CIIDefence: Defeating Adversarial Attacks by Fusing Class-specific Image Inpainting and Image Denoising

Puneet Gupta
IIT Indore
Indore, India
puneet@iiti.ac.in

Esa Rahtu
Tampere University
Finland
esa.rahtu@tuni.fi

Abstract

This paper presents a novel approach for protecting deep neural networks from adversarial attacks, i.e., methods that add well-crafted imperceptible modifications to the original inputs such that they are incorrectly classified with high confidence. The proposed defence mechanism is inspired by the recent works mitigating the adversarial disturbances by the means of image reconstruction and denoising. However, unlike the previous works, we apply the reconstruction only for small and carefully selected image areas that are most influential to the current classification outcome. The selection process is guided by the class activation map responses obtained for multiple top-ranking class labels. The same regions are also the most prominent for the adversarial perturbations and hence most important to purify. The resulting inpainting task is substantially more tractable than the full image reconstruction, while still being able to prevent the adversarial attacks.

Furthermore, we combine the selective image inpainting with wavelet based image denoising to produce a non-differentiable layer that prevents attacker from using gradient backpropagation. Moreover, the proposed nonlinearity cannot be easily approximated with simple differentiable alternative as demonstrated in the experiments with Backward Pass Differentiable Approximation (BPDA) attack. Finally, we experimentally show that the proposed Class-specific Image Inpainting Defence (CIIDefence) is able to withstand several powerful adversarial attacks including the BPDA. The obtained results are consistently better compared to the other recent defence approaches.

1. Introduction

Deep neural networks (DNN) have recently made tremendous impact in many fields, including audio analysis and computer vision. For instance, many state-of-the-art methods in speech recognition, face detection, fraud detection and autonomous driving nowadays rely on deep neural network architectures. In some of these applications, the system performance may have crucial financial and security impacts. In such cases, it becomes important that the system is resilient against engineered attacks targeting to influence the system behaviour.

Unfortunately, many modern DNN architectures are surprisingly vulnerable to well crafted attacks that utilise adversarial examples [30]. The adversarial example is created by perturbing the input image in such a way that it remains virtually indistinguishable from the original image, but unlike the original input, it will be misclassified by the system with high confidence. Such weakness could lead to severe consequences in many important applications [8]. These challenges have motivated researchers to propose defences that protect the networks against adversarial attacks [16].

One group of defence methods modify the network architecture [22] or the training set [31]. The key idea in these approaches is to minimize the original loss function while increasing the perturbation space around the clean inputs. The main drawback is that the network needs to be re-trained. Moreover, these defences are usually effective only against the attacks considered during the training. Another group of defence methods circumvents the need for network re-training by modifying only the input image to mitigate the possible adversarial perturbations. These methods are usually referred as transformation based defences. Our approach is also built according to this paradigm.

The transformation based defences are typically relying on gradient masking techniques [21], which aim at preventing the attacker from using the gradient descend optimisation for constructing the adversarial examples. However, recent works have introduced approaches that make such defences vulnerable [2]. Furthermore, the transformation based defences reduce the quality of the original input image, which may lead to reduced performance.

In this paper, we propose a novel defence called Class-specific Image Inpainting Defence (CIIDefence). The main contributions of our work include: i) introduction of class-

specific image inpainting approach for preventing adversarial attacks, ii) method for extracting image areas that are potential for adversarial perturbations, and iii) fusing inpainting based reconstruction method with traditional image denoising to improve the performance and provide a non-differentiable layer for gradient masking. In the experiments, we show that the proposed method outperform previous well-known defences and is able to withstand several state-of-the-art attacks, including Backward Pass Differentiable Approximation (BPDA) [2] to a large extend. Furthermore, the original classification accuracy is only minimally impacted by the application of our defence.

2. Preliminaries

In this section, we will outline a few well known attack and defence approaches that are useful for understanding the proposed method and the related work. Let X_c and $f(\cdot)$ denote the input image and the classifier (e.g. neural network), respectively. Further, assume that the classification result of X_c, denoted as x^c, is obtained as $x^c = \text{argmax}_i f(X_c)_i$, where $f(X_c)_i$ denotes the probability of X_c belonging to the class i. The classifier f is learned by minimising a loss function l with respect to the training set.

