FDA: Feature Disruptive Attack

Aditya Ganeshan *
Preferred Networks Inc.,
Tokyo, Japan
aditya@preferred.jp

Vivek B.S.
Video Analytics Lab,
Indian Institute of Science, India
svivek@iisc.ac.in

R. Venkatesh Babu
Video Analytics Lab,
Indian Institute of Science, India
venky@iisc.ac.in

Abstract

Though Deep Neural Networks (DNN) show excellent performance across various computer vision tasks, several works show their vulnerability to adversarial samples, i.e., image samples with imperceptible noise engineered to manipulate the network's prediction. Adversarial sample generation methods range from simple to complex optimization techniques. Majority of these methods generate adversaries through optimization objectives that are tied to the pre-softmax or softmax output of the network. In this work we, (i) show the drawbacks of such attacks, (ii) propose two new evaluation metrics: Old Label New Rank (OLNR) and New Label Old Rank (NLOR) in order to quantify the extent of damage made by an attack, and (iii) propose a new adversarial attack FDA: Feature Disruptive Attack, to address the drawbacks of existing attacks. FDA works by generating image perturbation that disrupt features at each layer of the network and causes deep-features to be highly corrupt. This allows FDA adversaries to severely reduce the performance of deep networks. We experimentally validate that FDA generates stronger adversaries than other state-of-the-art methods for image classification, even in the presence of various defense measures. More importantly, we show that FDA disrupts feature-representation based tasks even without access to the task-specific network or methodology.

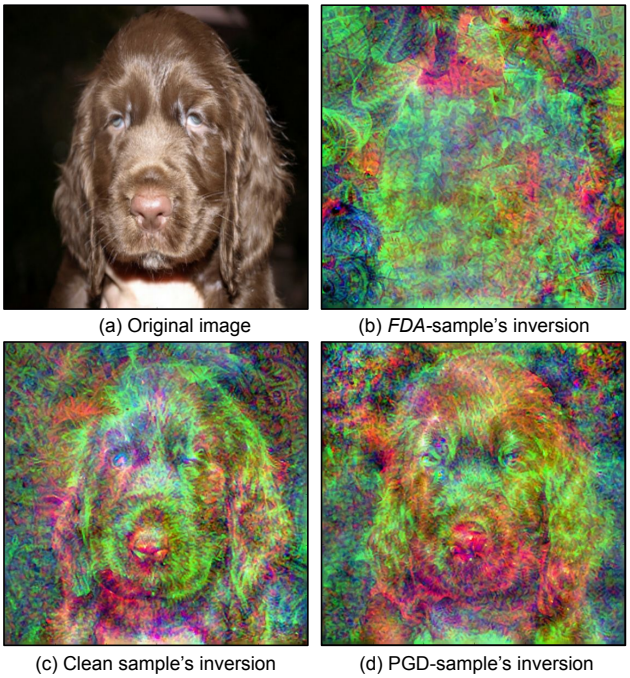

Figure 1. Using feature inversion [34], we visualize the $Mixed_{7b}$ representation of Inception-V3 [48]. The inversion of PGD-attacked sample (d) is remarkably similar to inversion of clean sample (c). In contrast, inversion of FDA-attacked sample (b) completely obfuscates the clean sample's information.

1. Introduction

With the advent of deep-learning based algorithms, remarkable progress has been achieved in various computer vision applications. However, a plethora of existing works [9, 49, 8, 39], have clearly established that Deep Neural Networks (DNNs) are susceptible to *adversarial samples*: input data containing imperceptible noise specifically crafted to manipulate the network's prediction. Further, Szegedy *et al*. [49] showed that adversarial samples transfer across models i.e., adversarial samples generated for one model can adversely affect other unrelated models as well. This transferable nature of adversarial samples further increases the vulnerability of DNNs deployed in real world. As DNNs become more-and-more ubiquitous, especially in decision-critical applications, such as Autonomous Driving [2], the necessity of investigating *adversarial samples* has become paramount.

Majority of existing attacks [49, 13, 33, 16], generate adversarial samples via optimizing objectives that are tied to the pre-softmax or softmax output of the network. The sole objective of these attacks is to generate adversarial samples which are misclassified by the network with very high con-

*Work done as a member of Video Analytics Lab, IISc, India.

fidence. While the classification output is changed, it is unclear what happens to the internal deep representations of the network. Hence, we ask a fundamental question:

Do deep features of adversarial samples retain usable clean sample information?

In this work, we demonstrate that deep features of adversarial samples generated using such attacks retain high-level semantic information of the corresponding clean samples. This is due to the fact that these attacks optimize only pre-softmax or softmax score based objective to generate adversarial samples. We provide evidence for this observation by leveraging feature inversion [35], where, given a feature representation $\psi(x)$, we optimize to construct the approximate inverse $\psi^{-1}(\psi(x))$. Using the ability to visualize deep features, we highlight the retention of clean information in deep features of adversarial samples. The fact that deep features of adversarial samples retain clean sample information has important implications:

- First, such deep features may still be useful for various feature driven tasks such as caption generation [52, 58], and style-transfer [21, 25, 51].
- Secondly, these adversarial samples cause the model to either predict semantically similar class or to retain high (comparatively) probability for the original label, while predicting a very different class. These observations are captured by using the proposed metrics i.e., New Label's Old Rank (NLOR) and Old Label's New Rank (OLNR), and statistics such as fooling rate at kth rank.

