},\mathbf{g})/\overline{d_1}(\mathbf{g})$ (c) $d(\boldsymbol{\alpha},\mathbf{s})/\overline{d}(\mathbf{s})$

Figure 3: Histogram of the Euclidean distances between FC7 adversarial encodings ($\boldsymbol{\alpha}$) and corresponding source ($\mathbf{s}$) and guide ($\mathbf{g}$), for optimizations targetting FC7. Here, $d(x, y)$ is the distance between x and y, $\overline{d}(\mathbf{s})$ denotes the average pairwise distances between points from images of the same class as the source, and $\overline{d_1}(\mathbf{g})$ is the average distance to the nearest neighbor encoding among images with the same class as the guide. Histograms aggregate over all source-guide pairs.

DNN representations of source, guide and adversarial images, whenever there is no confusion about the layer of the representations.

Euclidean Distance: As a means of quantifying the qualitative results in Fig. 2, for a large ensemble of source-guide pairs, all optimized at layer FC7, Fig. 3(a) shows a histogram of the ratio of Euclidean distance between adversarial $\boldsymbol{\alpha}$ and guide $\mathbf{g}$ in FC7, to the distance between source $\mathbf{s}$ and guide $\mathbf{g}$ in FC7. Ratios less than 0.5 indicate that the adversarial FC7 encoding is closer to $\mathbf{g}$ than $\mathbf{s}$. While one might think that the L_∞ norm constraint on the perturbation will limit the extent to which adversarial encodings can deviate from the source, we find that the optimization fails to reduce the FC7 distance ratio to less than 0.8 in only 0.1% of pairs when $\delta = 5$. Figure 6 below shows that if we relax the L_∞ bound on the deviation from the source image, then α is even closer to $\mathbf{g}$, and that adversarial encodings become closer to $\mathbf{g}$ as one goes from low to higher layers of a DNN.

Figure 3(b) compares the FC7 distances between $\boldsymbol{\alpha}$ and $\mathbf{g}$ to the average FC7 distance between representations of all ILSVRC training images from the same class as the guide and their FC7 nearest neighbors (NN). Not only is $\boldsymbol{\alpha}$ often the 1-NN of $\mathbf{g}$, but the distance between $\boldsymbol{\alpha}$ and $\mathbf{g}$ is much smaller than the distance between other points and their NN in the same class. Fig. 3(c) shows that the FC7 distance between $\boldsymbol{\alpha}$ and $\mathbf{s}$ is relatively large compared to typical pairwise distances between FC7 encodings of images of the source class. Only 8% of adversarial images (at $\delta = 10$) are closer to their source than the average pairwise FC7 distance within the source class.

Intersection and Average Distance to Nearest Neighbors: Looking at one's nearest neighbors provides another measure of similarity. It is useful when densities of points changes significantly through feature space, in which case Euclidean distance may be less meaningful. To this end we quantify similarity through rank statistics on near neighbors. We take the average distance to a point's K NNs as a scalar score for the point. We then rank that point along with all other points of the same label class within the training set. As such, the rank is a non-parametric transformation of average distance, but independant of the unit of distance. We denote the rank of a point x as $r_K(x)$; we use $K = 3$ below. Since $\boldsymbol{\alpha}$ is close to $\mathbf{g}$ by construction, we exclude $\mathbf{g}$ when finding NNs for adversarial points $\boldsymbol{\alpha}$.

Table 1 shows 3NN intersection as well as the difference in rank between adversarial and guide encodings, $\Delta r_3(\boldsymbol{\alpha}, \mathbf{g}) = r_3(\boldsymbol{\alpha}) - r_3(\mathbf{g})$. When $\boldsymbol{\alpha}$ is close enough to $\mathbf{g}$, we expect the intersection to be high, and rank differences to be small in magnitude. As shown in Table 1, in most cases they share exactly the same 3NN; and in at least 50% of cases their rank is more similar than 90% of data points in that class. These results are for sources and guides taken from the ILSVRC training set. The same statistics are observed for data from test or validation sets.

4.2 Similarity to Natural Representations

Having established that internal representations of adversarial images ($\boldsymbol{\alpha}$) are close to those of guides ($\mathbf{g}$), we then ask, to what extent are they typical of natural images? That is, in the vicinity of $\mathbf{g}$, is $\boldsymbol{\alpha}$ an inlier, with the same characteristics as other points in the neighborhood? We answer this question by examining two neighborhood properties: 1) a probabilistic parametric measure giving the log

Model	Layer	$\cap$3NN $= 3$	$\cap$3NN ≥ 2	Δr_3 median, [min, max] (%)
CaffeNet (Jia et al., 2014)	FC7	71	95	$-5.98, [-64.69, 0.00]$
AlexNet (Krizhevsky et al., 2012)	FC7	72	97	$-5.64, [-38.39, 0.00]$
GoogleNet (Szegedy et al., 2015)	pool5/7 $\times$ 7_s1	87	100	$-1.94, [-12.87, 0.10]$
VGG CNN S (Chatfield et al., 2014)	FC7	84	100	$-3.34, [-26.34, 0.00]$
Places205 AlexNet (Zhou et al., 2014)	FC7	91	100	$-1.24, [-18.20, 8.04]$
Places205 Hybrid (Zhou et al., 2014)	FC7	85	100	$-1.25, [-8.96, 8.29]$

Table 1: Results for comparison of nearest neighbors of the adversarial and guide. We randomly select 100 pairs of guide and source images such that the guide is classified correctly and the source is classified to a different class. The optimization is done for a maximum of 500 iterations, with $\delta = 10$. The statistics are in percentiles.

likelihood of a point relative to the local manifold at $\mathbf{g}$; 2) a geometric non-parametric measure inspired by high dimensional outlier detection methods.

For the analysis that follows, let $\mathcal{N}_K(x)$ denote the set of K NNs of point x. Also, let N_{ref} be a set of reference points comprising 15 random points from $\mathcal{N}_{20}(\mathbf{g})$, and let N_c be the remaining "close" NNs of the guide, $N_c = \mathcal{N}_{20}(\mathbf{g}) \setminus N_{ref}$. Finally, let $N_f = \mathcal{N}_{50}(\mathbf{g}) \setminus \mathcal{N}_{40}(\mathbf{g})$ be the set of "far" NNs of the guide. The reference set N_{ref} is used for measurement construction, while $\boldsymbol{\alpha}$, N_c and N_f are scored relative to $\mathbf{g}$ by the two measures mentioned above. Because we use up to 50 NNs, for which Euclidean distance might not be meaningful similarity measure for points in a high-dimensional space like P5, we use cosine distance for defining NNs. (The source images used below are the same 20 used in Sec. 4.1. For expedience, the guide set is a smaller version of that used in Sec. 4.1, comprising three images from each of only 30 random classes.)

