Adversarial Fooling Beyond "Flipping the Label"

Konda Reddy Mopuri*, Vaisakh Shaj*, R. Venkatesh Babu
Video Analytics Lab
Indian Institute of Science, Benguluru
kondamopuri@iisc.ac.in,vshaj@lincoln.ac.uk,venky@iisc.ac.in

Abstract

Recent advancements in CNNs have shown remarkable achievements in various CV/AI applications. Though CNNs show near human or better than human performance in many critical tasks, they are quite vulnerable to adversarial attacks. These attacks are potentially dangerous in real-life deployments. Though there have been many adversarial attacks proposed in recent years, there is no proper way of quantifying the effectiveness of these attacks. As of today, mere fooling rate is used for measuring the susceptibility of the models, or the effectiveness of adversarial attacks. Fooling rate just considers label flipping and does not consider the cost of such flipping, for instance, in some deployments, flipping between two species of dogs may not be as severe as confusing a dog category with that of a vehicle. Therefore, the metric to quantify the vulnerability of the models should capture the severity of the flipping as well. In this work we first bring out the drawbacks of the existing evaluation and propose novel metrics to capture various aspects of the fooling. Further, for the first time, we present a comprehensive analysis of several important adversarial attacks over a set of distinct CNN architectures. We believe that the presented analysis brings valuable insights about the current adversarial attacks and the CNN models.

1. Introduction

Machine learning (ML) models are observed (e.g. [1, 2]) to be unstable to addition of structured noises known as *adversarial perturbations*. These perturbations, despite being mild, tend to severely alter the inference of the ML models, generally referred to as *fooling* the models. Over the time, number of different adversarial attacks (algorithms to fool) were proposed (e.g. [23, 13, 3, 15, 16]) to demonstrate the vulnerability of the current ML systems, particularly the deep neural networks (DNNs). In case of a recognition model, it is understood that an adversarial attack is successful when the model predicts a different label upon adding the perturbation. Thus, all the existing works treat this *label flipping* as *fooling* the model. Therefore, they quantify the effectiveness of the underlying attack in terms of its *fooling or success rate*, which is the percentage of successful flips.

However, fooling rate is a weak metric which fails to capture various important aspects of *fooling* and ends up giving only partial picture about the attack or the target model. Specifically, it does not consider what the 'post-attack' label is, and therefore fails to quantify the severity of the attack on either semantic or visual scale. Consequently, the fooling rate becomes apathetic to different flippings of the label and treats them identical. Though, from the robustness perspective, all label flippings should be treated equally, and a robust ML system should avoid any such susceptibility, in practice, different flippings (misclassifications) may inflict in different costs. For instance, in certain deployments, confusion between a pair of dog breeds is acceptable and not as severe as wrongly recognizing the stop sign on a highway.

Particularly, datasets with unwanted bias towards a set of semantically similar categories (e.g. ImageNet [20] has 12% dog categories) need sophisticated metrics for better analysis of the attacks and models. In such cases, weak metrics such as fooling rate could be misleading by providing only an incomplete picture of the models' vulnerability. Existing evaluation (e.g. [19, 10, 12]) to compare various adversarial attacks is based solely on their fooling rate performance. However, the spectrum of existing attacks should be understood and analysed from not only the fooling rate perspective but also various other aspects of fooling, for instance, the actual semantic damage incurred due to the adversarial attack, etc. Moreover, fooling rate alone fails to bring out useful insights about the learning and the classification hyper-planes learned by these models. Therefore, the metric to quantify the effectiveness of the adversarial attacks should apprehend the severity of the flipping and provide better information about both the model and the underlying attack.

Hence, in this work, we present various important as-

* contributed equally

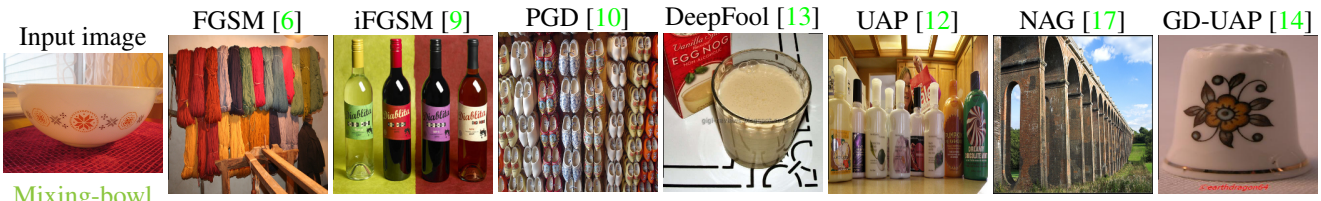


Figure 1. Fooling rate does not measure the extent of confusion caused by an adversarial attack. First column is the input image presented to GoogLeNet [22] below which the *pre-attack* label is mentioned in green. The subsequent columns show the representative images for the predicted *post-attack* labels for 7 different adversarial attacks. Note that the attacks are mentioned on top of the corresponding representative images taken from the ILSVRC dataset.

