INDIRECT SUPERVISION TO MITIGATE PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vulnerability of state-of-the-art computer vision models to image perturbations has drawn considerable attention recently. Often these perturbations are imperceptible to humans because they target the perception of deep neural networks (DNNs) employed in the corresponding computer vision task. Recent studies have revealed that DNNs, which are unable to handle targeted perturbation often fail to handle untargeted perturbations as well, such as Gaussian noise. Various techniques, ranging from classical preprocessing to current supervised and self-supervised deep discriminative and generative model based approaches, have been explored in past to mitigate both these types of perturbations. However, a common challenge with most of these is that they try to solve the problem from a quality enhancement point of view, which is primarily driven by human perception. In addition, the supervised models require a large volume of gold standard unperturbed data, whereas others fail to take into account the feedback of the targeted downstream DNN. We propose to model this problem in indirect supervision framework, where we assume that the gold standard data is missing, however, a variable dependent on it is available and the dependency of the observed variable is stated by the considered downstream DNN. The proposed method maintains the advantages of supervised models while relaxing the requirement of gold standard unperturbed data. To prove its utility, we conduct several experiments with various network architectures for downstream tasks of classification and medical image segmentation. We used MNIST, CIFAR-10-C and ISIC skin lesion dataset in our experiments. In all the experiments, a considerable restoration in the performance of the considered downstream model is observed along with the reduction in image perturbations.

1 INTRODUCTION

Current deep learning models are able to match human level accuracy in various computer vision tasks, however, their performance degrades to subhuman levels in presence of small image perturbation which are often imperceivable to human observers. A lot of research has happened in this direction to find techniques, called as adversarial attacks, for creating targeted perturbations and develop defences against such attacks (Szegedy et al. (2013); Biggio et al. (2013); Goodfellow et al. (2014); Carlini & Wagner (2017); Madry et al. (2017); Sinha et al. (2017)). Recently, the vulnerability of deep models against common untargeted attacks, such as Gaussian noise corruption, has been exposed while establishing its connection with targeted adversarial perturbations (Gilmer et al. (2019)). Thus, rather than having a differentiation between targeted and untargeted image perturbations, a strategy for reducing both to improve the performance of the models is desirable.

Our motivation comes from the behaviour of classical approaches of prepossessing, such as denoising (Xie et al. (2012)), deblurring (Eigen et al. (2013)), rescaling (Braun & Fairchild (1999)) etc. These approaches are driven by quality improvements as perceived by a human, which may not be same as what is seen by a deep learning model. We argue that if the quality improvements of perturbed images are driven by the perceptual quality of the downstream model, it can restore the model performance irrespective of the nature of the perturbation. In past attempts have been made to improve the perception of the models either during training (Zhang et al. (2017a)) or post deployment by re-training (Meng & Chen (2017)) using a large volume of perturbed data, however, their

generalizability is always a concern. Further, the re-training becomes infeasible in many scenarios after the deployment of a trained model.

An alternate to classical image enhancement approaches, used in general for preprocessing, is the combination of analytical methods with deep learning techniques, for example, convolutional neural network (CNN) based denoiser prior (Zhang et al. (2017c)). These techniques have been surpassed by denoising CNNs used under the plain discriminative settings (Zhang et al. (2017b)), which use a large number of paired perturbed and clean samples to learn a mapping between them. In real world scenarios, availability of clean samples corresponding to the perturbed samples often becomes bottleneck due various limitations, such as hardware induced noise. A potential alternate to these supervised approaches are the self-supervised approaches (Krull et al. (2019)), however their scope has yet been limited to restricted types of perturbations, such as unstructured and uncorrelated noise.

To alleviate the aforementioned problems, we propose ways to relax following constraints of supervised discriminative image enhancement models.

1. Instead of asking for clean samples corresponding to the perturbed samples, we look for information which remains invariant to the perturbation and is comparatively easy to obtain, for example, class labels.
2. Instead of improving the quality as perceived by humans, we propose to employ a reformer model (autoencoder) to improve the image quality as perceived by the targeted downstream model.
3. We reduce the requirement of a large volume of training data by pretraining the reformer model on the clean data used by the targeted downstream model for the corresponding downstream task. We show that post pretraining a small set of perturbed samples with information invariant to perturbation become sufficient to fine-tune the reformer model for the desired enhancement.

These relaxations also highlight our contributions under which our proposed approach becomes an indirect supervision approach. It enjoys the benefits of their supervised counterparts while reducing the requirements of large volume of ground truth data. In order to demonstrate the efficiency of the proposed approach we perform experiments on a variety of data for a different downstream task including classification of adversarial MNIST digit images and CIFAR-10-C (Hendrycks & Dietterich (2019)) images which are the corrupted version of standard CIFAR-10 images, and segmentation of skin lesion images under adversarial attacks.