These implications are major drawbacks of existing attacks which optimize only pre-softmax or softmax score based objectives. Based on these observations, in this work, we seek adversarial samples that can corrupt deep features and inflict severe damage to feature representations. With this motivation, we introduce *FDA: Feature Disruptive* attack. *FDA* generates perturbation with the aim to cause disruption of features at each layer of the network in a principled manner. This results in corruption of deep features, which in turn degrades the performance of the network. Figure 1 shows feature inversion from deep features of a clean, a PGD [33] attacked, and a *FDA* attacked sample, highlighting the lack of clean sample information after our proposed attack.

Following are the benefits of our proposed attack: (i) *FDA* invariably flips the predicted label to highly unrelated classes, while also successfully removing evidence of the clean sample's predicted label. As we elaborate in section 5, other attacks [49, 30, 13] only achieve one of the above objectives. (ii) Unlike existing attacks, FDA disrupts feature-representation based tasks e.g., caption generation, even without access to the task-specific network or methodology i.e., it is effective in a *gray-box* attack setting. (iii) FDA generates stronger adversaries than other state-of-the-art methods for Image classification. Even in the presence of various recently proposed defense measures (including adversarial training), our proposed attack consistently outperforms other existing attacks.

In summary, the major contributions of this work are:

- We demonstrate the drawbacks of existing attacks.
- We propose two new evaluation metrics i.e., NLOR and OLNR, in order to quantify the extent of damage made by an attack method.
- We introduce a new attack called *FDA* motivated by corrupting features at every layer. We experimentally validate that *FDA* creates stronger white-box adversaries than other attacks on ImageNet dataset for state-of-the-art classifiers, even in the presence of various defense mechanisms.
- Finally, we successfully attack two feature based-tasks, namely caption generation and style transfer where current attack methods either fail or are exhibit weaker attack than *FDA*. A novel "Gray-Box" attack scenario is also presented where *FDA* again exhibits stronger attacking capability.

2. Related Works

Attacks: Following the demonstration by Szegedy *et al*. [49] on the existence of adversarial samples, multiple works [22, 38, 29, 16, 4, 33, 13, 10] have proposed various techniques for generating adversarial samples. Parallelly, works such as [36, 57, 6] have explored the existence of adversarial samples for other tasks.

The works closest to our approach are Zhou *et al*. [59], Sabour *et al*. [43] and Mopuri *et al*. [41]. Zhou *et al*. create black-box transferable adversaries by simultaneously optimizing multiple objectives, including a final-layer cross entropy term. In contrast, we only optimize for our formulation of feature disruption (refer section 4.3). Sabour *et al*. specifically optimize to make a specific layer's feature arbitrarily close to a target image's features. Our objective is significantly different entailing disruption at every layer of a DNNs, without relying on a 'target' image representation. Finally, Mopuri *et al*. provide a complex optimization setup for crafting UAPs whereas our method yields image-specific adversaries. We show that a simple adaptation of their method to craft image-specific adversaries yields poor results (refer supplementary material).

Defenses: Goodfellow *et al*. [22] first showed that including adversarial samples in the training regime increases robustness of DNNs to adversarial attacks. Following this work,

multiple approaches [30, 50, 27, 54, 17, 33] have been proposed for adversarial training, addressing important concerns such as Gradient masking, and label leaking.

Recent works [40, 31, 1, 46, 15, 11], present many alternative to adversarial training. Crucially, works such as [23, 56, 53] propose defense techniques which can be easily implemented for large scale datasets such as ImageNet. While Guo *et al*. [23] propose utilizing input transformation as a defense technique, Xie *et al*. [56] introduce randomized transformations in the input as a defense.

Feature Visualization: Feature inversion has a long history in machine learning [55]. Mahendran *et al*. [34] proposed an optimization method for feature inversion, combining feature reconstruction with regularization objectives. Contrarily, Dosovitskiy *et al*. [18] introduce a neural network for imposing image priors on the reconstruction. Recent works such as [45, 20] have followed the suit. The reader is referred to [19] for a comprehensive survey.

Feature-based Tasks: DNNs have become the preferred feature extractors over hand-engineered local descriptors like SIFT or HOG [5, 7]. Hence, various tasks such as captioning [52, 58], and image-retrieval [12, 24] rely on DNNs for extracting image information. Recently, tasks such as style-transfer have been introduced which rely on deep features as well. While works such as [25, 51] propose a learning based approach, Gatys *et al*. [21] perform an optimization on selected deep features.

We show that previous attacks create adversarial samples which still provide useful information for feature-based tasks. In contrast, *FDA* inflicts severe damage to feature-based tasks without any task-specific information or methodology.

3. Preliminaries

We define a classifier $f : x \in R^m \rightarrow y \in Y^c$, where x is the m dimensional input, and y is the c dimensional score vector containing pre-softmax scores for the c different classes. Applying softmax on the output y gives us the predicted probabilities for the c classes, and $argmax(y)$ is taken as the predicted label for input x. Let y_{GT} represent the ground truth label of sample x. Now, an adversarial sample $\tilde{x}$ can be defined as any input sample such that:

$$argmax(f(\tilde{x})) \neq y_{GT} \;\; \& \;\; d(x, \tilde{x}) < \epsilon, \qquad (1)$$

where $d(x, \tilde{x}) < \epsilon$ acts as an imperceptibility constraint, and is typically considered as a l_2 or l_∞ constraint.