The adversary aims at generating an adversarial image X_a by adding small a perturbation δ to X_c such that: i) the perturbation is imperceptible to human eyes (i.e., δ is small), and ii) the classification of X_a (denoted as x_a^c) is incorrect. This kind of attack is referred as untargeted as opposed to targeted attack, where we additionally require X_a to be classified in the predefined target class y (i.e., $x_a^c = y$).

One computationally efficient attack is the Fast Gradient Sign Method (FGSM) presented in [9]. The FGSM computes gradients of the loss function l with respect to the image pixels and then subtracts or add a fixed value ε from each pixel depending on the sign of the corresponding gradient. More specifically,

$$X_a = clip\left(X_c + sign\left(\nabla l\left(f\left(X_c\right), x^c\right)\right)\right) \qquad (1)$$

where $clip$, $sign$ and ∇ denote image clipping (applied to restrict the maximum perturbation), sign function and gradient operator, respectively. The outlined process can be repeated until the desired adversarial image is obtained [13]. Such iterative version of the FGSM is known as Iterative Gradient Sign Method (IGSM) [13]. Another iterative version of the FGSM is Projected Gradient Descent (PGD) [18] which considers only the allowable perturbations and hence does not require image clipping.

Deep Fool (DFool) [19] attack treats the classifier f as a set of linear decision boundaries and produces perturbations that push the adversarial image across these boundaries, which eventually leads to mis-classification. Carlini & Wagner (C&W) [5] is another powerful attack which produces an adversarial image by jointly minimizing the adversarial perturbations and the difference between logits of the true and adversarial classes. That is, it minimizes

$$\|\delta\|_p + c * max\left(-\kappa, Z(X_a)_{x^c} - max\{Z(X_a)_{y \neq x^c}\}\right) \qquad (2)$$

where $Z(\alpha)_b$ denotes the logits of f for class b when image α is given as input, and κ, c and p are the margin parameters (i.e. constant defining contribution of different constraints and norm type, respectively).

A notable similarity in the above methods is the fact that they all need the gradients of the loss function l to construct the adversarial image. In fact, if these gradients are available, the attacks are able to breach any defence. Similarly, from defence perspective, most of these attacks can be prevented by making the gradients inaccessible [2]. This observation is leveraged in many transformation based defences by adding a preprocessing layer that prevents an attacker from obtaining the gradient values (gradient masking). Mathematically, if the preprocessing layer is denoted as $g(\cdot)$, the classification is performed as $x^c = \text{argmax}_i f\left(g\left(X_c\right)\right)_i$ and the attack has to be made on $f\left(g\left(\cdot\right)\right)$ rather than $f\left(\cdot\right)$.

Common approaches for gradient masking [2] include: i) *shattered gradients* where incorrect gradients are provided due to numeric instability or by adding non-differentiable layers; ii) *stochastic gradients* where randomness is used to provide different gradients in every iteration; and iii) *vanishing/exploding gradients* where impractical gradients are produced due to deep architectures.

Xu et al. [33] utilised the shattered gradient technique by adding non-differentiable layers like JPEG compression. This type of defence was later breached by the BPDA attack [2], where the non-differentiable layers are approximated by differentiable alternative during the backpropagation. BPDA is also able to breach defences that are based on vanishing and exploding gradient techniques. Xie et al. [32] presented a defence using the stochastic gradient technique. In their method, the input image is randomly padded and cropped before the classification. Such defences were later breached by expectation over transformation attack [2].

3. Related Work

Several works have utilized image denoising for the defences such as bit depth slicing [33], median smoothing [33], non-local mean filtering [33], JPEG compression [6], and high guided denoisers [14]. Liao et al. [14] used autoencoder to produce a denoised image that was similar to the clean input image at the high levels of the CNN output.