pects of fooling caused by adversarial attacks. Specifically, we consider the Convolutional Neural Networks (CNN) trained for visual object recognition. The major contributions of this work can be listed as:

- We present the shortcomings of the existing evaluation in order to emphasize the need for more sophisticated tools for analysis
- We propose a set of useful metrics (as baselines) to understand the attacks and models more comprehensively
- We present a detailed analysis and comparison of several important adversarial attacks over a set of distinct CNN architectures

The paper is organized as follows: section 2 discusses the shortcomings in the existing evaluation and proposes multiple novel metrics to bring out the various important aspects of the adversarial fooling, section 3 presents comprehensive empirical analysis on several important image agnostic and image specific adversarial attacks, section 4 narrates some of the important observations, and finally section 5 concludes the paper.

2. Fooling beyond *Flipping*

In this section we demonstrate how fooling rate fails to capture various important aspects of confusing a DNN and present a set of useful metrics to better understand it.

2.1. Shortcomings of Fooling Rate (FR)

Fooling rate, by definition is the percentage of success for an adversarial attack, i.e., expected number of times the attack is able to flip the label because of the added perturbation. For the rest of the paper, we define the label predicted on a clean image as *pre-attack* label and that predicted on the corresponding adversarial image as *post-attack* label. Thus, the fooling rate (FR) is defined as

$$\frac{\sum_{i=1}^{N} \mathbb{1}(\textit{pre-attack label}(i) \neq \textit{post-attack label}(i))}{N} \tag{1}$$

where $\mathbb{1}$ is the indicator function that returns 1 if the argument is true, else returns 0, and N is the total number of samples on which the attack is evaluated.

Clearly the fooling rate ignores what the *post-attack* label is. All that it matters is, if it is different from the *pre-attack* label or not. Thus, it does not measure the extent of the confusion caused by the adversarial attack. Since, some of the categories in the underlying dataset can be visually very close compared to others, not all mistakes are similar. For example, Figure 1 shows the *post-attack* labels predicted by GoogLeNet [22] for 7 different adversarial attacks. The input image is shown in the first column, who's *pre-attack* label is *mixing-bowl*. Note that the subsequent columns show (hand-picked) representative images for the predicted *post-attack* labels mentioned below them in red. Also, the corresponding adversarial attacks are mentioned above the representative images.

It is important to note that, though all the *post-attack* labels are different from the *pre-attack* label, visual patterns in some of the representative images are closer to the input image than others. For instance, *eggnog* and *thimble* will have bowl-like object patterns. On the other hand, *wool* and *viaduct* are visually very dissimilar to the *pre-attack* label, mixing-bowl. However, fooling rate treats them all as successful fooling without considering the perceptual distance. In the following subsections, we present multiple metrics that reveal more information about the adversarial fooling.

2.2. FR@K

We know that the fooling rate does not consider the rank of the *pre-attack* label after the attack. However, it would be interesting to know how strong an attack can demote the *pre-attack* label from rank-1, for instance, to compare different attacks or to understand the nature of the attack, etc. Therefore, we extend the definition of existing fooling rate in order to consider the rank of the *pre-attack* label using "Fooling rate at rank K" (FR@K). This means, for a given rank K, an attack is considered successful only if it assigns a rank $> K$ to the *pre-attack* label. Therefore, after the attack, there will be at least K-1 other labels with greater con-

fidence than the *pre-attack* label. Intuitively, the $FR@K$ metric quantifies the extent of damage caused to the visual features discriminative to the *pre-attack* label due to the attack. Thus, $FR@K$ is defined as

$$\frac{\sum_{i=1}^{N} \mathbb{1}(\textit{pre-attack label}(i) \notin \{\textit{top-K post-attack labels}(i)\})}{N} \tag{2}$$

Note that when K=1, FR@K becomes the fooling rate. A similar metric in spirit has been proposed by Ganeshan *et al.* [5] where they quantify the shifts in the ranks of *pre* and *post-attack* labels during the attack.