Attacks such as [49, 22, 16, 33], find adversarial samples by different optimization methods, but with the same optimization objective: maximizing the cross-entropy loss $J(f(\tilde{x}), y_{GT})$ for the adversarial sample $\tilde{x}$. Fast Gradient Sign Method (FGSM) [49] performs a single step optimization, yielding an adversary :

$$FGSM(x) = \tilde{x} = x + \epsilon \cdot sign\big(\nabla_x(J(f(x), y_{GT}))\big) \quad (2)$$

On the other hand, PGD [33] and I-FGSM [22] performs a multi-step signed-gradient ascent on this objective. Works such as [16, 4], further integrate Momentum and ADAM optimizer for maximizing the objective.

Kurakin *et al*. [30] discovered the phenomena of label leaking and use predicted label instead of y_{GT}. This yields a class of attacks which can be called most-likely attacks, where the loss objective is changed to $J(f(\tilde{x}), y_{ML})$ (where y_{ML} represents the class with the maximum predicted probability).

Works such as [27, 50] note that above methods yield adversarial samples which are weak, in the sense of being misclassified into a very similar class (for e.g., a hound misclassified as a terrier). They posit that targeted attacks are more meaningful, and utilize least likely attacks, proposing minimization of Loss objective $J(f(\tilde{x}), y_{LL})$ (where y_{LL} represents the class with the least predicted probability). We denote the most-likely and the least-likely variant of any attack by the suffix ML and LL.

Carlini *et al*. [13] propose multiple different objectives and optimization methods for generating adversaries. Among the proposed objectives, they infer that the strongest objective is as follows:

$$Objective(\tilde{x}) = (max_{i \neq ML}(f(\tilde{x})_i) - f(\tilde{x})_{ML})^+, \quad (3)$$

where $(e)^+$ is short-hand for $max(e, 0)$. For a l_∞ distance metric adversary, this objective can be integrated with PGD optimization to yield PGD-CW. The notation introduced in this section is followed throughout the paper.

Feature inversion: Feature inversion can be summarized as the problem of finding the sample whose representation is the closest match to a given representation [55]. We use the approach proposed by Mahendran *et al*. [34]. Additionally, to improve the inversion, we use Laplacian pyramid gradient normalization. We provide additional information in the supplementary.

4. Feature Disruptive Attack

4.1. Drawbacks of existing attacks

In this section, we provide qualitative evidence to show that deep features corresponding to adversarial samples generated by existing attacks (i.e., attacks that optimize objectives tied to the softmax or pre-softmax layer of the network), retain high level semantic information of its corresponding clean sample. We use feature inversion to provide evidence for this observation.

Figure 2 shows the feature inversion for different layers of VGG-16 [44] architecture trained on ImageNet dataset,

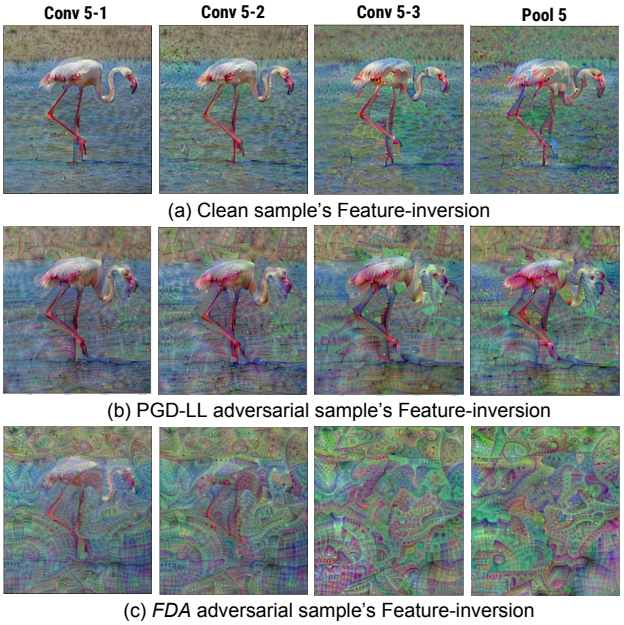

Figure 2. Feature Inversion: Layer-by-layer Feature Inversion [34] of clean, PGD-LL-adversarial and *FDA*-adversarial sample. Note the significant removal of clean sample information in later layers of *FDA*-adversarial sample.

for the clean and its corresponding adversarial sample. From Fig. 2, it can be observed that the inversion of adversarial features of PGD-LL sample [33] is remarkably similar to the inversion of features of clean sample. Further, in section 5.1, we statistically show the similarity between intermediate feature representations of clean and its corresponding adversarial samples generated using different existing attack methods. Finally, in section 5.2 we show that as a consequence of retaining clean sample information, these adversarial samples cause the model to either predict semantically similar class or to retain high (comparatively) probability for the original label, while predicting a very different class. These observations are captured by using the proposed metrics i.e., New Label Old Rank (NLOR) and Old Label New Rank (OLNR), and statistics such as fooling rate at k-th rank.

4.2. Proposed evaluation metrics

An attack's strength is typically measured in terms of *fooling rate* [37], which measures the percentage (%) of images for which the predicted label was changed due to the attack. However, only looking at fooling rate does not present the full picture of the attack. On one hand, attacks such as PGD-ML may result in flipping of label into a semantically similar class, and on the other hand, attacks such as PGD-LL may flip the label to a very different class, while still retaining high (comparatively) probability for the original label. These drawbacks are not captured in the existing evaluation metric i.e., *Fooling rate*.