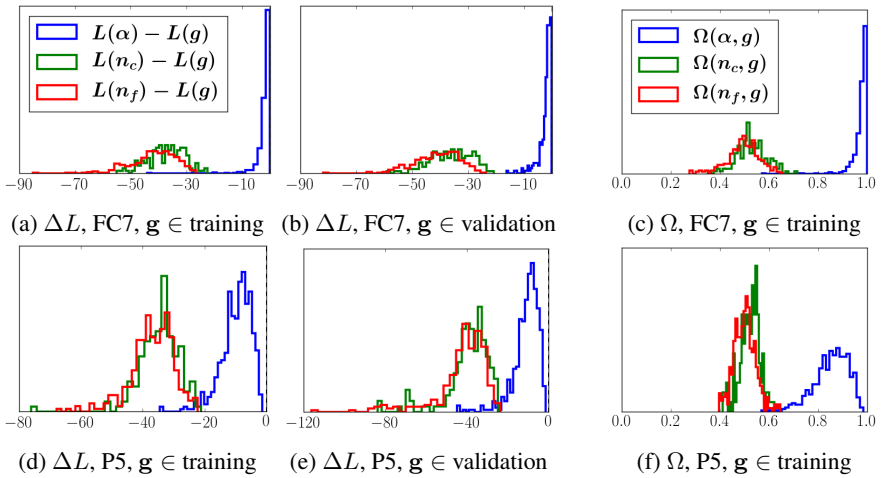
Manifold Tangent Space: We build a probabilistic subspace model with probabilistic PCA (PPCA) around $\mathbf{g}$ and compare the likelihood of $\boldsymbol{\alpha}$ to other points. More precisely, PPCA is applied to N_{ref}, whose principal space is a secant plane that has approximately the same normal direction as the tangent plane, but generally does not pass through $\mathbf{g}$ because of the curvature of the manifold. We correct this small offset by shifting the plane to pass through $\mathbf{g}$; with PPCA this is achieved by moving the mean of the high-dimensional Gaussian to $\mathbf{g}$. We then evaluate the log likelihood of points under the model, relative to the log likelihood of $\mathbf{g}$, denoted $\Delta L(\cdot, \mathbf{g}) = L(\cdot) - L(\mathbf{g})$. We repeat this measurement for a large number of guide and source pairs, and compare the distribution of ΔL for $\boldsymbol{\alpha}$ with points in N_c and N_f.

For guide images sampled from ILSVRC training and validation sets, results for FC7 and P5 are shown in the first two columns of Fig. 4. Since the Gaussian is centred at $\mathbf{g}$, ΔL is bounded above by zero. The plots show that $\boldsymbol{\alpha}$ is well explained locally by the manifold tangent plane. Comparing $\boldsymbol{\alpha}$ obtained when $\mathbf{g}$ is sampled from training or validation sets (Fig. 4(a) vs 4(b), 4(d) vs 4(e)), we observe patterns very similar to those in plots of the log likelihood under the local subspace models. This suggests that the phenomenon of adversarial perturbation in Eqn. (1) is an intrinsic property of the representation itself, rather than the generalization of the model.

Angular Consistency Measure: If the NNs of $\mathbf{g}$ are sparse in the high-dimensional feature space, or the manifold has high curvature, a linear Gaussian model will be a poor fit. So we consider a way to test whether $\boldsymbol{\alpha}$ is an inlier in the vicinity of $\mathbf{g}$ that does not rely on a manifold assumption. We take a set of reference points near a $\mathbf{g}$, N_{ref}, and measure directions from $\mathbf{g}$ to each point. We then compare the directions from $\mathbf{g}$ with those from $\boldsymbol{\alpha}$ and other nearby points, e.g., in N_c or N_f, to see whether $\boldsymbol{\alpha}$ is similar to other points around $\mathbf{g}$ in terms of *angular consistency*. Compared to points within the local manifold, a point far from the manifold will tend to exhibit a narrower range of directions to others points in the manifold. Specifically, given reference set N_{ref}, with cardinality k, and with z being $\boldsymbol{\alpha}$ or a point from N_c or N_f, our angular consistency measure is defined as

$$\Omega(z, \mathbf{g}) = \frac{1}{k} \sum_{x_i \in N_{ref}} \frac{\langle x_i - z,\, x_i - \mathbf{g} \rangle}{\|x_i - z\| \, \|x_i - \mathbf{g}\|} \tag{3}$$

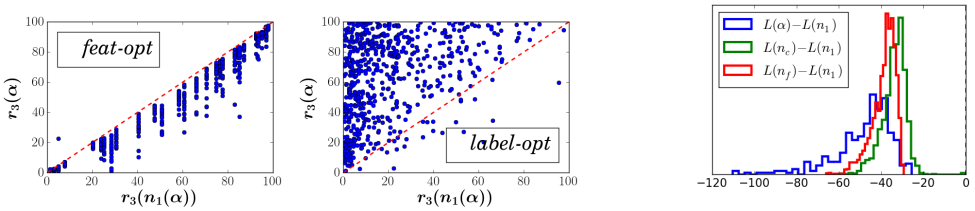
Fig. 4(c) and 4(f) show histograms of $\Omega(\boldsymbol{\alpha}, \mathbf{g})$ compared to $\Omega(n_c, \mathbf{g})$ where $n_c \in N_c$ and $\Omega(n_f, \mathbf{g})$ where $n_f \in N_f$. Note that maximum angular consistency is 1, in which case the point behaves like $\mathbf{g}$. Other than differences in scaling and upper bound, the angular consistency plots 4(c) and 4(f) are strikingly similar to those for the likelihood comparisons in the first two columns of Fig. 4, supporting the conclusion that $\boldsymbol{\alpha}$ is an inlier with respect to representations of natural images.



(a) ΔL, FC7, $\mathbf{g} \in$ training (b) ΔL, FC7, $\mathbf{g} \in$ validation (c) Ω, FC7, $\mathbf{g} \in$ training

(d) ΔL, P5, $\mathbf{g} \in$ training (e) ΔL, P5, $\mathbf{g} \in$ validation (f) Ω, P5, $\mathbf{g} \in$ training

Figure 4: Manifold inlier analysis: the first two columns (4(a),4(b),4(d),4(e)) for results of manifold tangent space analysis, showing distribution of difference in log likelihood of a point and $\mathbf{g}$, $\Delta L(\cdot, \mathbf{g}) = L(\cdot) - L(\mathbf{g})$; the last column (4(c)),(4(f)) for angular consistency analysis, showing distribution of angular consistency $\Omega(\cdot, g)$, between a point and $\mathbf{g}$. See Eqn. 3 for definitions.



(a): Rank of adversaries vs rank of $n_1(\alpha)$: Average distance of 3-NNs is used to rank all points in predicted class (excl. guide). Adversaries with same horizontal coordinate share the same guide.