2.3. Mean semantic confusion: QI-Wup

Existing evaluation completely ignores to measure the "semantic damage" caused by the adversarial attacks. This is because the fooling rate is apathetic to different flippings by adversarial attacks. However, there exist various attacks that are designed with very different objective functions though the ultimate goal is to fool the target model. Thus, it is quite possible that a given model incurs varying levels of confusion for different adversarial attacks. For instance, in Figure 1, GoogLeNet confuses *Mixing-bowl* to a range of different labels from *wine bottle* to *viaduct*. Note that the *post-attack* labels resulted by different attacks lie at different semantic distance to the *pre-attack* label. Also, different deployment environments (e.g. household, commercial, military, etc.) would work with varied levels of acceptable confusion. It is beneficial to have a useful metric that can quantify the actual semantic damage incurred by a given model for various attacks.

In this subsection we introduce an intuitive metric named "Mean semantic confusion", that can quantify the semantic damage caused by an adversarial attack (or in other words, the semantic damage incurred by a given model). We adapt the familiar word similarity metrics such as Wu-Palmer [24] to measure the severity of the flipping on a semantic scale. We define the Quantized Inverse Wup similarity (QI-Wup) as

$$QI\text{-}Wup = \begin{cases} 1, & \text{if } Wup(\textit{pre-attacklabel}, \textit{post-attacklabel}) < T_s. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where $Wup(x, y)$ is the Wup similarity between the words x and y and T_s is a threshold chosen based on the target deployment environment. Note that the proposed $QI\text{-}Wup$ metric deems a *flipping* as *fooling* only when the semantic similarity is less than an acceptable threshold T_s. Different thresholds can be chosen for various deployment scenarios based on the acceptable semantic confusion.

Therefore, the mean semantic confusion can be computed as the average $QI\text{-}Wup$ score over a set of evaluation samples. Note that Wup measure is one choice of word (semantic) similarity and we can chose any other similarity.

2.4. Mean visual confusion: QI-Vis

Another very important aspect of adversarial fooling is to understand if the CNN models get confused only among "visually" closer labels or even the dissimilar ones (refer to section 4.1). For instance, as shown in Figure 1, GoogLeNet gets confused to recognize the *Mixing-bowl* to various labels. The Deepfool [13] attack flips the label to *eggnog*. Note that the eggnog images always have a bowl like container to hold it. Therefore, in this case the attack tries to fool the model to predict another class that has similar visual patterns. While in case of other attacks, particularly the image-agnostic attacks such as UAP [12], GD-UAP [14] it is less observed.

In case of measuring the semantic damage, we have a hierarchical structure (graph) such as WordNet [11] to understand how the labels are semantically related. However, such a data structure for 'visual' relations does not exist. It is very difficult to collect the visual similarities among the categories via human annotations given large number of classes and intra class variations. Thus, (similar to [18]) we collect these visual similarities from the learned model itself. The final layer of any classification layer will be a fully connected (fc) layer with a softmax nonlinearity. Each neuron in this layer corresponds to a class (c) and its activation is treated as the confidence/probability (S_c) predicted by the model to that class. The weights connecting previous layer to this neuron (W_c) can be considered as the template of the class (c) learned by the network. This is because, the confidence predicted (S_c) is proportional to the alignment of the previous layer's output with the template (W_c). It becomes maximum when the previous layer's output is a positive scaled version of this template (W_c). On the other hand, if the output of the previous layer is misaligned with the template W_c, the confidence S_c is reduced.

Therefore, we compute the visual similarity (as perceived by the target model) between a pair of classes i and j as

$$Vis(i, j) = \frac{{W_i}^T W_j}{\|W_i\| \, \|W_j\|} \tag{4}$$

Using this similarity score, we define the Quantized Inverse Visual Similarity $(QI\text{-}Vis)$ as

$$QI\text{-}Vis = \begin{cases} 1, & \text{if } Vis(\textit{pre-attack} \text{ label}, \textit{post-attack} \text{ label}) < T_v. \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Where T_v is a threshold chosen in order to impose a desired tolerance in terms of visual confusion. Note that $Vis(i, j)$ may not lie in $[0, 1]$ unlike the Wup measure. Also, Vis is one of the possible visual similarity measures, and $QI\text{-}Vis$ can be computed over any such measure. Further, for a given dataset, our visual similarities are model specific, however it is still a valid candidate for visual similarity since

we observe that multiple models closely agree upon these similarities. For the 4 models considered in our experiments, the variance of class similarities computed across the 1000 ILSVRC categories is very small with a mean value of $6 \times 10^{-4} (\pm 5.95 \times 10^{-4})$.