Hence, we propose two new evaluation metrics, *New Label Old Rank (NLOR)* and *Old Label New Rank (OLNR)*. For a given input image, the softmax output of a C-way classifier represents the confidence for each of the C classes. We sort these class confidences in descending order (from rank 1 to C). Consider the prediction of the network before the attack as the *old label* and after the attack as the *new label*. Post attack, the rank of the *old label* will change from 1 to say 'p'. This new rank 'p' of the *old label* is defined as OLNR (Old Label's New Rank). Further, post attack, the rank of the *new label* would have changed from say 'q' to 1. This old rank 'q' of the *new label* is defined as NLOR (New Label's Old Rank). Hence, a stronger attack should flip to a label which had a high old rank (which will yield high *NLOR*), and also reduce probability for the clean prediction (which will yield a high *OLNR*). These metrics are computed for all the mis-classified images and the mean value is reported.

4.3. Proposed attack

We now present *Feature Disruptive Attack (FDA)*, our proposed attack formulation explicitly designed to generate perturbation that contaminate and corrupt the internal representations of a DNN. The aim of the proposed attack is to generate image specific perturbation which, when added to the image should not only flip the label but also disrupt its inner feature representations at each layer of the DNN. We first note that activations supporting the current prediction have to be lowered, whereas activations which do not support the current prediction have to be strengthened and increased. This can lead to feature representations which, while hiding the true information, contains high activations for features not present in the image. Hence, for a given i^{th} layer l_i, our layer objective $\mathcal{L}$, which we want to increase is given by:

$$\begin{aligned}\mathcal{L}(l_i) = D\left(\{l_i(\tilde{x})_{N_j} | N_j \notin S_i\}\right) \\ - D\left(\{l_i(\tilde{x})_{N_j} | N_j \in S_i\}\right),\end{aligned} \tag{4}$$

where $l_i(\tilde{x})_{N_j}$ represents the N_jth value of $l_i(\tilde{x})$, S_i represents the set of activations which support the current prediction, and D is a monotonically increasing function of activations $l(\tilde{x})_{N_j}$ (on the partially ordered set $R^{|S_i|}$). We define D as the l_2-norm of inputs $l_i(\tilde{x})$.

Finding the set S_i is non-trivial. While all high activations may not support the current prediction, in practice, we find it to be usable approximation. We define the support set S_i as:

$$S_i = \{N_j \quad | \quad l_i(x)_{N_j} > C\}, \tag{5}$$

where C is a measure of central tendency. We try various choices of C including $median(l_i(x))$ and $inter\text{-}quartile\text{-}mean(l_i(x))$. Overall, we find $spatial\text{-}mean(l_i(x)) =$

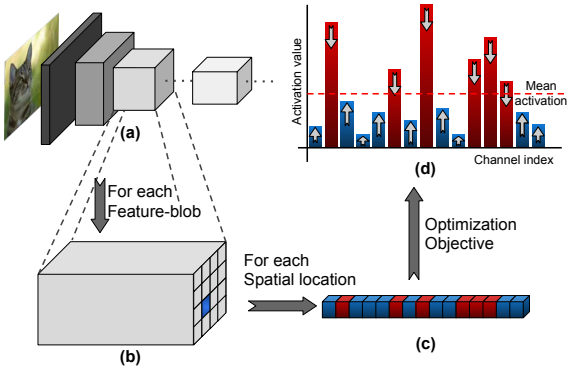

Figure 3. Overview Image: From network (a), for each selected feature blob (b) we perform the optimization (d) as explained in equation 6. (c) shows a spatial feature, where the support set S_i is colored red, and the remaining is blue.

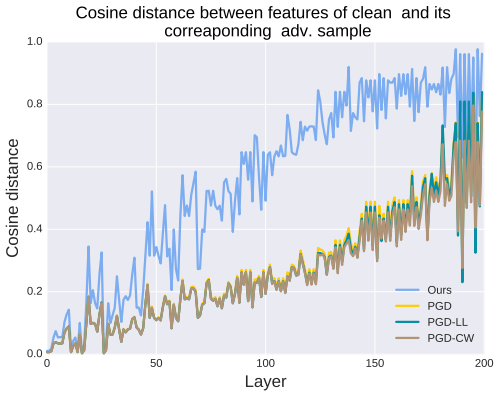

Figure 4. Cosine distance between features of clean image and its corresponding adversarial sample, at different layer of P-NasNet [32] architecture.

$C(h, w)$ (mean across channels) to be the most effective formulation. Finally, combining Eq. (4) and (5), our layer objective $\mathcal{L}$ becomes:

$$\mathcal{L}(l_i) = \log\left(D\left(\{l_i(\tilde{x})_{(h,w,c)} | l_i(x)_{(h,w,c)} < C_i(h,w)\}\right)\right) - \log\left(D\left(\{l_i(\tilde{x})_{(h,w,c)} | l_i(x)_{(h,w,c)} > C_i(h,w)\}\right)\right), \tag{6}$$

We perform this optimization at each non-linearity in the network, and combine the per-layer objectives as follows:

$$Objective = -\sum_{i=1}^{K} \mathcal{L}(l_i), \quad \text{such that} \quad \|\tilde{x} - x\|_\infty < \epsilon, \tag{7}$$

Figure 3 provides a visual overview of the proposed method. In supplementary document we provide results for ablation study of the proposed attack i.e., different formulation of C such as median, Inter-Quartile-mean etc.