(b): Manifold analysis for label-opt adversaries, at layer FC7, with tangent plane through $n_1(\alpha)$.

Figure 5: Label-opt and feature-opt PPCA and rank measure comparison plots.

4.3 Comparisons and Analysis

We now compare our feature adversaries to images created to optimize mis-classification (Szegedy et al., 2014), in part to illustrate qualitative differences. We also investigate if the linearity hypothesis for mis-classification adversaries of Goodfellow et al. (2014) is consistent with and explains with our class of adversarial examples. We hereby refer to our results as *feature adversaries via optimization (feature-opt)*. The adversarial images designed to trigger mis-classification via optimization (Szegedy et al., 2014), described briefly in Sec. 2, are referred to as *label adversaries via optimization (label-opt)*.

Comparison to label-opt: To demonstrate that label-opt differs qualitatively from feature-opt, we report three empirical results. First, we rank $\boldsymbol{\alpha}$, $\mathbf{g}$, and other points assigned the same class label as $\mathbf{g}$, according to their average distance to three nearest neighbours, as in Sec. 4.1. Fig. 5(a) shows rank of $\boldsymbol{\alpha}$ versus rank of its nearest neighbor-$n_1(\boldsymbol{\alpha})$ for the two types of adversaries. Unlike feature-opt, for label-opt, the rank of $\boldsymbol{\alpha}$ does not correlate well with the rank of $n_1(\boldsymbol{\alpha})$. In other words, for feature-opt $\boldsymbol{\alpha}$ is close to $n_1(\boldsymbol{\alpha})$, while for label-opt it is not.

Second, we use the manifold PPCA approach in Sec. 4.2. Comparing to peaked histogram of standardized likelihood of feature-opt shown in Fig. 4, Fig. 5(b) shows that label-opt examples are not represented well by the Gaussian around the first NN of $\boldsymbol{\alpha}$.

Third, we analyze the sparsity patterns on different DNN layers for different adversarial construction methods. It is well known that DNNs with ReLU activation units produce sparse activations (Glorot et al. (2011)). Therefore, if the degree of sparsity increases after the adversarial perturbation, the

	ΔS		I/U with s	
	feature-opt	label-opt	feature-opt	label-opt
FC7	7 ± 7	13 ± 5	$\mathbf{12 \pm 4}$	39 ± 9
C5	0 ± 1	0 ± 0	33 ± 2	70 ± 5
C3	2 ± 1	0 ± 0	60 ± 1	85 ± 3
C1	0 ± 0	0 ± 0	78 ± 0	94 ± 1

Table 2: Sparsity analysis: Sparsity is quantified as a percentage of the size of each layer.

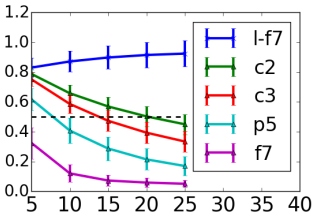


Figure 6: Distance ratio $d(\boldsymbol{\alpha},\mathbf{g})/d(\mathbf{s},\mathbf{g})$ vs δ. C2, C3, P5, F7 are for *feature-opt* adversaries; ℓ-f7 denotes FC7 distances for *feature-linear*.

adversarial example is using additional paths to manipulate the resulting represenation. We also investigate how many activated units are shared between the source and the adversary, by computing the intersection over union I/U of active units. If the I/U is high on all layers, then two represenations share most active paths. On the other hand, if I/U is low, while the degree of sparsity remains the same, then the adversary must have closed some activation paths and opened new ones. In Table 2, ΔS is the difference between the proportion of non-zero activations on selected layers between the source image represenation for the two types of adversaries. One can see that for all except FC7 of label-opt, the difference is significant. The column "I/U with s" also shows that feature-opt uses very different activation paths from s when compared to label-opt.

Testing The Linearity Hypothesis for feature-opt: Goodfellow et al. (2014) suggests that the existence of label adversaries is a consequence of networks being too linear. If this linearity hypothesis applies to our class of adversaries, it should be possible to linearize the DNN around the source image, and then obtain similar adversaries via optimization. Formally, let $J_s = J(\phi(I_s))$ be the Jacobian matrix of the internal layer encoding with respect to source image input. Then, the linearity hypothesis implies $\phi(I) \approx \phi(I_s) + J_s^\top (I - I_s)$. Hence, we optimize $\|\phi(I_s) + J_s^\top (I - I_s) - \phi(I_g)\|_2^2$ subject to the same infinity norm constraint in Eqn. 2. We refer to these adversaries as *feature-linear*.

As shown in Fig. 6, such adversaries do not get particularly close to the guide. They get no closer than 80%, while for *feature-opt* the distance is reduced to 50% or less for layers down to C2. Note that unlike *feature-opt*, the objective of *feature-linear* does not guarantee a reduction in distance when the constraint on δ is relaxed. These results suggest that the linearity hypothesis may not explain the existence of *feature-opt* adversaries.

Networks with Random Weights: We further explored whether the existence of *feature-opt* adversaries is due to the learning algorithm and the training set, or to the structure of deep networks per se. For this purpose, we randomly initialized layers of Caffenet with orthonormal weights. We then optimized for adversarial images as above, and looked at distance ratios (as in Fig. 3). Interestingly, the distance ratios for FC7 and Norm2 are similar to Fig. 6 with at most 2% deviation. On C2, the results are at most 10% greater than those on C2 for the trained Caffenet. We note that both Norm2 and C2 are overcomplete representations of the input. The table of distance ratios can be found in the Supplementary Material. These results with random networks suggest that the existence of *feature-opt* adversaries may be a property of the network architecture.

5 Discussion

We introduce a new method for generating adversarial images that appear perceptually similar to a given source image, but whose deep representations mimic the characteristics of natural guide images. Indeed, the adversarial images have representations at intermediate layers appear quite natural and very much like the guide images used in their construction. We demonstrate empirically that these imposters capture the generic nature of their guides at different levels of deep representations. This includes their proximity to the guide, and their locations in high density regions of the feature space. We show further that such properties are not shared by other categories of adversarial images.

We also find that the linearity hypothesis (Goodfellow et al., 2014) does not provide an obvious explanation for these new adversarial phenomena. It appears that the existence of these adversarial images is not predicated on a network trained with natural images per se. For example, results on random networks indicate that the structure of the network itself may be one significant factor.

Nevertheless, further experiments and analysis are required to determine the true underlying reasons for this discrepancy between human and DNN representations of images.

Another future direction concerns the exploration of failure cases we observed in optimizing feature adversaries. As mentioned in supplementary material, such cases involve images of hand-written digits, and networks that are fine-tuned with images from a narrow domain (e.g., the Flicker Style dataset). Such failures suggest that our adversarial phenomena may be due to factors such as network depth, receptive field size, or the class of natural images used. Since our aim here was to analyze the representation of well-known networks, we leave the exploration of these factors to future work. Another interesting question concerns whether existing discriminative models might be trained to detect feature adversaries. Since training such models requires a diverse and relatively large dataset of adversarial images we also leave this to future work.