Guo et al. [10] apply image quilting and total variance minimization (TVM) techniques to prevent the attacks. Image quilting produces an approximation of the input by replacing small image areas with similar looking patches obtained from the external dataset using k-nearest neighbours classifier. The TVM technique [10] removes a small set of

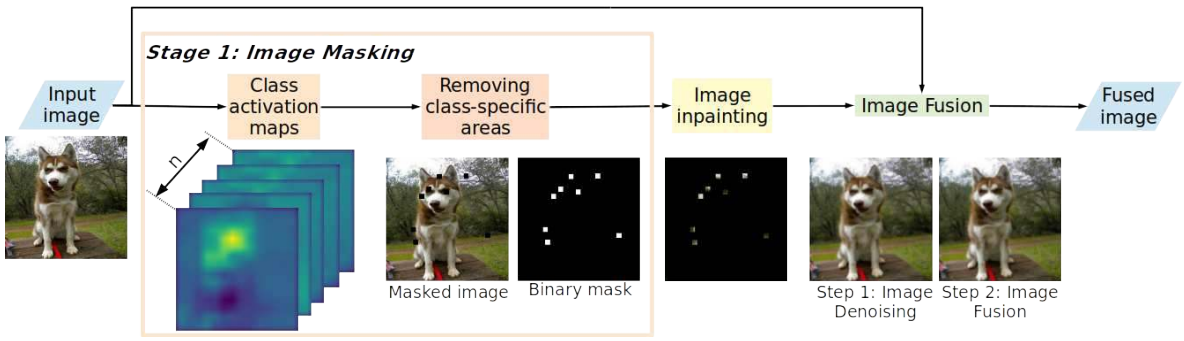

Figure 1. Flow-graph of the proposed defence, $CIIDefence$.

randomly chosen image pixels and reconstructs them to mitigate possible adversarial perturbations. A similar approach was proposed in [23]. This method uses semantic maps and randomization to select a small number of pixels that are subsequently replaced with randomly selected neighbours. The procedure creates noise that is later removed using a wavelet denoising filter. This kind of approach is referred as pixel deflection (PD). The classification performance with TVM and PD defence is further improved by running the classifier multiple times and aggregating the results.

Another group of transformation based defences uses generative adversarial network (GAN) to reconstruct the input image. In this way, the adversarial perturbations are likely to be removed. Samangouei et al. [25] proposed a GAN based method that can mitigate several adversarial attacks when tested on MNIST dataset. Their approach obtained the reconstruction using GAN with random number input. Similar technique was proposed in [27], which aims at reconstructing the unperturbed realistic image using the adversarial example rather than random number. Unfortunately, these approaches have high negative impact on the original classification accuracy as the dimensionality of the input space and number of classes increase [5].

All the existing GAN based defences aim at reconstructing the full image, which is a challenging task considering the large dimensionality of the input space. In contrast, we do not aim at reconstructing the entire image, but only a small set of carefully selected image patches (see Fig. 1). Therefore, the reconstruction task is considerably easier compared to the full image reconstruction in [25, 27].

One of the key ideas in CIIDefence is to select those using Class Activation Map (CAM) technique that pinpoints the image parts most influential to the classification outcome. The same image parts are also the most prominent for the adversarial perturbations and hence it makes sense to target the reconstruction on them rather than random locations as in [17]. CAM technique was also applied in [23], but they produced only a single activation map using weighted average over different classes, whereas we utilise class specific maps without any randomization procedure. Moreover, due to lack of randomized components, we do not need any time consuming ensemble based classification procedures as in [10] and [23]. Finally, we fuse our inpainting based defence with wavelet based image denoising [23] to further improve the results. In addition, this combination provides a non-differentiable layer that turns out to be difficult to approximate with simple differentiable alternatives.

4. Proposed Defence

In this section, we will present the proposed defence. An adversarial attack analyses the classifier behaviour to fool the defence by employing two strategies: i) finding image areas which help the classifier to make correct classification and modify them to reduce the score of correct class; and ii) finding the areas corresponding to the incorrect classes and modifying them to increase the corresponding scores.