3. Experiments

In this section we present the experimental analysis to show the effectiveness of the aforementioned metrics. We performed all our experiments on the models trained for object recognition on ILSVRC [20] dataset. To be comprehensive, we considered models from different architecture families, namely, CaffeNet [8], GoogLeNet [22], VGG-19 [21], ResNet-152 [7]. Note that the evaluation is performed on 10000 correctly classified images from the validation set. We considered a range of adversarial attacks that include image specific, image agnostic and iterative variants. Specifically, we evaluated on the following

- Fast Gradient Sign Method (FGSM) [6]
- Projected Gradient Descent (PGD) [10]
- DeepFool [13]
- Carlini and Wagner (CW) [3]
- Univeral Adversarial Perturbations (UAP) [12]
- Generalizable Data-free UAP (GD-UAP) [14]

We briefly introduce these attacks along with the required notation for the ease of reference.

- X : clean image from the dataset.
- X^{adv} : potential adversarial image crafted from X.
- y_{true} : ground truth label corresponding to X.
- y_{pred} : prediction of the neural network for X.
- f : mapping function that represents neural network
- ϵ : strength of perturbation added to the clean image.
- J : loss function used to train the neural network.
- ∇J : gradient of the loss J with respect to image X.

Fast Gradient Sign Method (FGSM): is a simple way to craft adversaries. They linearly approximate the loss function and compute the gradient as the adversarial direction to perturb the input:

$$X^{adv} = X + \epsilon.sign(\nabla J(X, y_{true})) \tag{6}$$

I-FGSM-LL is a variety of the FGSM attack, in which we iteratively (with small steps) compute the perturbation in order to decrease the loss for predicting a 'least-likely' label.

Projected Gradient Descent (PGD): One can think of FGSM attack as a single-step scheme to maximize the loss function within the ϵ ball around X, which is represented by $\mathcal{S}$. A powerful attack would be an iterative variation, $FGSM^k$, which is essentially performing Projected Gradient Descent on the negative loss function:

$$X^{t+1} = \Pi_{X+\mathcal{S}}(X^t + \alpha.sign(\nabla J(X, y_{true}))) \tag{7}$$

where t is the iteration, and α is the maximum perturbation at each iteration.

DeepFool: defines an adversarial perturbation as the minimal perturbation v that is sufficient to change the inference of the classifier:

$$min_v \text{ subject to } f(X+v) \neq f(X) \tag{8}$$

Their algorithm is a greedy method that approximates the non-linear class boundaries as hyper-planes and in practice (generally) yields a small and effective perturbation.

Carlini and Wagner attack (CW): makes the perturbations quasi-imperceptible by minimizing the l_p norm. This is achieved by solving the following optimization problem:

$$\text{minimize } ||v||_p + c.g(X+v) \tag{9}$$
$$\text{such that } X + v \in [0, 1] \tag{10}$$

Here g is a surrogate objective function such that $g(x+v) \leq 0 \iff f(x+v) = t$. Here v is the adversarial perturbation, c is a constant, t is the target label and f is a mapping function that represents the neural network.

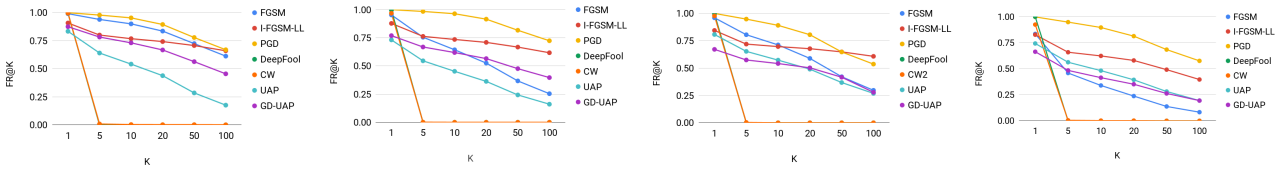
Universal Adversarial Perturbation (UAP): presented an algorithm to compute a perturbation agnostic to the input samples, known as 'Universal' Adversarial Perturbation (v).

$$||v||_p \leq \epsilon \tag{11}$$
$$\mathcal{P}_{X \sim \mu}(f(X+v) \neq f(X)) \geq 1 - Th \tag{12}$$

where, Th quantifies the desired fooling rate. Their algorithm simply accumulates the individual sample specific DeepFool [13] adversarial perturbations and regularly projects into the feasible ball of the perturbations.