ACKNOWLEDGMENTS Financial support for this research was provided, in part, by MITACS, NSERC Canada, and the Canadian Institute for Advanced Research (CIFAR). We would like to thank Foteini Agrafioti for her support. We would also like to thank Ian Goodfellow, Xavier Boix, as well as the anoynomous reviewers for helpful feedback.

REFERENCES

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 3, 6

Deng, J, Dong, W, Socher, R, Li, LJ, Li, K, and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pp. 248–255, 2009. 3

Fawzi, A, Fawzi, O, and Frossard, P. Fundamental limits on adversarial robustness. In *ICML*, 2015. 1, 2

Glorot, X, Bordes, A, and Bengio, Y. Deep sparse rectifier neural networks. In *AISTATS*, volume 15, pp. 315–323, 2011. 7

Goodfellow, IJ, Shlens, J, and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR (arXiv:1412.6572)*, 2014. 1, 2, 7, 8, 11

Gu, S and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. In *Deep Learning and Representation Learning Workshop (arXiv:1412.5068)*, 2014. 1

Jia, Y, Shelhamer, E, Donahue, J, Karayev, S, Long, J, Girshick, R, Guadarrama, S, and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *ACM Int. Conf. Multimedia*, pp. 675–678, 2014. 2, 6

Krizhevsky, A, Sutskever, I, and Hinton, GE. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012. 3, 6

Mahendran, A and Vedaldi, A. Understanding deep image representations by inverting them. In *IEEE CVPR (arXiv:1412.0035)*, 2014. 3, 4

Nguyen, A, Yosinski, J, and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE CVPR (arXiv:1412.1897)*, 2015. 1, 2

Szegedy, C, Zaremba, W, Sutskever, I, Bruna, J, Erhan, D, Goodfellow, I, and Fergus, R. Intriguing properties of neural networks. In *ICLR (arXiv:1312.6199)*, 2014. 1, 2, 3, 7

Szegedy, C, Liu, W, Jia, Y, Sermanet, P, Reed, S, Anguelov, D, Erhan, D, Vanhoucke, V, and Rabinovich, A. Going deeper with convolutions. In *CVPR*, 2015. 3, 6

Tabacof, P and Valle, E. Exploring the space of adversarial images. *arXiv preprint arXiv:1510.05328*, 2015. 1

Wang, Z, Bovik, AC, Sheikh, HR, and Simoncelli, EP. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. PAMI*, 3(4):600–612, 2004. 2

Zhou, B, Lapedriza, A, Xiao, J, Torralba, A, and Oliva, A. Learning deep features for scene recognition using places database. In *NIPS*, pp. 487–495, 2014. 3, 6

SUPPLEMENTARY MATERIAL

S1 ILLUSTRATION OF THE IDEA

Fig. S1 illustrates the achieved goal in this paper. The image of the fancy car on the left is a training example from the ILSVRC dataset. On the right of it, there is an adversarial image that was generated by guiding the source image by an image of Max (the dog). While the two fancy car images are very close in image space, the activation pattern of the adversarial car is almost identical to that of Max. This shows that the mapping from the image space to the representation space is such that for each natural image, there exists a point in a small neighborhood in the image space that is mapped by the network to a point in the representation space that is in a small neighborhood of the representation of a very different natural image.

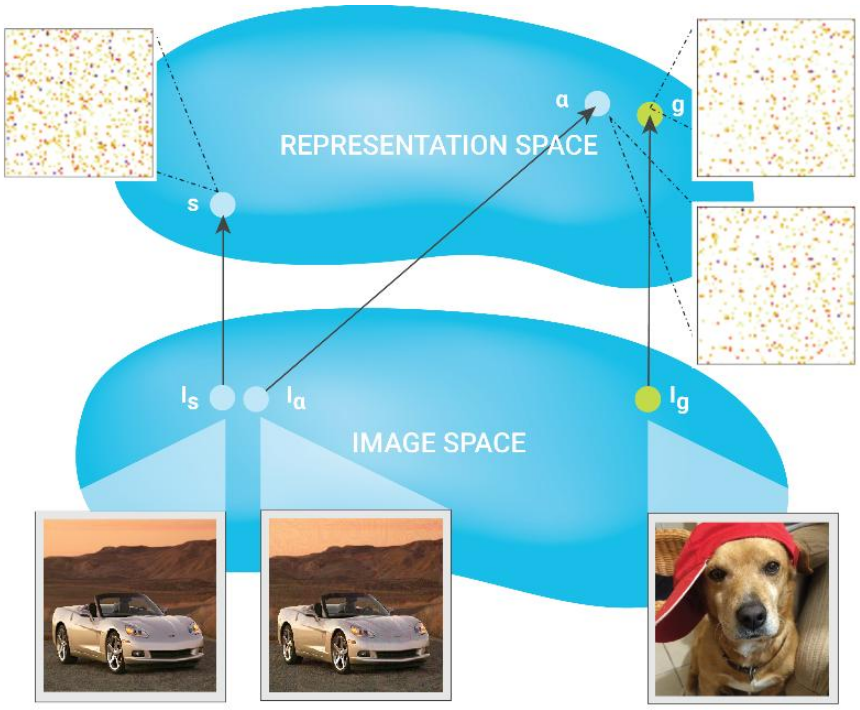


Figure S1: Summary of the main idea behind the paper.

S2 DATASETS FOR EMPIRICAL ANALYSIS

Unless stated otherwise, we have used the following two sets of source and guide images. The first set is used for experiments on layer FC7 and the second set is used for computational expedience on other layers (e.g. P5). The source images are guided by all guide images to show that the convergence does not depend on the class of images. To simplify the reporting of classification behavior, we only used guides from training set whose labels are correctly predicted by Caffenet.

In both sets we used 20 source images, with five drawn at random from each of the ILSVRC train, test and validation sets, and five more selected manually from Wikipedia and the ILSVRC validation set to provide greater diversity. The guide set for the first set consisted of three images from each of 1000 classes, drawn at random from ILSVRC training images, and another 30 images from each of the validation and test sets. For the second set, we drew guide images from just 100 classes.

S3 EXAMPLES OF ADVERSARIES

Fig. S2 shows a random sample of source and guide pairs along with their FC7 or Pool5 adversarial images. In none of the images the guide is perceptable in the adversary, regardless of the choice of source, guide or layer. The only parameter that affects the visibility of the noise is δ.

S4 DIMENSIONALITY OF REPRESENTATIONS

The main focus of this study is on the well-known Caffenet model. The layer names of this model and their representation dimensionalities are provided in Tab. S1.