5. Experiments

In this section, we first present statistical analysis of the features corresponding to adversarial samples generated using existing attacks and the proposed attack. Further, we show the effectiveness of the proposed attack on (i) image recognition in white-box (Sec. 5.2) and black-box settings (shown in supplementary document), (ii) Feature-representation based tasks (Sec. 5.4) i.e., caption generation and style-transfer. We define optimization budget of an attack by the tuple $(\epsilon, nb_{iter}, \epsilon_{iter})$, where ϵ is the L_∞ norm limit on the perturbation added to the image, nb_{iter} defines the number of optimization iterations used by the attack method, and ϵ_{iter} is the increment in the L_∞ norm limit of the perturbation at each iteration.

5.1. Statistical analysis of adversarial features

In this section, we present the analysis which fundamentally motivates our attack formulation. We present various experiments, which posit that attack formulations tied to pre-softmax based objectives retain clean sample information in deep features, whereas *FDA* is effective at removing them. For all the following experiments, all attacks have been given the same optimization budget ($\epsilon = 8, nb_{iter} = 10, \epsilon_{iter} = 1$). Reported numbers have been averaged over 1000 image samples.

First, we measure the similarity between intermediate feature representations of the clean and its corresponding adversarial samples generated using different attack methods. Figure 4, shows average cosine distance between intermediate feature representations of the clean and its corresponding adversarial samples, for various attack methods on PNasNet [32] architecture. From Fig. 4 it can be observed that for the proposed attack, feature dis-similarity is much higher than to that of the other attacks. The significant difference in cosine distance implies that contamination of intermediate feature is much higher for the proposed attack. We observe similar trend in other models at different optimization budgets $(\epsilon, nb_{iter}, \epsilon_{iter})$ as well (refer supplementary).

Now, we measure the similarity between features of clean and adversarial samples at the *pre-logits* layer (i.e., input to the classification layer) of the network. Apart from cosine distance, we also measure the Normalized Rank Transformation (NRT) distance. NRT-distance represents the average shift in the rank of the k^{th} ordered statistic $\forall\ k$. Primary benefit of NRT-distance measure is its robustness to outliers.

Table 1, tabulates the result for *pre-logit* output for multiple architectures. It can be observed that our proposed attack shows superiority to other methods. Although the *pre-logits* representations from other attacks seem to be cor-

Table 1. Metrics for measuring the dissimilarity between adversarial *pre-logits* and clean *pre-logits* on different networks. Our method *FDA* exhibits stronger dissimilarity.

		PGD-ML	PGD-CW	PGD-LL	Ours
Res-152	Cosine Dist.	0.49	0.37	0.60	**0.81**
	NRT Dist.	15.00	13.56	16.29	**19.17**
Inc-V3	Cosine Dist.	0.51	0.41	0.49	**0.55**
	NRT Dist.	16.11	14.97	17.38	**19.01**

rupted, in section 5 we show that these *pre-logits* representation provide useful information for feature-based tasks.

5.2. Attack on Image Recognition

ImageNet [42] is one of the most frequently used large-scale dataset for evaluating adversarial attacks. We evaluate our proposed attack on five DNN architectures trained on ImageNet dataset, including *state-of-the-art* PNASNet [32] architecture. We compare *FDA* to the strongest white-box optimization method (PGD), with different optimization objective, resulting in the following set of competing attacks: PGD-ML, PGD-LL, and PGD-CW.

We present our evaluation on the ImageNet-compatible dataset introduced in NIPS 2017 challenge (contains $5,000$ images). To provide a comprehensive analysis of our proposed attack, we present results with different *optimization budgets*. Note that attacks are compared only when they have the same optimization budget.

Table 2: top section presents the evaluation of multiple attack formulations across different DNN architectures with the optimization budget ($\epsilon = 4, nb_{iter} = 5, \epsilon_{iter} = 1$) in white-box setting. A crucial inference is the partial success of other attacks in terms of *NLOR* and *OLNR*. They either achieve significant *NLOR* or *OLNR*. This is due to the singular objective of either lowering the maximal probability, or increasing probability of the least-likely class. Table 2 also highlights the significant drop in performance of other attack for deeper networks (PNASNet [32] and Inception-ResNet [47]) due to vanishing gradients.

In Figure 5, we present *Generalizable Fooling Rate* [41] with respect to Top-k accuracy as a function of k. The significantly higher *Generalizable Fooling Rate* at high k values further establishes the superiority of our proposed attack on networks trained on ImageNet dataset.

5.3. Evaluation against Defense proposals

Now, we present evaluation against defenses mechanisms which have been scaled to ImageNet (experiments on defense mechanisms in smaller dataset (CIFAR-10) [28] are provided in the supplementary document).

Adversarial Training: We test our proposed attack against three adversarial training regimes, namely: Simple (adv) [30], Ensemble ($ens3$) [50] and Adversarial-logit-pairing (alp) [27] based adversarial training. We set the optimization budget of ($\epsilon = 8, nb_{iter} = 5, \epsilon_{iter} = 2$) for all the attacks on adv and $ens3$ models. Table 2: bottom section presents the results of our evaluation. Further, to show effectiveness at different optimization budgets, alp models are tested with different optimization budget, as show in Table 3.