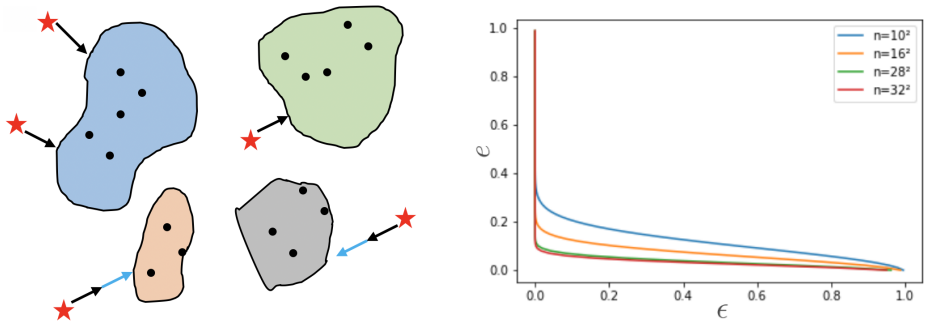
2 Proposed Method

We are interested in a general case where no assumption is made on the characteristics of the observed perturbation and the gold standard data (clean/unperturbed) may or may not be available. We assume that a variable dependent on gold standard data is observed and a perturbation is considerable only if it affects the value of the observed dependent variable. The effect of perturbation on the dependent variable is used as indirect supervision to solve the problem in hand. Although approaches have been reported in past which implicitly use indirect supervision in form of noisy annotations (Natarajan et al. (2013)), partial observations (Cour et al. (2011); Raghunathan et al. (2016)) or external world feedback (Berant et al. (2013)), however, the indirect supervision has been formally considered under the leaning premises very recently in (Wang et al. (2019)). We follow similar notations as (Wang et al. (2019)) with some minor modification to describe the considered problem and proposed approach.

Preliminaries: Let us assume X is a variable which takes value in space $\mathcal{X}$, where $\mathcal{X} = \{x_1, x_2, x_3...\}$ is a finite set of observed (perturbed) images with corresponding set of gold standard clean images as $\mathcal{Y} = \{y_1, y_2, y_3...\}$. Let us denote another variable Y which takes value in space $\mathcal{Y}$ and is unobserved, however, we have an observed variable $Z \in \mathcal{Z} = \{z_1, z_2, z_3...\}$ which is dependent on Y. This dependency is modelled using the mapping $g_\phi : \mathcal{Y} \mapsto \mathcal{Z}$ which is parametrized by ϕ. The sets of perturbed and clean images obey the following relationship

$$x_i = \eta(y_i) + \epsilon \tag{1}$$

where η and ϵ respectively induce the effects of correlated and uncorrelated perturbations. Estimation of Y from X is considered as an inverse problem and has seen several analytical and more



(a) (b)

Figure 1: **(a)** Pictorial representation of perturbation reduction done by f_θ in case g_ϕ is a classifier. Regions with different colors depict the span of different classes with samples (y_i) shown as black dots. Red stars show the perturbed samples (x_i) which are pushed towards the class boundaries by pretrained f_θ, represented by black arrows. A further push or reduction in perturbation is obtained after fine-tuning using indirect sparse supervision, represented by blue arrows. **(b)** Effect of perturbation with varying ϵ for different values n (input dimension). Note that $n = 32^2$ is the dimension of CIFAR-10 samples, for which $e = \exp\left(-\frac{n-1}{2}\epsilon^2\right)$ vanishes even for very small values of ϵ.

recently learning based solutions. In particular, deep learning has emerged as an attractive choice due to the large capacity of (convolutional) neural networks and fast computations. Initial attempts were made to combine the rich analytical knowledge in this domain with the power deep neural networks (DNNs). The inverse problem is divided into sub-problems, connected with a fidelity term and a regularization term, where DNNs were used to solve the regularization sub-problem. However, it is now slowly becoming a well-accepted fact that DNNs employed in plain discriminative settings to estimate Y from the given X has comparatively better performance.

Problem setting: In our problem setting, we first define a *relevant perturbation* – a targeted or untargeted perturbation is said to be relevant if $g_\phi(x_i) \neq g_\phi(y_i)$ given $x_i, y_i \in \mathbb{R}^n$ with the relation between the two as in equation 1. Our objective is to learn a mapping $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ such that $g_\phi\left(f_\theta(x_i)\right) = g_\phi(y_i)$, parametrized by θ. We propose to solve this problem using *sparse* indirect supervision provided in form of $z_i = g_\phi(y_i)$. This is particularly useful in scenarios where obtaining z_i is easier as compared to y_i for *few samples*, for example class labels in case of natural images, organ boundaries in medical images etc. We realize f_θ using a DNN based reformer model. g_ϕ depends on Z and if not provided a *priori*, it is also realized using a DNN, for example in case of classification, g_ϕ is a discriminative CNN. We assume that even though paired samples $\{x_i, y_i\}$ are not available, we have a set of unperturbed data point $\mathcal{Y}_0 = \{y | y \sim P_Y\}$ with corresponding set $\mathcal{Z}_0$, where P_Y is the distribution of unperturbed data. We use $\mathcal{Y}_0$ to estimate ϕ. We also use $\mathcal{Y}_0$ for pretraining of the reformer network f_θ. Subsequently, we minimize a loss $L\left(g_\phi\left(f_\theta(x_i)\right), z_i\right)$ w.r.t. θ to learn the desired mapping f_θ.

We observed that the reformer network, pretrained on $\mathcal{Y}_0$, tries to reduce perturbations from X to some extent, similar observations were made in (Meng & Chen (2017)), however, (Meng & Chen (2017)) chose to ignore samples with large perturbations due to the limited reduction produced by the reformer. The indirect supervision in our approach helps f_θ to handle even large perturbations by using the information provided in form of Z. This is pictorially shown using a 2D diagram in Fig. 1(a).