Layer Name	Input	Conv2	Norm2	Conv3	Pool5	FC7
Dimensions	$3 \times 227 \times 227$	$256 \times 27 \times 27$	$256 \times 13 \times 13$	$384 \times 13 \times 13$	$256 \times 6 \times 6$	4096
Total	154587	186624	43264	64896	9216	4096

Table S1: Caffenet layer dimensions.

S5 RESULTS FOR NETWORKS WITH RANDOM WEIGHTS

As described in Sec. 4.3, we attempt at analyzing the architecture of Caffenet independent of the training by initializing the model with random weights and generating feature adversaries. Results in Tab. S2 show that we can generate feature adversaries on random networks as well. We use the ratio of distances of the adversary to the guide over the source to the guide for this analysis. In each cell, the mean and standard deviation of this ratio is shown for each of the three random, orthonormal random and trained Caffenet networks. The weights of the random network are drawn from the same distribution that Caffenet is initialized with. Orthorgonal random weights are obtained using singular value decomposition of the regular random weights.

Results in Tab. S2 indicate that convergence on Norm2 and Conv2 is almost similar while the dimensionality of Norm2 is quite smaller than Conv2. On the other hand, Fig. 6 shows that although Norm2 has smaller dimensionality than Conv3, the optimization converges to a closer point on Conv3 rather than Conv2 and hence Norm2. This means that the relation between dimensionality and the achieved distance of the adversary is not straightforward.

Layer	$\delta = 5$	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 25$
conv2	T:0.79 ± 0.04 OR:0.89 ± 0.03 R:0.90 ± 0.02	T:0.66 ± 0.06 OR:0.78 ± 0.05 R:0.81 ± 0.04	T:0.57 ± 0.06 OR:0.71 ± 0.07 R:0.74 ± 0.06	T:0.50 ± 0.07 OR:0.64 ± 0.09 R:0.67 ± 0.08	T:0.45 ± 0.07 OR:0.58 ± 0.10 R:0.61 ± 0.09
norm2	T:0.80 ± 0.04 OR:0.82 ± 0.05 R:0.85 ± 0.03	T:0.66 ± 0.05 OR:0.69 ± 0.08 R:0.73 ± 0.06	T:0.57 ± 0.06 OR:0.59 ± 0.10 R:0.63 ± 0.08	T:0.50 ± 0.06 OR:0.51 ± 0.11 R:0.55 ± 0.09	T:0.45 ± 0.06 OR:0.44 ± 0.11 R:0.48 ± 0.10
fc7	T:0.32 ± 0.10 OR:0.34 ± 0.12 R:0.52 ± 0.09	T:0.12 ± 0.06 OR:0.12 ± 0.09 R:0.26 ± 0.11	T:0.07 ± 0.04 OR:0.07 ± 0.06 R:0.13 ± 0.10	T:0.06 ± 0.03 OR:0.05 ± 0.04 R:0.07 ± 0.08	T:0.05 ± 0.02 OR:0.05 ± 0.02 R:0.04 ± 0.06

Table S2: Ratio of $d(\boldsymbol{\alpha},\mathbf{g})/d(\mathbf{s},\mathbf{g})$ as δ changes from 5 to 25 on randomly weighted(R), orthogonal randomly weighted(OR) and trained(T) Caffenet optimized on layers Conv2, Norm2 and FC7.

S6 ADVERSARIES BY FAST GRADIENT

As we discussed in Sec. 4.3, Goodfellow et al. (2014) also proposed a method to construct label adversaries efficiently by taking a small step consistent with the gradient. While this *fast gradient* method shines light on the label adversary misclassifications, and is useful for adversarial training, it is not relevant to whether the linearity hypothesis explains the feature adversaries. Therefore we omitted the comparison in Sec. 4.3 to fast gradient method, and continue the discussion here.

The fast gradient method constructs adversaries (Goodfellow et al. (2014)) by taking the perturbation defined by $\delta \text{sign}(\nabla_I loss(f(I), \ell))$, where f is the classifier, and ℓ is an erroneous label

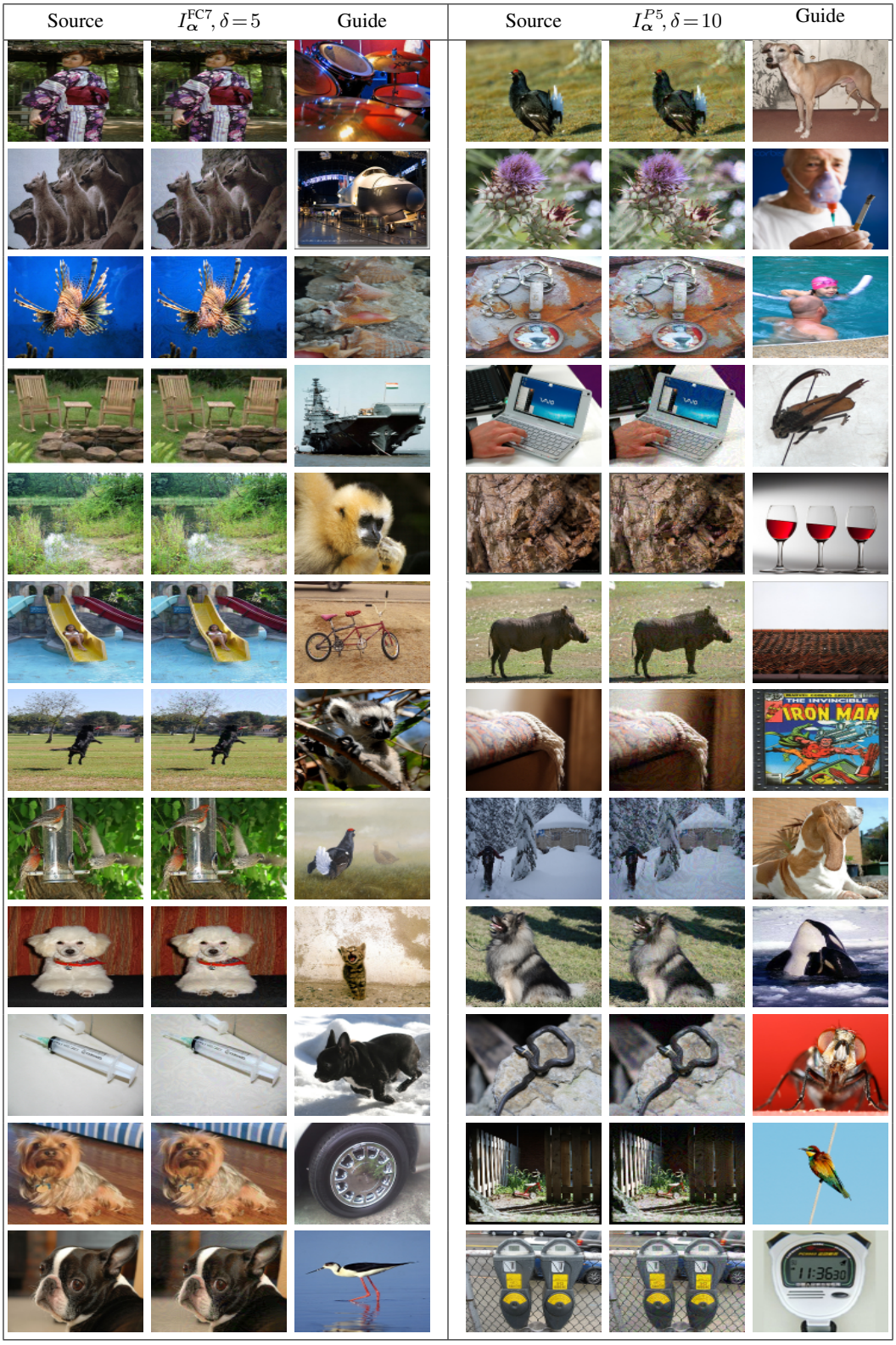
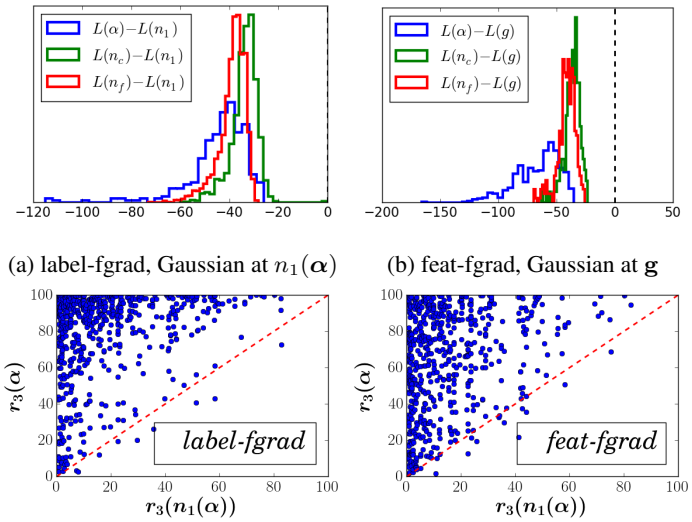