Defense Mechanisms: We also test our model against defense mechanisms proposed by Guo *et al*. [23] and Xie *et al*. [56]. Table 4 shows fooling rate achieved in Inception-ResNet V2 [47], under the presence of various defense mechanisms. The above results confirm the superiority of our proposed attack for white-box attack.

5.4. Attacking Feature-Representation based tasks

5.4.1 Caption Generation

Most DNNs involved in real-world applications utilize transfer learning to alleviate problems such as data scarcity and efficiency. Furthermore, due to the easy accessibility of trained models on ImageNet dataset, such models have become the preferred starting point for training task-specific models. This presents an interesting scenario, where the attacker may have the knowledge of which model was fine-tuned for the given task, but may not have access to the fine-tuned model.

Due to the partial availability of information, such a scenario in essence acts as a "Gray-Box" setup. We hypothesize that in such a scenario, feature-corruption based attacks should be more effective than softmax or pre-softmax based attacks. To test this hypothesis, we attack the caption-generator "Show-and-Tell" (SAT) [52], which utilizes a ImageNet trained Inception-V3 (IncV3) model as the starting point, using adversaries generated from only the ImageNet-trained IncV3 network. Note that the IncV3 in SAT has been fine-tuned for 2 Million steps (albeit with a smaller learning rate).

Table 5 presents the effect of adversarial attacks on caption generation. We attack "Show-and-Tell" [52]. Similar performance can be expected on advanced models such as [26, 58]. We clearly see the effectiveness of *FDA* in such a "Gray-Box" scenario, validating the presented hypothesis. Additionally, we note the content-specific metrics such as SPICE [3], are far more degraded. This is due to the fact that other attacks may change the features only to support a similar yet different object class, whereas *FDA* aims to completely removes the evidence of the clean sample.

We further show results for attacking SAT in a "White-box" setup in Table 6. We compare against Hongge *et al*. [14] as well, an attack specifically formulated for caption generation. While the prime benefit of Hongge *et al*. is the ability to perform targeted attack, we observe that we are comparable to Hongge *et al*. in the untargeted scenario.

Table 2. Evaluation of various attacks on networks trained on ImageNet dataset, in white-box setting. Top: Comparison on normally trained architectures, with the optimization budget (refer section 5) of ($\epsilon = 4, nb_{iter} = 5, \epsilon_{iter} = 1$). Bottom: Comparison on adversarially trained models (adv & ens), with the budget ($\epsilon = 8, nb_{iter} = 5, \epsilon_{iter} = 2$). The salient feature of our attack is high performance on all metrics at the same time.

Metrics	**Fooling Rate**				**NLOR**				**OLNR**			
	PGD-ML	PGD-CW	PGD-LL	Ours	PGD-ML	PGD-CW	PGD-LL	Ours	PGD-ML	PGD-CW	PGD-LL	Ours
VGG-16	99.90	**99.90**	93.80	97.80	57.26	6.17	**539.92**	433.33	308.34	29.19	217.98	**455.26**
ResNet-152	99.50	**99.60**	88.15	97.69	20.62	5.12	**593.64**	412.52	247.22	21.84	89.58	**380.04**
Inc-V3	99.20	99.10	89.06	**99.80**	61.73	21.95	**599.49**	549.57	524.65	63.86	92.45	**669.31**
IncRes-V2	94.18	94.58	74.30	**99.60**	75.43	44.51	314.20	**492.95**	314.14	44.46	67.02	**487.76**
PNasNet-Large	92.60	92.40	81.40	**99.00**	123.93	59.44	319.18	**473.54**	335.63	70.67	118.73	**512.21**
Inc-V3$_{adv}$	97.89	97.69	80.62	**99.70**	68.03	34.56	346.59	**545.89**	281.75	39.08	77.80	**629.93**
Inc-V3$_{ens3}$	98.69	97.49	88.76	**100.00**	114.96	68.76	450.66	**533.49**	386.16	106.58	142.65	**634.55**
IncRes-V2$_{adv}$	91.27	89.66	61.65	**99.70**	81.80	39.68	284.36	**504.51**	234.66	33.20	67.27	**571.46**
IncRes-V2$_{ens3}$	98.69	97.49	88.76	**100.00**	114.96	68.76	450.66	**533.49**	386.16	106.58	142.65	**634.55**

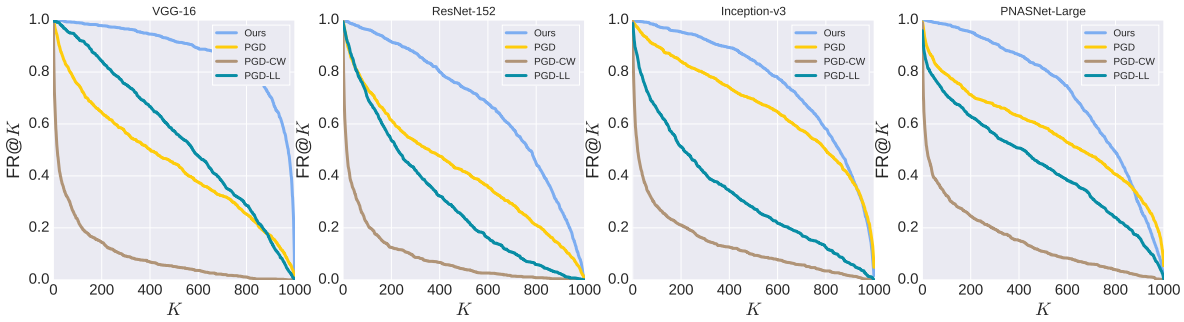

Figure 5. Fooling rate at K-th rank for various attacks in white-box setting with the optimization budget (refer section 5) ($\epsilon = 8, nb_{iter} = 10, \epsilon_{iter} = 1$). Attacks are performed on networks trained on ImageNet dataset. Column-1: VGG-16, Column-2: ResNet-152, Column-3: Inception-V3 and Column-4: PNASNet-Large.