To prevent such attacks, we aim at modifying the adversarial image in a way that the key areas involved in the classification decision would be similar to those in the original clean image. The proposed method consists of the following three stages: image masking, inpainting, and fusion. In the first stage, we detect and remove areas that play crucial role in the classification decision. These areas are reconstructed in the second stage using a GAN based image inpainting method by considering both local and global image characteristics.

In the last stage of the method, the input image is denoised and the reconstructed areas are merged with the denoised result. It is important to note that the denoising is not applied to the inpainted regions because it was found to result in providing blurry regions, which eventually degrade the classification performance. Eventually, the fused image is provided to the original classifier. The flow-graph of the $CIIDefence$ is shown in Figure 1 and the steps involved in the method are outlined in Algorithm 1. Further details are provided in the subsequent sections.

4.1. Image Masking

If the adversarial attack is successful, the new (mis)classification result is usually one of the top-k class la-

Algorithm 1 $CIIDefence(I_q, f)$

Require: Input image, I_q and classifier f.
Ensure: Fused Image, I_f.

```
# Stage 1: Image masking
Compute top-n classes using f(I_q)
Initialize images to store binary mask, M and masked image I using M = zeros(size(I_q)) and I = I_q
for each class c in top-n classes do
    Obtain CAM from f(I_q) for c using Equation (3).
    Find p̄ class-specific locations by extracting all the local maximas and choosing the p̄ maximas from them such that their CAM intensities are higher than the CAM intensities of the remaining ones.
    Corresponding to each location, mask the relevant areas by updating M and I using Equation (4).
end for
(Stage 2) Inpaint the selected areas using M and I. Store the resultant image in I_i.
# Stage 3
Denoise I_q and save it as I_d
Fuse the denoised and inpainted images using Equation (5) and save the fused image in I_f.
return (I_f)
```

bels of the original undisturbed classification output. Similarly, the original correct class is usually still within the top-k classes of the new result [23]. Inspired by this observation, we target our inpainting operations to areas that are most significant to the top-n classes according to CAM. The value of n is one of the hyperparameters of our method and in Section 5.1 we will experimentally evaluate how it affects to the performance of the defence. The best results were obtained with $n = 5$, although the performance was relatively stable over a range of different values.

4.1.1 Class Activation Maps (CAM)

In the case of real-world images, the image pixels have highly uneven contribution to the classification outcome [10]. Therefore, if the most influential image areas can be protected from perturbations, the classifier may able to work correctly despite of the adversarial attack. These influential pixels can be detected using either saliency detectors [12] or CAM [36]. In this paper, we utilise the CAM technique presented in [36], since the saliency models are usually optimised to predict human eye-fixation densities rather than reflecting the characteristics of the classifier. In the following, we will shortly describe the basic principle in the CAM technique and refer the reader to [36] for more details.

Zhou et al. [35] demonstrated that the CNN layers have potential to act as object detectors even if the network is trained only for the image classification task. The same

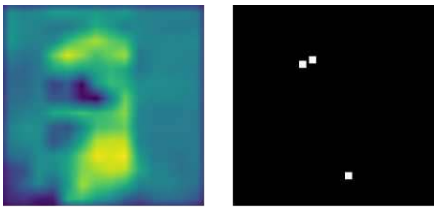

Figure 2. a) An example CAM response obtained using the input image in Figure 1; and b) the corresponding masked image area.

principle is utilised in the CAM technique [36] to perform weakly-supervised localisation of class specific discriminative areas. These areas are obtained by selecting a target class label and back-propagating the corresponding information throughout the CNN layers all the way to the input image.

Intuitively, each unit in CNN is activated by certain patterns that indicate the presence or absence of a class. Let $r_k(x, y)$ indicate the presence of class specific patterns in k^{th} unit at the spatial location (x, y). Now the CAM M_c for a class c is obtained as a weighted sum of these responses as

$$M_c(x, y) = \sum_k w_k^c \times r_k(x, y), \tag{3}$$

where w_k^c denotes the importance of the k^{th} channel for the class c (see [36] for more details). This kind of map does not capture the fine-grained details, but roughly highlights the image locations corresponding to the target class. Figures 1 and 2 illustrate examples of the obtained CAM maps.