Figure S2: Each row shows examples of adversarial images, optimized using different layers of Caffenet (FC7, P5), and different values of $\delta = (5, 10)$.

for input image I. We refer to the resulting adversarial examples *label-fgrad*. We can also apply the fast gradient method to an internal representation, i.e. taking the perturbation defined by $\delta \text{sign}(\nabla_I \|\phi(I) - \phi(I_g)\|^2)$. We call this type *feature adversaries via fast gradient (feat-fgrad)*.

The same experimental setup as in Sec. 4.3 is used here. In Fig. S3, we show the nearest neighbor rank analysis and manifold analysis as done in Sec. 4.2 and Sec. 4.3. Moreover, Figs. S3(a)-S3(b) in compare to Figs. 4(a)-4(b) from *feature-opt* results and Fig. 5(b) from *label-opt* results indicates that this adversaries are not represented as well as *feature-opt* by a Gaussian around the NN of the adversary too. Also, Figs. S3(c)-S3(d) in compare to Fig. 5(a) show the obvious difference in adversarial distribution for the same set of source and guide.



(a) label-fgrad, Gaussian at $n_1(\boldsymbol{\alpha})$ (b) feat-fgrad, Gaussian at $\mathbf{g}$

(c) label-fgrad (d) feat-fgrad

Figure S3: Local property analysis of label-fgrad and feat-fgrad on FC7: S3(a)-S3(b) manifold analysis; S3(c)-S3(d) neighborhood rank analysis.

S7 Failure Cases

There are cases in which our optimization was not successful in generating good adversaries. We observed that for low resolution images or hand-drawn characters, the method does not always work well. It was successful on LeNet with some images from MNIST or CIFAR10, but for other cases we found it necessary to relax the magnitude bound on the perturbations to the point that traces of guide images were perceptible. With Caffenet, pre-trained on ImageNet and then fine-tuned on the Flickr Style dataset, we could readily generate adversarial images using FC8 in the optimization (i.e., the unnormalized class scores), however, with FC7 the optimization often terminated without producing adversaries close to guide images. One possible cause may be that the fine-tuning distorts the original natural image representation to benefit style classification. As a consequence, the FC7 layer no longer gives a good generic image represenation, and Euclidean distance on FC7 is no longer useful for the loss function.

S8 More Examples with Activation Patterns

Finally, we dedicate the remaining pages to several pairs of source and guide along with their adversaries, activation patterns and inverted images as a complementary to Fig. 2. Figs. S4, S5, S6, S7 and S8 all have similar setup as it is discussed in Sec. 3.

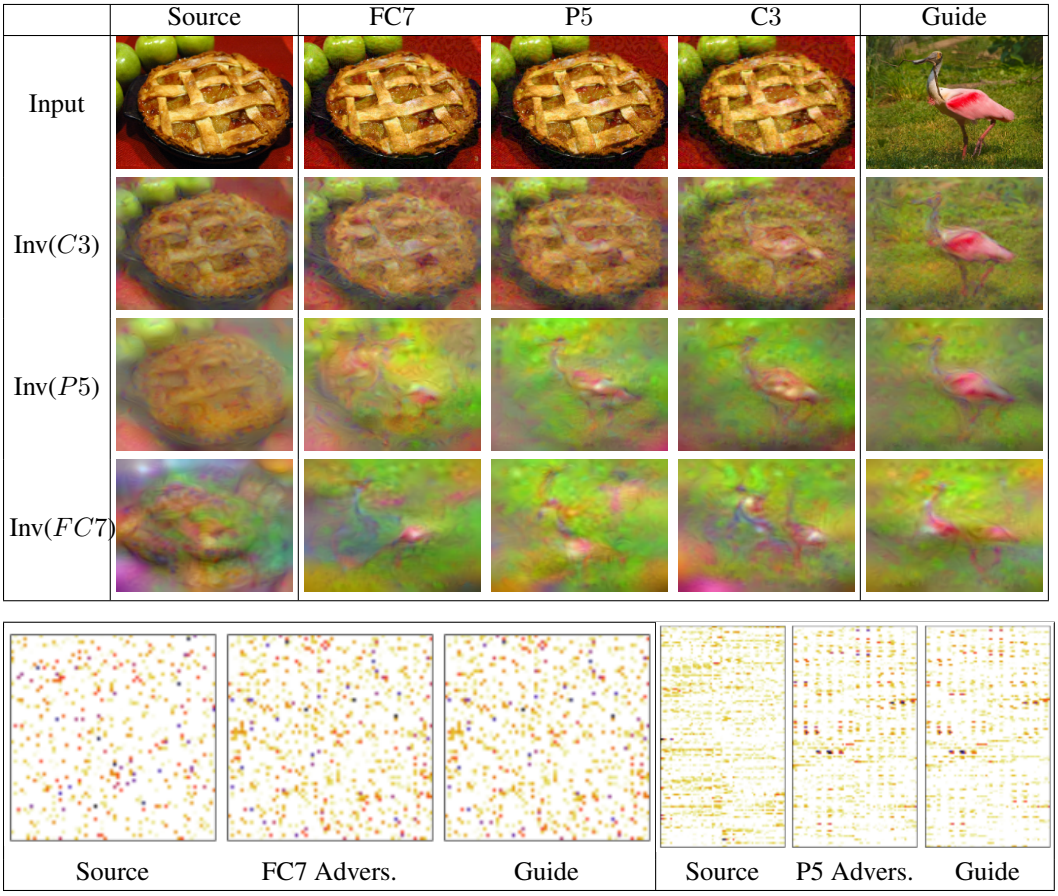


Figure S4: Inverted images and activation plot for a pair of source and guide image shown in the first row (Input). This figure has same setting as Fig. 2.