5.4.2 Style-transfer

From its introduction in [21], Style-transfer has been a highly popular application of DNNs, specially in arts. However, to the best of our knowledge, adversarial attacks on Style-transfer have not yet been studied.

Earlier method by Gatys *et al.* [21] proposed an optimization based approach, which utilizes gradients from trained networks to create an image which retains "content" from one image, and "style" from another. We first show that adversaries generated from other methods (PGD etc.) completely retain the structural content of the clean

Table 3. Evaluation on ALP [27]-adversarially trained model, with different optimization budget.

$\epsilon = 8, nb_{iter} = 5, \epsilon_{iter} = 2$				
	PGD-ML	PGD-CW	PGD-LL	Ours
Fooling Rate	85.04	**87.15**	51.10	80.02
NLOR	22.28	10.83	20.60	**119.41**
OLNR	77.55	11.14	14.90	**81.73**
$\epsilon = 16, nb_{iter} = 10, \epsilon_{iter} = 2$				
	PGD-ML	PGD-CW	PGD-LL	Ours
Fooling Rate	96.99	**98.29**	64.56	94.28
NLOR	41.51	12.26	77.40	**259.78**
OLNR	**302.03**	14.97	25.66	241.43

Table 4. Evaluation of various attacks in the presence of input transformation based defense measures with budget ($\epsilon = 16, nb_{iter} = 10, \epsilon_{iter} = 2$). While achieving higher fooling rate, we also achieve higher *NLOR* and *OLNR* (refer supplementary).

Defenses	**Fooling Rate**			
	PGD-ML	PGD-CW	PGD-LL	Ours
Gaussian Filter	81.93	36.95	68.57	**92.87**
Median Filter	50.40	23.19	38.45	**70.88**
Bilateral Filter	54.52	19.18	41.47	**70.18**
Bit Quant.	73.90	40.86	62.05	**91.77**
JPEG Comp.	79.82	31.83	66.67	**96.18**
TV Min.	38.96	17.67	27.81	**55.72**
Quilting	38.35	24.10	30.82	**56.63**
Randomize [56]	81.93	42.87	68.17	**98.19**

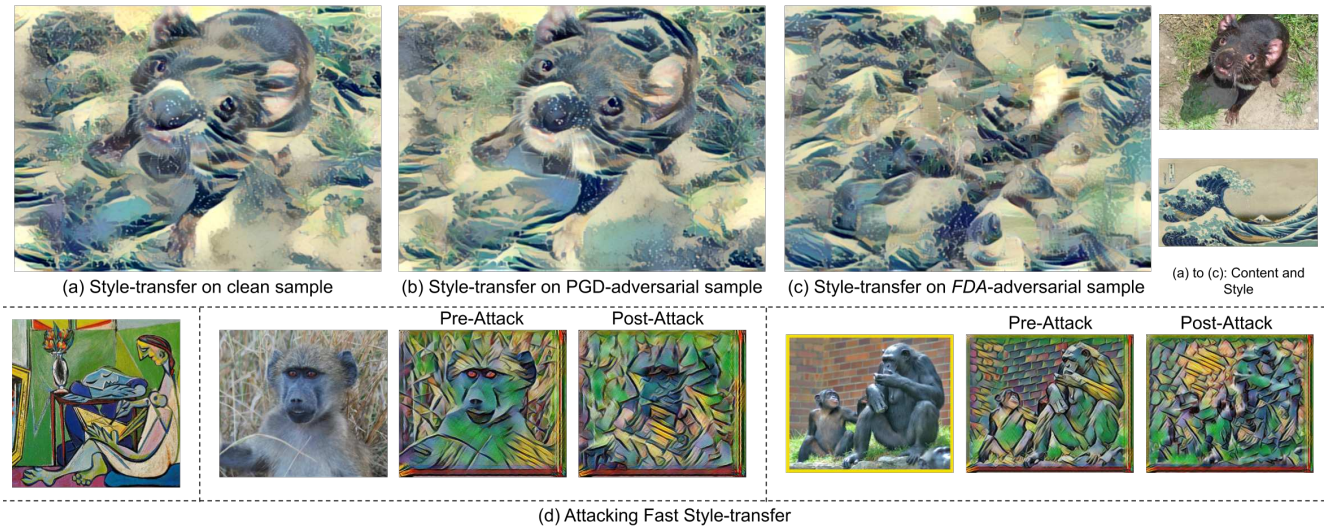

Figure 6. Attacking Style Transfer. Top: PGD adversaries provide clean sample information sufficient for effective style transfer, whereas *FDA* adversaries do not. (d): Generating adversaries for Johnson *et al*. [25] using *FDA*, where PGD formulation fails. Leftmost image presents the style, followed by a sequence of clean image, style-transfer before and after adversarial attack by FDA.

sample, allowing them to be used for style transfer without any loss in quality. In contrast, as *FDA* adversaries corrupts the clean information. Hence, apart from causing misclassification, *FDA* adversaries also severely damage style-transfer. Figure 6: top shows example of style-transfer on clean, PGD-adversarial, and *FDA* adversarial sample. More importantly, *FDA* disrupts style-transfer without utilizing any task-specific knowledge or methodology.