The fully connected layers are hampering the localisation performance of the CAM. Hence, they are usually replaced by either global average pooling (GAP) [15] or global max pooling (GMP) layers [20]. In our work, we adopt the GAP based approach since it consider all discriminative areas of an object, whereas GMP considers only the most discriminative part [35]. We are interested in recognising all these parts rather than just the most significant one. For instance, parts like face can be considered as a face if they contain some facial attributes like eyes, nose, eyebrows and so on [37]. Alternatively, one could use GradCAM [26], which is a generalisation of the basic CAM technique and does not require any modifications to the classifier. We did not experiment with GradCAM, since already the basic CAM [36] provide excellent performance.

We will calculate the CAM responses using each of the top n classes as a target class, one at a time. This results in n maps that are later used to determine the reconstructed image areas (see Fig. 1). We note that it is important to obtain the CAM using the same classifier f that is being protected from the adversarial attacks.

4.1.2 Removing class-specific areas

The reconstructed image area is determined using the following procedure. Firstly, we detect $\bar{p}$ highest scoring local maxima from the first CAM response map. We utilise non-maxima suppression to avoid nearly overlapping detections. Secondly, we remove all image data within $(2w+1) \times (2w+1)$ square box centred at each of the selected $\bar{p}$ local maxima. More specifically,

$$X_c(x, y) = 0 \text{ if } |x_i - x| \leq w \text{ and } |y_i - y| \leq w, \quad (4)$$

where X_c refers to the input image, (x_i, y_i) is the location of the selected maxima, $i = 1 \ldots \bar{p}$, and $|\cdot|$ denotes the absolute value. We repeat this process for all n CAM response maps. For later use, we will also create a binary mask M indicating the removed image areas (i.e., $M = 1$ if the pixel is masked and otherwise $M = 0$). Figures 1 and 2 illustrate the described workflow and the obtained binary masks.

4.2. Image Inpainting

The image areas, removed in the previous stage, are reconstructed using image inpainting method. In particular, we use the inpainting technique described in [34]. Their method is based on two-stage coarse-to-fine architecture for increasing the receptive fields and stabilising the training phase. In the first stage, coarse inpainting is carried out using a network that is explicitly trained for minimizing the reconstruction loss. In the next stage, the coarse prediction is refined to produce a realistic image using a network which minimizes reconstruction loss along with the adversarial losses. This network analyses both the global and local image characteristics for the refinement by incorporating the global and local Wasserstein GANs respectively. Figure 1 shows an example of the inpainting result.

4.3. Image Fusion

In some cases the inpainting may result in blurry reconstruction, that may re-focuse the classifier on the remaining image areas. These areas might also be subject to the adversarial perturbations and to mitigate this, we fuse the inpainted result with denoised version of the original input. That is, we apply a denoising technique to the original input image and then replace the pixels indicated by the mask M with the corresponding inpainted versions. In this way, the denoising is not applied to the inpainted areas as it can result in blurring which eventually degrade the classification performance. More specifically,

$$I_r(x, y) = M(x, y) \times I_i(x, y) + (1 - M(x, y)) \times I_d(x, y) \quad (5)$$

where I_r, M, I_i, and I_d represent the fused output image, the mask image, inpainted image, and denoised image, respectively. Finally, the fused image I_r is provided to the classifier f for classification.

We experimented with several denoising approaches and the corresponding results are outlined in Section 5.5. The most suitable technique was found to be the wavelet denoising presented in [23]. Therefore, we use this method in all other experiments.

5. Experiments

In this section, we will assess the performance of the proposed $CIIDefence$ and compare it with other well known defences. The evaluation will be done in terms of classification accuracy (i.e., top-1 prediction) as the adversarial defence can be considered successful when the system is able to correctly classify the input image. Following the common practice [23], we will concentrate on cases where the adversarial attack is applied to images that were originally correctly classified. The other cases, such as originally misclassfied images, would not provide a useful measure of the defence efficacy. However, we will evaluate the impact of the defence on the original classification accuracy (i.e., the case where no attack is applied).