Generalized Data-free UAP (GD-UAP): presented a data-free 'activation' loss that is generalizable across various vision tasks to compute a UAP. They attempt achieve an objective similar to eqn. (12) without utilizing any data samples via optimizing the following loss:

$$Loss = -\log(\prod_{i=1}^{K} ||l_i(v)||_2) \tag{13}$$
$$\text{such that } ||v||_p \leq \epsilon \tag{14}$$



CaffeNet GoogLeNet VGG−19 ResNet−152

Figure 2. FR@K computed for various CNN models for multiple adversarial attacks. Note that the attacks are mentioned in the legend and the model name is provided below the corresponding plot.

where, $l_i(v)$ is the response of i^{th} layer, and K is the total number of layers in the model.

We chose the best hyper-parameters for all the attacks (e.g. number of iterations, etc.) as mentioned in the corresponding works or via conducting ablation. Also, we consider l_∞ norm of 10 for restricting the strength of the perturbation (ϵ). However, note that the CW and DeepFool attacks because of their nature, do not impose the same max-norm restriction on the perturbations. In case of CW attack, we use a relatively large value of 15 for binary-search steps, which helps in determining the trade off-constant c.

In summary, we consider the best operating parameters for all the attacks in order to ensure the comparison is fair. We present some useful ablations over these hyper-parameters (e.g. CW) in the supplementary material.

3.1. FR@K

Fooling rate at rank K $(FR@K)$ gives the success rate of an attack to demote the *pre-attack* label beyond the first K ranks after the attack. Figure 2 shows $FR@K$ for various attacks computed on multiple models. We computed the results for $K = 1, 2, 5, 10, 20, 50,$ and, 100. Note that K can take a maximum value of 999 since the total number of categories in ILSVRC is 1000. We can notice that, as expected, the $FR@K$ falls with K, since it gets more difficult for the attack to demote the *pre-attack* label further.

However, we make a very important observation about the DeepFool [13] and CW [3] attacks. They could not demote the *pre-attack* label further. $FR@K$ becomes zero for all the higher values of K. Specifically, in all the 10000 correctly classified validation images that we considered for evaluation, the highest rank DeepFool could successfully demote to is 3. In almost all the cases, it makes the model to simply swap the first and second labels. Note that this behaviour is consistent across all the models. This behaviour can be explained from the design of the attack. The DeepFool algorithm searches for the nearest decision boundary to the input sample and finds a contamination in order to move the sample across that boundary. Therefore, it ends up fooling the model to predict the nearest class to the *pre-attack* label which is top-2 (or top-3) label before the attack. For the CW attack, we used a variant in which the adversary is crafted to make the *pre-attack* label least-likely (i.e. poorly ranked after the attack). We observed that the resulting attack's behaviour is similar to the 'best case' scenario proposed by the authors in which the *post-attack* label is the second most probable label before perturbing. Please note that the CW attack doesn't guarantee the l_p norm of the perturbation to be smaller than a predefined value, rather, it only minimises through the objective (eq. 9). Hence, we chose a variant whose average perturbation is comparable to the other attacks in order to provide a fair comparison. Please refer to the supplemental material for a stronger version of the CW attack.

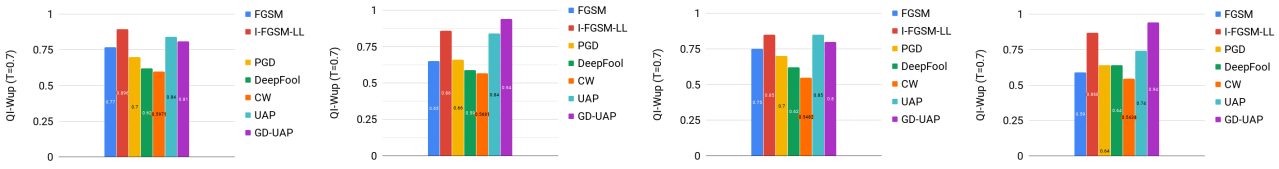
	CaffeNet	GoogLeNet	VGG-19	ResNet-152
FGSM	0.7	0.4	0.44	0.19
IFGSM-LL	0.66	0.61	0.59	0.43
PGD	**0.79**	**0.83**	**0.68**	**0.71**
DeepFool	0.02	0.02	0.02	0.02
CW	0.02	0.02	0.02	0.02
UAP	0.33	0.28	0.4	0.32
GD-UAP	0.58	0.5	0.42	0.29

Table 1. Area under the $FR@K$ curves for multiple models under various adversarial attacks.

Apart from extracting such hidden details about the attacks, we can also convert these graphs into metrics such as the area under the curve (AuC), which can be used for a direct quantitative comparison. For instance, if one looks for an attack that can demote the *pre-attack* label strongly, it should have a high AuC for the $FR@K$ plot. Similarly a robust model would have a low AuC for multiple attacks. Table 1 shows the AuC computed for the curves shown in Figure 2. In terms of performance with respect to AuC metric, attacks such as PGD and I-FGSM-LL inflict the maximum disruption to the visual features discriminative to the *pre-attack* label. Notice that the AuC generally decreases as the models get sophisticated from left to right. This increased robustness can be attributed to the advanced network architectures with efficient regularizers such as dropout and batchnorm.