Mitigating perturbations: To understand the motivation behind the proposed work, let us consider a simple scenario where $x_i = y_i + \epsilon$ are obtained with an additive perturbation applied on y_i and the indirect supervision is provided by a classifier. We borrow the setup from (Shafahi et al. (2018)) and assume that a given dataset contains m classes defined by probability density functions $\{p_c\}_{c=1}^m$ which are bounded as $U_c = \sup_Y p_c(X)$. All the data point lie on a unit sphere $\mathcal{S} = \{y \in \mathbb{R}^n | ||y||_2 = 1\}$. Our classifier g_ϕ partitions $\mathcal{S}$ into m disjoint subsets. In what follows, we show that for a given clean sample y_i belonging to class c, that is $g_\phi(y_i) = c$, the probability of the classifier g_ϕ correctly classifying perturbed sample x_i into the same class c becomes zero asymptotically even

for small values of ϵ. Thus, for correct classification, x_i need to be enhanced by the reformer, f_θ, to bring them closer to y_i.

Let $\mathcal{S}_c = \{y|g_\phi(y) = c\}$ is the portion of $\mathcal{S}$ labeled as class c and $\bar{\mathcal{S}}_c$ is its complement. If the fraction $\mathcal{S}_c$ is less than or equal to $\frac{1}{2}$, the fraction of $\bar{\mathcal{S}}_c$ will be at least half sphere, therefore ϵ-expansion of $\bar{\mathcal{S}}_c$, represented by $\bar{\mathcal{S}}_c(\epsilon)$, obeys the following

$$\mu[\bar{\mathcal{S}}_c(\epsilon)] \geq 1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{2}$$

where $\mu[\mathcal{S}_c]$ represents the normalized measure/surface area of $\mathcal{S}_c$. Right hand side of equation 2 comes from the fact (Shafahi et al. (2018); Milman & Schechtman (2009)) that the normalized measure of geodesic ϵ-expansion of a half sphere is at least

$$1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{3}$$

Now the data point x will only be classified correctly if it belongs to the complement of $\bar{\mathcal{S}}_c(\epsilon)$. Let us represent the complement of $\bar{\mathcal{S}}_c(\epsilon)$ by $\mathcal{R}$ then

$$\mu[\mathcal{R}] \leq \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{4}$$

Thus the probability of $x \in \mathcal{R}$ being correctly classified is bounded by

$$P\left(g_\phi(x) = c|g_\phi(y) = c\right) \leq V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{5}$$

where V_c is the normalized supremum of p_c. It is independent of ϵ and obtained by the multiplication of U_c and surface area of $\mathcal{S}$.

As can be seen in Fig. 1(b) that the upper bound in equation 5 quickly vanishes even for very small increment in values of ϵ when n is large such as 1024, which is the dimension of CIFAR-10 samples. Hence, the probability of the classifier g_ϕ correctly classifying x_i into class c asymptotically reaches to zero. This can extended for an autoencoder based reformer where $f_\theta(x_i) = x_i$ and the probability of the classifier g_ϕ correctly classifying $f_\theta(x_i)$ into class c also reaches to zero. In other words, unless $f_\theta(x_i)$ comes close to y_i, the classifier is unable to assign the class to $f_\theta(x_i)$ which is same as y_i. This is exactly what is done by f_θ in presence of indirect supervision when we reduce the error in decision of g_ϕ by adjusting θ, which in turn reduces perturbation from x_i to discover the desired y_i.

Design: In all our experiments we use U-net (Ronneberger et al. (2015)) based convolutional autoencoder as the reformer to realize f_θ with trainable parameters as θ. We first adjust these parameters during pretraining using a reconstruction loss (L_r) applied on the set $\mathcal{Y}_0$, as mentioned above.

$$\theta^0 = \arg\min_\theta L_r(\mathcal{Y}_0) \tag{6}$$

$$L_r(\mathcal{Y}_0) = \frac{1}{|\mathcal{Y}_0|} \sum_{y \in \mathcal{Y}_0} ||f_\theta(y) - y||_2 \tag{7}$$

Unlike f_θ, design of g_ϕ depends on Z and chosen accordingly, for example in case Z is class labels, given in form of 1-hot vectors, g_ϕ is designed as a CNN classifier, on the other hand when Z is organ boundary in medical images, g_ϕ is designed as a segmentation network. We use g_ϕ as a general notation, accordingly the loss which is used to adjust the parameters ϕ, denoted by L, depends on the design of g_ϕ, for example when g_ϕ is a classifier categorical crossentropy is used as the loss function. Similar to f_θ, g_ϕ is also pretrained using set $\mathcal{Y}_0$. Once trained on $\mathcal{Y}_0$, the parameters of g_ϕ are freezed (ϕ^0) and *not* adjusted subsequently. This is the key design component of the entire model. Post-pretraining f_{θ^0} and g_{ϕ^0} are combined to form an end-to-end trainable pipeline $h_{\theta^0,\phi^0}(.) = g_{\phi^0}\left(f_{\theta^0}(.)\right)$.