Table 5. Attacking "Show-and-Tell"(SAT) [52] in a "Gray-box" setup with budget ($\epsilon = 8, nb_{iter} = 10, \epsilon_{iter} = 1$). The rightmost column tabulates the metrics when complete white noise is given as input. *FDA* Adversaries generated from Inception-V3 are highly effective for disrupting SAT.

Metrics	No Attack	PGD-ML	PGD-LL	MI-FGSM	Ours	Noise
CIDEr	103.21	47.95	47.13	49.23	**4.90**	2.84
Blue-1	71.61	57.04	55.68	57.18	**39.80**	37.60
Rough$_L$	53.61	42.15	41.24	42.65	**30.70**	29.30
METEOR	25.58	17.507	16.78	17.34	**10.02**	7.84
SPICE	18.07	9.60	9.45	10.02	**2.04**	1.00

Table 6. Attacking (SAT) [52] in a "White-box" setup with budget ($\epsilon = 8, nb_{iter} = 10, \epsilon_{iter} = 1$). *FDA* is at-par with task-specific attack [14]

Metrics	No Attack	PGD-ML	MI-FGSM	[14]	Ours	Noise
CIDEr	94.90	31.70	31.21	10.80	**4.14**	2.84
Blue-1	69.13	51.64	51.36	**38.95**	39.80	37.60
Rough$_L$	51.68	38.20	38.20	**28.19**	31.00	29.30
METEOR	24.29	14.55	14.60	9.75	**9.30**	7.84
SPICE	17.08	7.30	7.00	3.38	**1.68**	0.99

In [25], Johnson *et al*. introduced a novel approach where a network is trained to perform style-transfer in a single forward pass. In such a setup it is infeasible to mount an attack with PGD-like adversaries as there is no final layer to derive loss-gradients from. In contrast, with the white-box access to the parameters of these networks, *FDA* adversaries can be generated to disrupt style-transfer, without any change in its formulation. Figure 6: bottom shows qualitative examples of disruption caused due to *FDA* adversaries in the model proposed by Johnson *et al*. Style-transfer has been applied to videos as well. We have provided qualitative results in the supplementary to show that *FDA* remains highly effective in disrupting stylized videos as well.

6. Conclusion

In this work, we establish the retention of clean sample information in adversarial samples generated by attacks that optimizes objective tied to softmax or pre-softmax layer of the network. This is found to be true even when these samples are misclassified with high confidence. Further, we highlight the weakness of such attacks using the proposed evaluation metrics: OLNR and NLOR. We then propose *FDA*, an adversarial attack which corrupts the features at each layer of the network. We experimentally validate that *FDA* generates one of the strongest white-box adversaries. Additionally, we show that feature of *FDA* adversarial samples do not allow extraction of useful information for feature-based tasks such as style-transfer, and caption-generation as well.

References

[1] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha. Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 89–96, June 2017.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *The European Conference on Computer Vision (ECCV)*, 2016.

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, July 2018.

[5] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[6] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. *arXiv preprint arXiv:1701:04143*, 2017.

[7] Alexander C Berg, Tamara L Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio 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.

[9] Battista Biggio, Giorgio Fumera, and Fabio Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(07):1460002, 2014.

[10] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018.

[11] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Deep visual-semantic quantization for efficient image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2017.

[13] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

[14] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[15] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. *CoRR*, abs/1708.05493, 2017.

[18] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] M. Du, N. Liu, and X. Hu. Techniques for Interpretable Machine Learning. *arXiv preprint arXiv: 1808.00033*, July 2018.

[20] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD, pages 1358–1367, 2018.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, June 2016.

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

[23] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018.

[24] Tuan Hoang, Thanh-Toan Do, Dang-Khoa Le Tan, and Ngai-Man Cheung. Selective deep convolutional features for image retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1600–1608, 2017.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[26] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018.

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[30] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[31] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[32] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[34] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[35] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision (IJCV)*, 120(3):233–255, 2016.

[36] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *International Conference on Computer Vision (ICCV)*, 2017.

[37] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[40] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Konda Reddy Mopuri, Aditya Ganeshan, and R. Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[43] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. *arXiv preprint arXiv:1511.05122*, 2015.

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

[45] A. Singh and A. Namboodiri. Laplacian pyramids for deep feature inversion. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 286–290, Nov 2015.

[46] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[47] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

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

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[52] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, April 2017.

[53] B. S. Vivek, Arya Baburaj, and R. Venkatesh Babu. Regularizer to mitigate gradient masking effect during single-step adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[54] B. S. Vivek, Konda Reddy Mopuri, and R. Venkatesh Babu. Gray-box adversarial training. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[55] R. J. Willaims. Inverting a connectionist network mapping by backpropagation of error. *Proc. of 8th Annual Conference of the Cognitive Science Society*, pages 859–865, 1986.

[56] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018.

[57] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision (ICCV)*, 2017.

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2048–2057, 07–09 Jul 2015.

[59] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *The European Conference on Computer Vision (ECCV)*, September 2018.