3.2. QI-Wup

Figure 3 presents the $QI\text{-}Wup$ measure computed for multiple models over various attacks. For all these experiments we have used a threshold (T_s) of 0.7 on the Wup



CaffeNet GoogLeNet VGG−19 ResNet−152

Figure 3. QI-Wup computed for various CNN models for multiple adversarial attacks. Note that the attacks are mentioned in the legend and the model names are provided below the corresponding plot.

similarity.

It can be observed that image agnostic attacks such as UAP [12], GD-UAP [14] consistently result in a strong semantic damage compared to the image specific counterparts. This can also be explained from the fooling patterns of these attacks. In case of image agnostic attacks, the existence of 'dominant' *post-attack* labels is observed. That is, after the attack, the *pre-attack* labels are generally mapped to a small set of sink classes. For instance, in case of UAP, *post-attack* labels computed for all the 50000 validation images on GoogLeNet comprise only 17% of the total categories. It is hypothesized [12] that the dominant labels occupy large space and hence represent good candidates for these attacks to fool the models. However, this is not the case with the image specific attacks. It is observed that the DeepFool and CW attacks generally inflict the smallest semantic damage compared to the others. Note that this is consistent with the observation in section 3.1 that they result in least $FR@K$.

Among the image specific attacks, I-FGSM-LL inflicts maximum semantic confusion. This is understandable, since this attempts to make the model predict the least-likely label which is generally (visually) far away from the *pre-attack* label. Hence, in general, the two labels should also be semantically very far away. However, it is interesting to note that PGD and FGSM attacks also inflict stronger semantic damage.

We experimented with different threshold values (T_s) for the Wup measure. Figure 4 shows the $QI\text{-}Wup$ values for different threshold values computed for the VGG-19 model. Note that as threshold value decreases it becomes difficult for the attacks to cause severe semantic confusion and the metric becomes close to indiscernible. On the other hand, higher value of T_s will bring out the subtle differences among the attacks.

3.3. QI-Vis

Figure 5 shows the $QI\text{-}Vis$ metric computed for various adversarial attacks over multiple CNN models. Note that we have used visual similarities extracted from the corresponding CNN models as detailed in section 2.4 and a threshold (T_v) value of 0.1. That means, a flipping is considered 'visually fooling' only if the visual similarity between the *pre-attack* label and *post-attack* label is less than 0.1. Note that this threshold is very small compared to T_s. This is because the Wup similarity scores in general are high even for a trivial case of semantically dissimilar labels. For e.g. *Brain Coral* and *Jack fruit* (Figure 6) have a Wup similarity of 0.46. On the other hand, the visual similarities computed from the network (sec. 2.4) are very low. Supplemental document provides the percentile graph computed for GoogLeNet visual similarities. We observe that 95% of the similarities are less than 0.1. Thus, we chose $T_v = 0.1$ for our analysis.

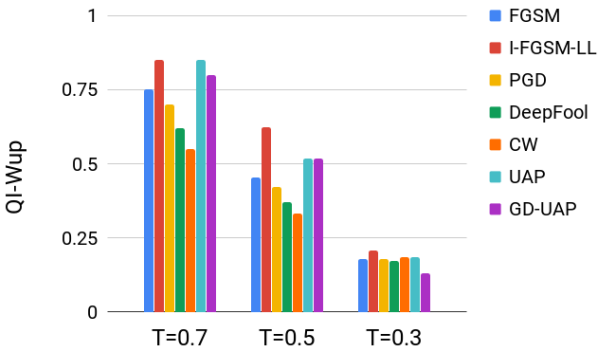


Figure 4. Mean semantic confusion (QI-Wup) caused by various attacks on VGG-19 computed with different threshold values for Wup similarity.

Due to the presence of dominant labels in the *post-attack* labels, the image agnostic attacks (UAP, GD-UAP) cause more visual confusion compared to the image specific attacks. Image specific attacks such I-FGSM-LL also inflict significant visual confusion to the models. Also, note that the visual confusion caused by FGSM and PGD is significantly higher than that by CW and DeepFool.

In summary, iterative attacks remove most evidence for the *pre-attack* label and cause severe demotion. Image agnostic, and targeted attacks can do severe damage with respect to all the three metrics on most of the CNN models. Interestingly, simple FGSM based attacks also cause significant damage (which is non-trivial given only their fooling rate performance) with respect to proposed metrics.