In the combined model h_{θ^0,ϕ^0}, only trainable parameters are θ^0. These are adjusted, as in equation 8, during fine-tuning on 10% samples, x_i's (sparse indirect supervision) separated out from the perturbed set. For these separated x_i's, corresponding z_i's are assumed to be known.

$$\theta^1 = \arg\min_\theta \left[L + L_r\right] \tag{8}$$

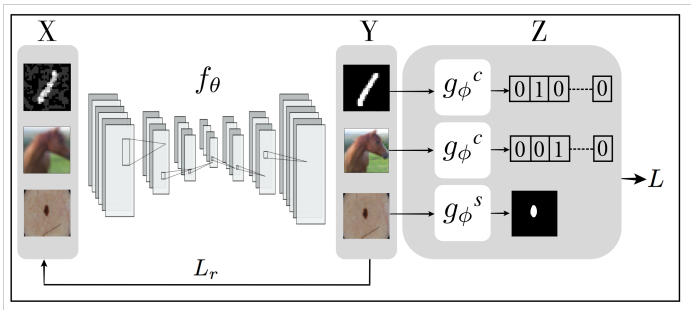


Figure 2: Architecture used in the proposed approach with multiple downstream models. X represents the perturbed input image which is given as input to a reformer (autoencoder) f_θ. Y represents the output (desired clean images) produced by the reformer and Z indicates output given by different downstream models. L_r is the reconstruction loss used by the reformer while L is the loss representing error in the decisions of the downstream models.

where the reconstruction loss (L_r) is included to avoid overfitting on the small set used during fine-tuning. We have used equal weights for both the losses in equation 8, however a weighted combination can also be explored. We also include samples from $\mathcal{Y}_0$, in equal number to perturbed samples, during fine-tuning to ensure that f_θ does not affect unperturbed samples. Note that the loss L is defined over the outputs of g_ϕ but minimized by adjusting the parameters of f_θ. This means, once g_ϕ is trained on $\mathcal{Y}_0$, it can be deployed on some device and need not to be touched again. f_θ can be trained with a copy of g_ϕ and used as a plug-in to the deployed version of g_ϕ. The pictorial representation of the proposed approach is shown in Fig. 2.

3 Related Work

In the related works, the first category of approaches is the analytical methods developed with the efforts of several years and have an important place in the domain of the problem considered in this work.

Classical Filtering Methods: The Gaussian filter, median filter, Wiener filter etc. are some of the well known classical filters in this domain. Some unconventional but very effective ones are methods based non-local averaging of pixels (Buades et al. (2005)). One of the most popular in this direction is BM3D (Dabov et al. (2007)) which groups similar patterns in the image and filters them jointly. (Lefkimmiatis (2017); Zeng et al. (2019); Liu et al. (2018)) have used non-local modules for denoising of images to improve the performance of the downstream model. However, a major limitation comes from the fact that the model settings need to be changed according to the noise levels, which is true for the other classical methods as well.

Discriminative Methods: In 2009, (Jain & Seung (2009)) proposed the idea of using deep discriminative models for image denoising which was later extended to other inverse problems (Ledig et al. (2017); Hradiš et al. (2015)). Improvements on the work of (Jain & Seung (2009)) were proposed in (Zhang et al. (2017b)), where a very deep residual network was trained to predict noise instead of clean images.

The techniques in this category work under fully supervised settings, therefore require a set of noisy images with respective clean images for training. In case of the unavailability of the ground truth clean images, these techniques lose their utility. There have attempted to combine deep CNN models with analytical approaches such as half-quadratic splitting (Schuler et al. (2015)), however the deep CNNs with plain discriminative settings give better performance.

Unsupervised and Self-Supervised Methods: Alternates to supervised CNN based enhancement models are unsupervised and self-supervised approaches. Denoising autoencoder (Vincent et al. (2008)) can be considered as one of the early unsupervised approaches in this category, however, the approaches like generative adversarial network (GAN) (Chen et al. (2018)) provide better performance in restoration of images. The model in (Chen et al. (2018)) requires random samples of noisy and clean images to produce a paired set of noisy images and corresponding clean image given as

input to the generator. The generated pairs are subsequently used to train a fully supervised discriminative network for noise reduction however the efficiency is limited to the performance of GAN which depends on the amount of data available for training.

Self supervised approaches overcome the limitation of GAN model. In 2018, (Lehtinen et al. (2018)) proposed an efficient self supervised technique named as Noise2noise which does not require clean images for training. Their network can produce clean images however, it requires a set of two noisy images obtained by two different kinds of noise induced on the same image. To remove the requirement of pairs of noisy images (Krull et al. (2019)) proposed an approach, named as Noise2void, which is also independent of the availability of ground truth data or clean images. However, it failed on structured perturbations due to the independence assumption on noise. Recently, in 2019 (Batson & Royer (2019)) proposed Noise2Self, which does not require any assumptions such as information of noise or availability of ground truth.

A common problem with all the aforementioned approaches is that these approaches fail to take into account the requirement of the subsequent downstream models. The perturbations are reduced to improve the quality of the images, where the quality improvement is driven by human perception, not by the perception of the model. In contrast, our approach brings into picture the perception of a downstream model using an indirect supervision approach.

4 Experimental Results

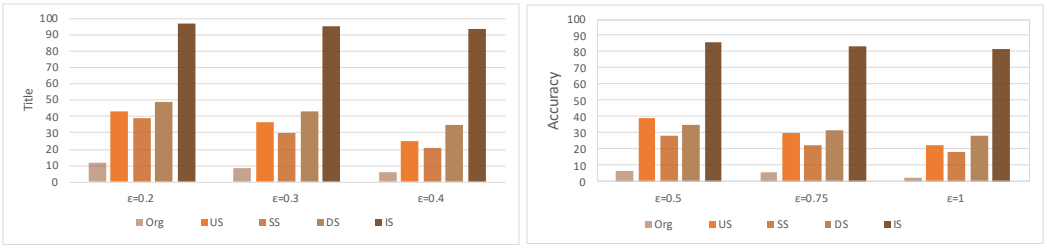
This section begins with details of the dataset used for experiments and training procedure adopted for the proposed approach. It is followed by baseline information and description of the obtained results.