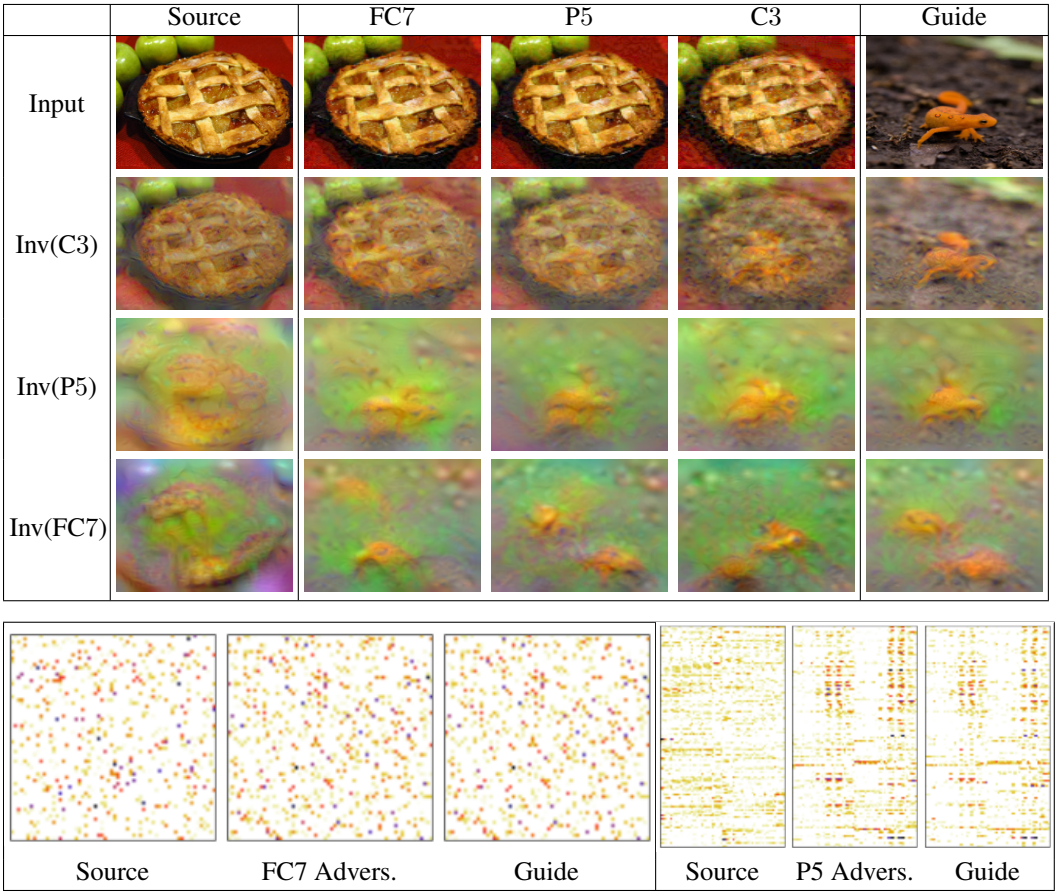


Figure S5: Inverted images and activation plot for a pair of source and guide image shown in the first row (Input). This figure has same setting as Fig. 2.

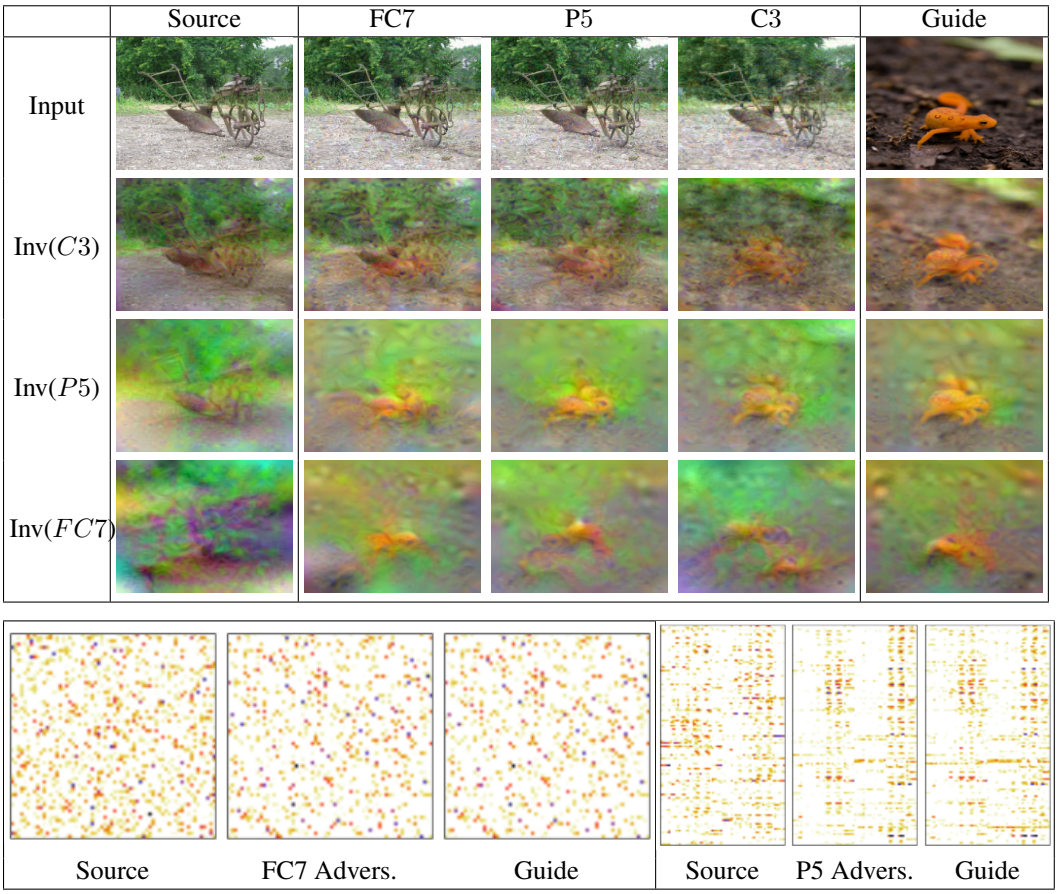


Figure S6: Inverted images and activation plot for a pair of source and guide image shown in the first row (Input). This figure has same setting as Fig. 2.