Among the models, ResNet-152 has the least AuC followed by VGG-19. Further, ResNet-152 has smallest

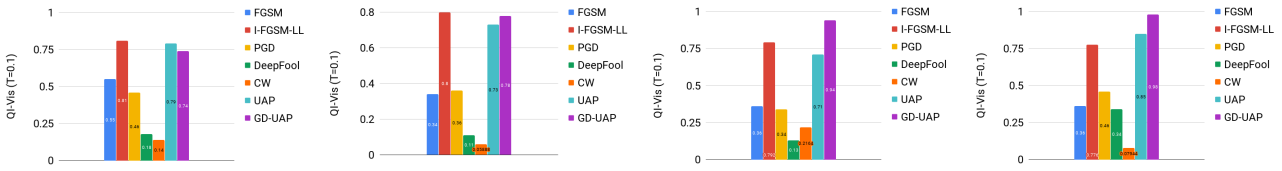


Figure 5. QI-Vis computed for various CNN models for multiple adversarial attacks. Note that the model is mentioned below the corresponding plot.

Figure 6. Similar 'object' patterns can cause severe semantic confusion. First column shows input samples and their *pre-attack* labels and the second column shows the representative images from the predicted *post-attack* labels.

Figure 7. Similar context can also cause models to easily confuse across classes. First column shows input samples and the corresponding *pre-attack* labels and the second column shows the representative images from the predicted *post-attack* labels.

$QI\text{-}Wup$ followed by GoogLeNet. In terms of $QI\text{-}vis$, GoogLeNet incurs least damage followed by VGG-19. Interestingly, ResNet-152 demonstrates the highest $QI\text{-}Vis$.

4. Discussion

In this section, we present some interesting observations that we have come across while analysing the attacks.

4.1. Visual vs Semantic similarity

It is as challenging as it is interesting to discuss 'if the visual similarity correlates to the semantic similarity' or vice versa. In computer vision, it is often taken for granted that they both are correlated [4]. While analysing the confusion of the CNN models to adversarial attacks, we have come across interesting examples about the visual and semantic similarity. Figure 6 shows a pair of example images presented to a trained GoogLeNet under the DeepFool adversarial attack. The first column shows the original images with the ground truth label mentioned in green below them. In the second column we show the representative samples for the *post-attack* labels predicted by the model. Note that the *post-attack* labels are mentioned below them in red.

We intentionally chose the DeepFool attack for it causes the least semantic confusion. However, in this case it is very clear that the *post-attack* labels are semantically far away (Wup similarities are 0.35 and 0.46) from the *pre-attack* labels. However, upon investigating, we found that the visual patterns of the *post-attack* labels are similar to those of the *pre-attack* labels. Their visual similarities given by the model (refer to sec. 2.4) are 0.17 and 0.23 respectively. Please note that they are large compared to the 95% percentile similarity of 0.1. In both the cases, the *post-attack* labels are ranked 2 before the attack. It can be explained with the learning procedure. The only input that the CNN model has received about the object categories is the images belonging to those categories. Therefore, the model tries to learn the discriminative visual patterns for each category from the corresponding samples. Visually similar patterns thus can cause the model to confuse across categories that are semantically far apart, though such cases are not very common. In case of the examples presented in Figure 6, due

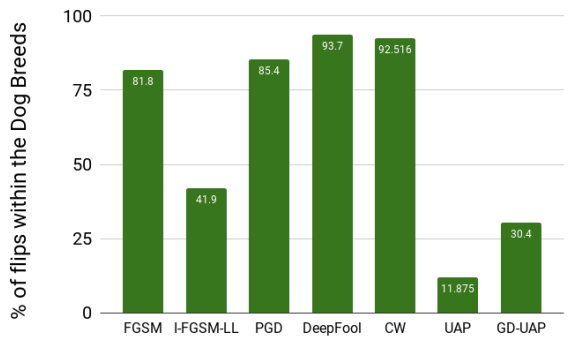


Figure 8. Confusion among fine-grained categories. Figure shows the % of dog samples that are misclassified as another dog category. Note that the analysis is performed on GoogLeNet for 800 validation images from 117 dog categories of ILSVRC dataset.

to the similar visual patterns DeepFool attack could successfully inflict a severe semantic confusion. However, in general we observe the top ranked labels to be both visually and semantically similar.