Dataset and Training: We perform experiments with MNIST, CIFAR-10-C Hendrycks & Dietterich (2018) and ISIC 2018 skin lesion[1] dataset. MNIST is a dataset of handwritten digit images, which contains well sets for training validation and testing. We perform pretraining of both, f_θ and g_ϕ, using the training set. Subsequently, we use pretrained and froze downstream model g_{ϕ^0} to generate adversarial images using the MNIST test set images by applying a fast gradient sign attack (FGSM) attack (Goodfellow et al. (2014)). From the generated set of images, we use 10% images for finetuning of f_θ using indirect supervision to obtain f_{θ^1} and remaining adversarial images for testing the accuracy of g_{ϕ^0} on the outputs of f_{θ^1}.

CIFAR-10-C is the perturbed version of CIFAR-10 test set, which has been generated by adding different types of corruptions with varying severity level in five categories namely noise, blur, weather, digital, and extrathe . We consider five different most effective corruptions with highest severity level which in total becomes a set of 50,000 corrupted images (10,000 per corruption). We use the clean CIFAR-10 training set for pretraining and the set of corrupted images (CIFAR-10-C) for finetuning and testing in the same manner as MNIST. No adversarial attack is considered due to the avaialbility of the corrupted data. ISIC is a collection of high resolution dermoscopic skin lesion images (Codella et al. (2019)). It contains 2,596 images for which segmentation masks of lesions are provided, whereas for the remaining images, which belong to test set, the segmentation masks are not made publicly available. We use 2,000 images for pretraining and remaining 596 images for validation during pretraining. Since test set annotations are not available, we generate perturbed images from all the 2,596 available clean images using FGSM to create a reasonable test set. This is motivated from the observations made during experiments on MNIST and CIFAR-10-C, where the pretrained models do well on unseen clean test images but fail miserably on perturbed versions of the test images. Since the objective here is to remove *relevant perturbation*, as defined in section 2, the performance of g_{ϕ^0} on the outputs of f_{θ^1} becomes the preferred choice of evaluation metric and used in all the experiments. Various hyperparameter values used during different experiments are mentioned in appendix C.

Baseline: We have compared our approach results with four different baselines. (i) Original (Org) model: performance of the original downstream model on perturbed images is the first baseline, mainly considered to find the degradation in the performance of the model. (ii) Unsupervised (US) approach: we consider a GAN model (DC-GAN) to represent an unsupervised approach of mitigating perturbation where the generator of GAN produces unperturbed images. (iii) Self Supervision

[1] https://challenge2018.isic-archive.com



(a) (b)

Figure 3: Experimental results on MNIST data for $(\mathbf{a})$ LeNet $(\mathbf{b})$ Resnet-18. For both networks the proposed approach (IS: indirect supervision) outperforms the other approaches.

Model	CIFAR-10					CIFAR-10-C				
	Org	**US**	**SS**	**DS**	**IS**	**Org**	**US**	**SS**	**DS**	**IS**
WRN-28-10	89.16	51.48	83.14	88.16	84.82	56.50	30.58	43.48	42.19	**69.77**
WRN-40-2	91.93	53.02	85.22	89.86	88.88	54.20	31.32	44.96	44.72	**70.99**
RN-50	92.04	50.40	85.68	89.90	85.0	54.25	30.16	44.62	42.02	**67.24**

Table 1: Comparison of experimental results on CIFAR-10-C dataset with three different classifiers. The combined model, h_{θ^1,ϕ^0}, finetuned with indirect supervision (IS), gives best performance on corrupted images, however there is some degradation in the clean image performance.

(SS): we consider the recently proposed Noise2Self approach as our third baseline for comparison. (iv) Direct Supervision (DS): our final baseline is a discriniminative filtering approach in which gold standard clean images are used as ground truth to finetune f_{θ^0}. DS is considered only for comparison, consider its superior performance as compared to other baseline approaches. Wherever required, we use the same setting, as our proposed approach (10% of the total perturbed images for sparse supervision, remaining for testing) for all the baseline approaches, for example US and DS. In addition, for DS we assume that the gold standard clean images, corresponding to the 10% perturbed images, are also available for direct supervision.

MNIST Results: MNIST data is considered for the classification task, where we have considered two well-known models, LeNet and Resnet-18 for the downstream task. Subsequent to pretraining on the training set of MNIST with crossentropy loss, both the networks are used for adversarial image generation from test set images using FGSM. Severity level of the adversarial attacks are determined using a scaling parameter ε (Goodfellow et al. (2014)). We use $\varepsilon \in \{0.2, 0.3, 0.4\}$ for LeNet and $\varepsilon \in \{0.5, 0.75, 1\}$ for Resnet-18. Fig. 3 presents a summary of the performances of the considered approaches. Both the models achieved more than 90% accuracy on a clean test set, however, their accuracy came down to 10% or lower than that on adversarial images. The baseline approaches restore the model performances to some extent, but the proposed approach outperforms all the baselines with considerable margin while using sparse indirect supervision. AN interesting observation here is the difference in the performance of the proposed approach and DS, where DS gives suboptimal performance, mainly due to overfitting on the small set available for finetuning. In contrast, the indirect supervision in the proposed approach allows only that enhancement which is relevant to the downstream model. Some of the output images are shown in appendix B.