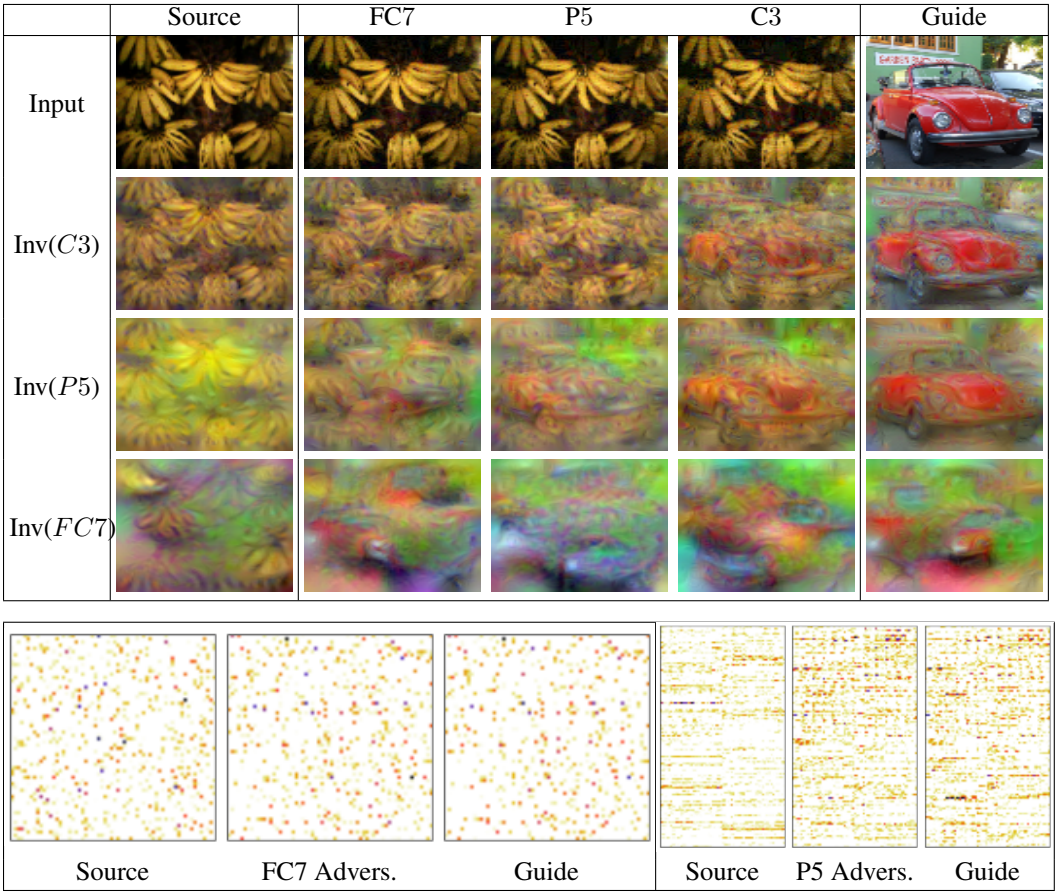


Figure S7: Inverted images and activation plot for a pair of source and guide image shown in the first row (Input). This figure has same setting as Fig. 2.

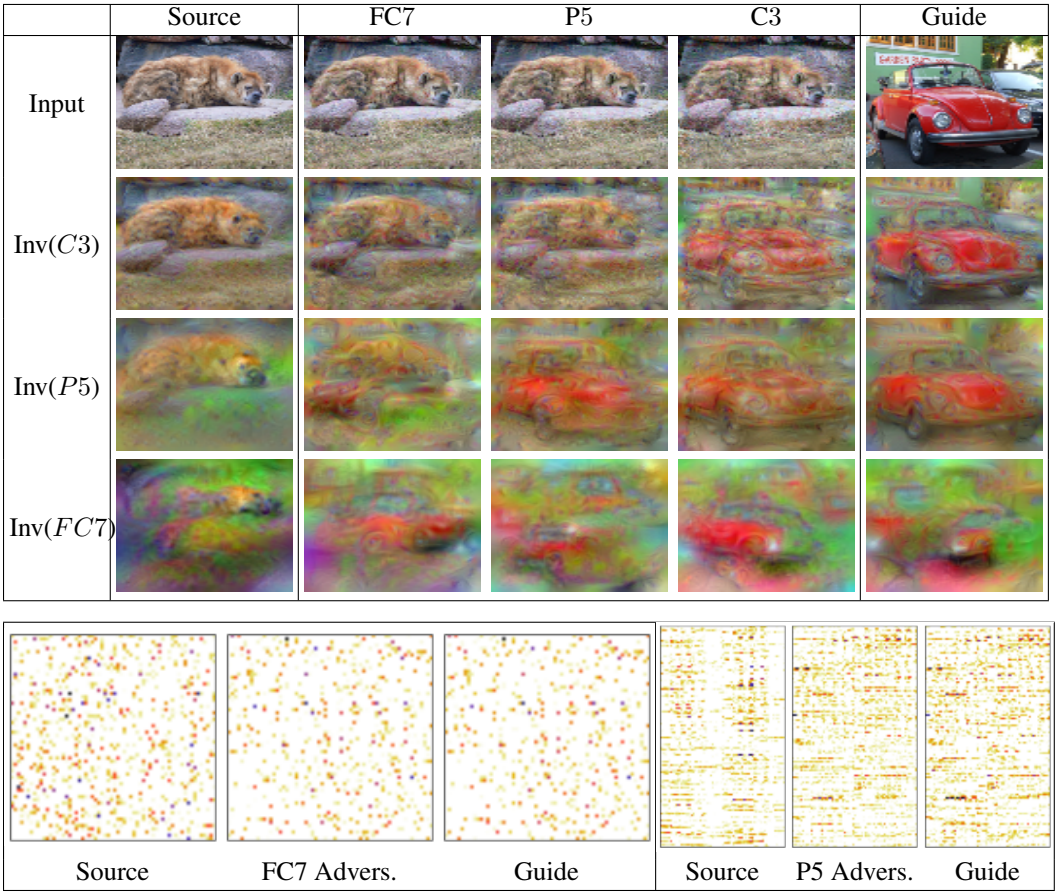


Figure S8: Inverted images and activation plot for a pair of source and guide image shown in the first row (Input). This figure has same setting as Fig. 2.