4.2. Influence of context

As an extension to the previous subsection, we present the confusion caused by visual patterns from the context. Figure 7 shows the images and corresponding ground truth labels on the left. On the right, the representative image for the predicted *post-attack* labels corresponding to them are shown. In the first case, presence of *sky* region, landscapes makes the model to consider the two classes to be visually closer. In case of other samples of *Yurt* category, we observe that the shape of the *Yurt* being very similar to that of the *Alp*. The visual similarity given by the model for the two classes (sec. 2.4) is 0.137. Similarly, in the second example, presence of the bowl(s) in the *eggnog* sample makes the model to confuse between the two classes. Almost all the samples from *eggnog* class have a very close context as that of *mixing-bowl*. In this case, the visual similarity is 0.25.

4.3. Confusion among the fine-grained categories

In this subsection, we analyse the confusion of a CNN model among the fine-grained visual categories under various adversarial attacks. In particular, we chose the 117 dog breed categories and GoogLeNet for this analysis. We consider 800 validation samples belonging to these categories and compute the percentage of intra-dog category confusions. Figure 8 shows the % of samples that are confused among these fine-grained categories, i.e., foolings in which a dog sample is misclassified as another dog category. Note that the confusion caused within the fine-grained categories by the image specific attacks is significantly high. Also, in spite of the existence of 'dominant labels', image agnostic attacks also fool the CNN among the fine-grained categories.

5. Conclusion

In this paper, we challenge the current consensus in the field of using *fooling rate* alone as a metric for evaluating the quality of an adversarial attack. We introduce three additional metrics $FR@K$, $QI\text{-}Wup$, and $QI\text{-}Vis$ that capture three different aspects of the fooling. They helped in bringing out previously unknown strengths and weaknesses of these attacks which would be helpful while deploying the CNN models in real-world environments. To the best of our knowledge, none of the existing works evaluates with metrics other than 'fooling rate'. We list some of the important inferences drawn from our work:

- Our experimental results bring out the usefulness of the new metrics by clearly differentiating attack behaviours. For instance, AuC computed from the $FR@K$ graphs (Tab. 1) reveal that the attacks vary in their ability to reduce the confidence assigned to the *pre-attack* label. PGD demonstrates significantly higher AuC values suggesting its superiority over the others. This is interesting and in line with the observation that PGD attack is the most robust against current adversarial defenses [10], which might be attributed to its ability to reduce the confidence to the *pre-attack* label. Similarly, the observation that Deep Fool (and CW) attack generally swaps the top 2 labels is non-trivial with only 'fooling rate'.
- *QI-Wup* metric (Fig. 3) reveals that some of the attacks (e.g. FGSM based) that are less effective with respect to 'fooling rate', are comparatively more severe on a semantic scale than their counter parts. On the other hand, attacks such as Deep Fool while achieving a high 'fooling rate' inflict least amount of semantic damage. Clearly, when analysing the semantic damage inflicted by the attacks (or incurred by models), fooling rate *can not* serve the need.
- Our experiments reveal (Sec. 4.3, Fig. 3) that some of the strongest adversarial attacks such as PGD achieve significant fooling ($> 85\%$) via confusing the models among visually similar, fine-grained categories which are only $\sim 12\%$ of the total categories. Without this information, higher fooling rates achieved by these attacks may project the classifiers as unsophisticated to the community and more importantly to the policy makers. Similarly, attacks such as CW and DeepFool, in spite of resulting a very high (top-1) fooling rate, cause significantly lesser visual and semantic confusion. On the other hand, relatively simple attacks such as FGSM and FGSM-LL cause higher visual and semantic damage. These aspects throw new light on the attacks and the way they achieve the fooling.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402, 2013.

[2] B. Biggio, G. Fumera, and F. Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(07), 2014.

[3] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[4] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1777–1784, 2011.

[5] A. Ganeshan, B. S. Vivek, and R. V. Babu. FDA: Feature disruptive attack. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances inNeural Information Processing Systems (NIPS)*. 2012.

[9] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[10] A. Madry, A. Makelov, L. Schmidt, T. Dimitris, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[11] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[12] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[14] K. R. Mopuri, A. Ganeshan, and R. V. Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[15] K. R. Mopuri, U. Garg, and R. V. Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[16] K. R. Mopuri, P. Krishna, and R. V. Babu. Ask, acquire, and attack: Data-free uap generation using class impressions. In *European Conference on Computer Vision (ECCV)*, 2018.

[17] K. R. Mopuri, U. Ojha, U. Garg, and R. V. Babu. NAG: Network for adversary generation. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty. Zero-shot knowledge distillation in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[19] NIPS 2018 Challenge. NIPS 2018 : Adversarial vision challenge.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[23] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[24] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, 1994.