CIFAR-10-C Results: Here we consider three different Resnet variants (Zagoruyko & Komodakis (2016)) for classification - Wide Resnet-28-10 (WRN-28-10), Wide Resnet-40-2 (WRN-40-2), and Resnet-50 (RN-50). These variants differ from each other in terms of depth and number of parameters. Similar to MNIST experiments, all these models are pretrained using crossentropy loss on clean training set from CIFAR-10 dataset. However, the key difference here is in finetuning and testing. As mentioned above, unlike MNIST, perturbed CIFAR-10 images are obtained from CIFAR-10-C data. Among the 16 different types of corruption, which were used to create CIFAR-10-C, we select elastic transformation, frost, impulse noise, shot noise and zoom blur, as these considerably degrade the performance of classification network. Table 1 shows a comparison of different approaches, where the proposed approach restores the classifier performances to a considerable extent. On the

Approach	Classifier Parameters	PSNR
Org	-	18.48
US	-	19.28
SS	-	22.68
DS	-	26.48
IS (WRN-28-10)	36.47 M	25.48
IS (WRN-40-2)	2.24 M	25.10
IS (RN-50)	23.53 M	24.24

Table 2: Mean PSNR values for enhanced CIFAR-10-C images from different approaches.

Approach	Mean IoU
Org	0.17
US	0.25
SS	0.64
DS	0.67
IS	**0.69**

Table 3: Mean IoU values observed during testing the segmentation network on the outputs of different approaches.

other hand, US and SS degrade the classification performances due to their own limitations. DS also degrades performance due to overfitting, as the training accuracy during finetuning reaches values larger than 85% for all the classifiers, whereas testing accuracy remains under 50%.

An interesting observation here is the performance WRN-40-2, which has a considerably smaller number of parameters as compared to its counterparts, shown in Table 2. WRN-40-2 performance on CIFAR-10-C images is lower than the other two classifiers in its original form, however, when plugged in with the reformer network in the proposed approach, its performance becomes better as compared to its counterparts. Another interesting observation is the enhancement of the images evaluated using mean PSNR values shown in Table-2. DS, while using gold standard clean images, results in the best PSNR values, which is consistent with the literature. However, poor performances of all the classifiers on the outputs of DS (Table 5) provides evidence that the enhancement without taking into account the feedback of downstream models may not be useful. The same can also be observed from the enhanced images shown in appendix C. Note that the proposed approach results in PSNR values close to DS without even using the gold standard clean images.

ISIC Skin Lesion Segmentation Results: Here we considered segmentation as our downstream task and used a U-net based segmentation network. The network is pretrained with dice loss using the procedure mentioned above. Subsequently, we generate 2,596 perturbed images from all the available clean images using FGSM with severity controlling parameter value $\varepsilon = 0.15$. To evaluate the segmentation accuracy we consider mean intersection over union (IoU) values and observe that the segmentation network performance comes down from 0.76 to 0.17 on adversarial images. We apply all the considered baseline approaches along with the proposed approach to mitigate adversarial perturbations and use 250 perturbed images ($\sim 10\%$ of the total perturbed images) for finetuning. We test the segmentation network performance on the outputs of all the approaches obtained from the remaining perturbed images and present the observation in Table 3. The proposed approach outperforms the existing approaches in this experiment as well, even though the existing approach seems to have benefitted from the overlap between the sets of images used for pretraining and adversarial image generation. Few sample images with corresponding outputs obtained from the proposed approach are shown in appendix D.

5 Conclusion

In this work we showed that the problem of mitigating targeted and untargeted perturbations can be modeled in the indirect supervision framework. The proposed approach matches the enhancement performance of supervised models without demanding gold standard data. Also it does not require any information about the perturbation characteristics. It uses the information provided by a variable depended on gold standard data for sparse indirect supervision and outperforms the existing approaches in various downstream tasks. This is a very useful property for application such as medical image analysis where getting a variable dependent on clean data, such as segmentation mask, is easier than getting the clean data itself. Various experiments performed in the work highlight the importance of downstream model's feedback in the perturbation mitigation process. Observed results are promising and we believe that this work would provide motivation for considering indirect supervision for other applications. In future, we will take the work forward by exploring the possibility of combining indirect supervision with other approaches.

REFERENCES

Joshua Batson and Loïc Royer. Noise2self: Blind denoising by self-supervision. *CoRR*, abs/1901.11365, 2019. URL http://arxiv.org/abs/1901.11365.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

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*, pp. 387–402. Springer, 2013.

Gustav J Braun and Mark D Fairchild. Image lightness rescaling using sigmoidal contrast enhancement functions. *Journal of Electronic Imaging*, 8(4):380–393, 1999.

Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 60–65. IEEE, 2005.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164, 2018.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16 (8):2080–2095, 2007.

David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE international conference on computer vision*, pp. 633–640, 2013.

Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pp. 2280–2289, 2019.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Michal Hradiš, Jan Kotera, Pavel Zemcık, and Filip Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of BMVC*, volume 10, pp. 2, 2015.

Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In *Advances in neural information processing systems*, pp. 769–776, 2009.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3587–3596, 2017.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pp. 1673–1682, 2018.

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

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 135–147, 2017.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer, 2009.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pp. 1196–1204, 2013.

Aditi Raghunathan, Roy Frostig, John Duchi, and Percy Liang. Estimation from indirect supervision with linear moments. In *International conference on machine learning*, pp. 2568–2577, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, 2015.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Aichen Wang, Wen Zhang, and Xinhua Wei. A review on weed detection using ground-based machine vision and image processing techniques. *Computers and electronics in agriculture*, 158: 226–240, 2019.

Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pp. 341–349, 2012.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5949–5957, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017b.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017c.

A HYPERPARAMETERS:

In all our experiments, we have used publicly available models for both f_θ and g_ϕ. We considered Adam optimizer for both training and finetuning. Number of samples used at various stages along with various hyperparameter values are shown below. We used Pytorch Python library for all our experiments.

Dataset	Sample set size			Initial LR		Batch Size		No of epochs			L
	P	F	T	P	F	P	F	P(f_θ)	P(g_ϕ)	F(f_θ)	
MNIST	50K	1K	9K	1e-2	1e-3	128	32	1K	1K	750	CCE
CIFAR-10	50K	5K	45K	1e-3	1e-4	128	32	1.5K	2.5K	1K	CCE
ISIC Skin Lesion Dataset	2596	250	2346	1e-4	1e-5	16	4	500	350	1K	DC

Table 4: The table represents the information of datasets used in the experiments, where P, F, and T represent pretraining, finetuning, and testing phase. * Learning rate in pretraining and fine-tuning is reduced by the factor of 10 after every 50 and 100 epochs respectively. * LR stands for learning rate, CCE for Categorical Cross Entropy and DC for Dice Coefficient.

B MNIST IMAGE OUTPUTS

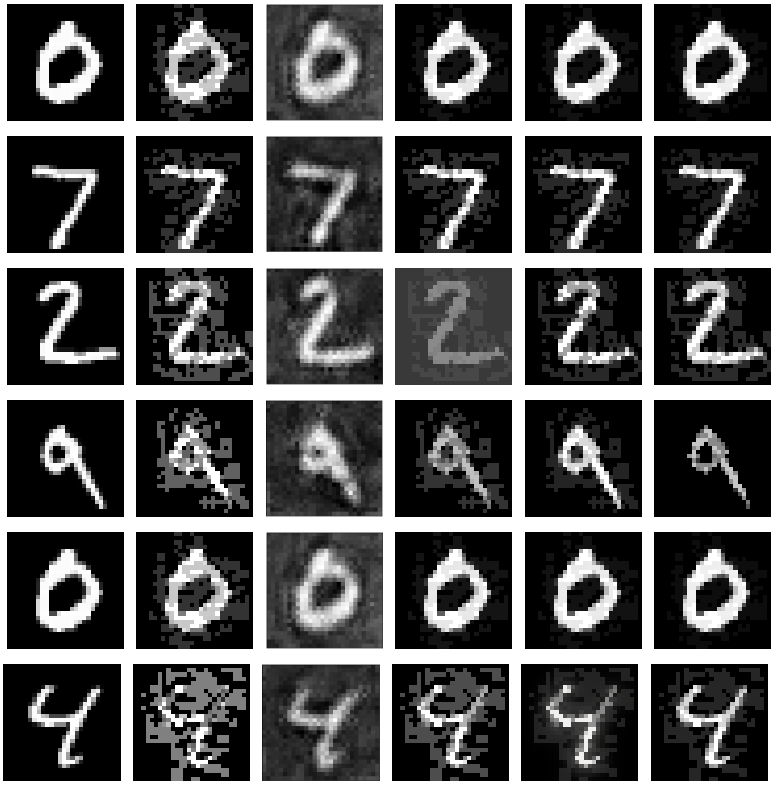


Figure 4: Upper 3 rows shows the images for LeNet for ε values 0.2, 0.3 and 0.4 respectively and remaining 3 for the Resnet-18 for ε values 0.5, 0.75 and 1. First column shows the original images from MNIST dataset. Second column shows perturbed images generated using FGSM attack. Third column shows the images generated using unsupervised approach. Fourth columns shows the images generated by self supervision (noise2self). Fifth column shows direct supervision images and the last column shows the images generated by our approach. For some samples indirect supervision performs better than direct supervision.

C CIFAR-10-C IMAGE OUTPUTS

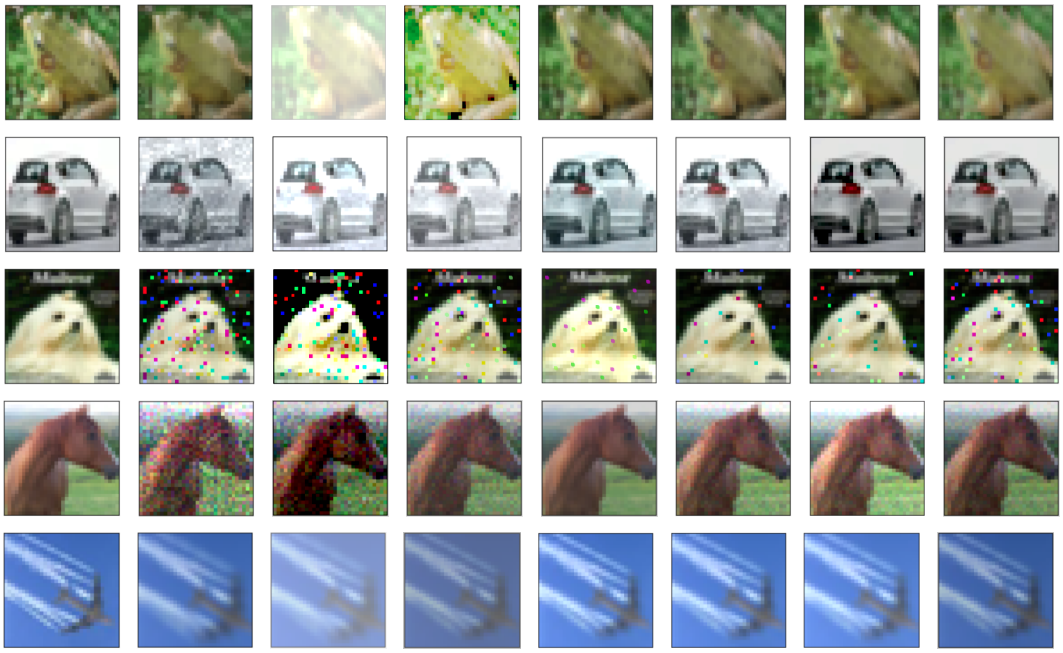


Figure 5: First column shows the original images from CIFAR-10 dataset. Second column shows corresponding the perturbed images from CIFAR-10-C dataset. Third column shows the images generated using unsupervised (GAN) approach. Fourth column shows the images generated using Self supervised (noise2self) approach. Fifth column shows the images generated using direct supervision. Remaining three columns shows the images generated by our approach using Resnet-28-10, Resnet-40-2 and Resnet-50 respectively. All the rows represents different perturbations from CIFAR-10-C dataset. Elastic transformation, frost, impulse noise, shot noise and zoom blur are the perturbations shown in the rows from top to down.

D ISIC LESION SEGMENTATION OUTPUTS

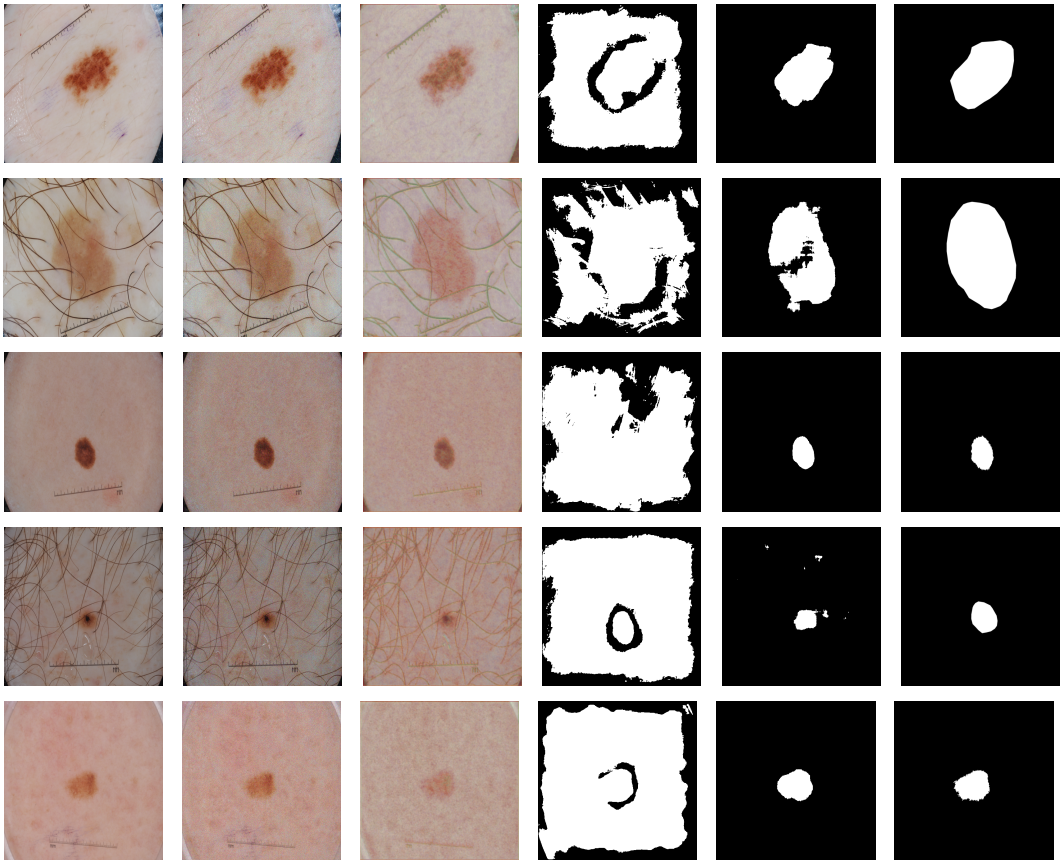


Figure 6: First column shows the original images from ISIC skin lesion dataset. Second column represents the adversarial images generated using FGSM attack. Third column shows the images generated by our approach. Fourth column shows the segmentation model's output on adversarial perturbed images. Fifth and sixth column shows improved segmentation output using our approach and corresponding masks respectively.