The evaluation dataset contains 3500 randomly selected examples from the ImageNet [7] validation set. The evaluation images are further divided into two non-overlapping subparts referred as training and validation set. The training set consists of 500 images and it is used for parameter selection, analysing BPDA attack and for ablation studies. The remaining 3000 images make the validation set that is used for analysing the classification performance in the other experiments. Since only correctly classified images are considered, the classification accuracy of the original set is 100%.

Similar to most previous works [33, 32, 23], we concentrate on white-box setup (attacker has access to the network architecture) as opposed to black box attacks (attacker has no access to the network architecture). White-box attack is generally more challenging to defend as the attacker has access to the internal structure of the system. This choice is also supported by the recent results in [4] stating that a defence can be robustly evaluated by providing the complete knowledge of the defence to the adversary (i.e. white box setup). This is known as Kerckhoffs principle [3].

The $CIIDefence$ is not restricted to any specific image classification architecture. In our experiments, we consider ImageNet pre-trained VGG-16[1] [28], ResNet-101 [11][2], and Inception-v3 [29][3]. It is important to note that the same classifier should be used for obtaining the CAM responses for the defence and for the final image classification. Finally, we use Foolbox library[4] [24] to implement the adversarial attacks with normalised RMSE, $|L_2|$ equal to 0.03.

[1] https://www.cs.toronto.edu/ frossard/vgg16/vgg16_weights.npz
[2] http://download.tensorflow.org/models/resnet_v1_101_2016_08_28.tar.gz
[3] http://download.tensorflow.org/models/inception_v3_2016_08_28.tar.gz
[4] https://github.com/bethgelab/foolbox

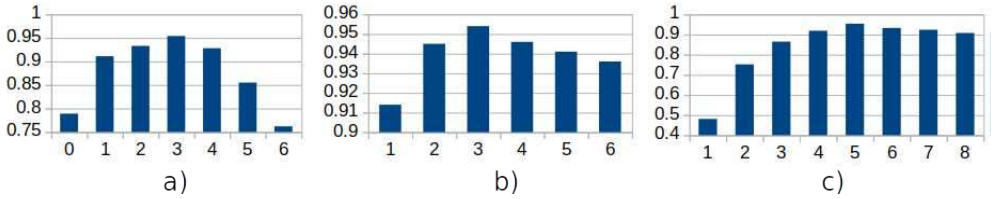

Figure 3. Top-1 classification accuracies, obtained by varying different hyperparameters: a) $\bar{p}$, b) w (rather than $(2w+1)$), and c) n. The x-axis corresponds to the hyperparameter value and y-axis denotes the classification accuracy. Note the different scales on y-axis.

5.1. Parameter Selection

The proposed $CIIDefence$ contains three hyperparameters related to image masking. These are $\bar{p}$, w and n which control the number of masked windows per class, the size of the masked image patch, and the number of considered top ranking classes, respectively. For the inpainting and denoising methods, we use the parameters from the original publications. Unfortunately, the optimal hyperparameters values depend on the specific adversarial attack. For instance, different L_p norm values and objective functions result in different characteristics in the adversarial images. Nevertheless, we should select a fixed set of hyperparameters since the specific adversarial attack is usually not known a-priori.

In this section, we analyse the effect of the hyperparameter selection using our training set of 500 images and pre-trained ResNet-101 classification network. We utilise line search for each hyperparameter value and evaluate the performance with respect to well known FGSM [9], IGSM [13], DFool [19], PGD [18] and C&W [5] attacks. For the C&W attack, we use only the L_2 loss. The results are obtained by setting the corresponding hyperparameter to the designated value and finding the maximum performance over the remaining hyperparameters. Figure 3 depicts the mean classification accuracies for different configurations.

Figure 3(a) illustrates the results over different values of $\bar{p}$ (number of masked windows per class). One can observe that the performance first quickly increases, then remains relatively stable, and later declines at higher values of $\bar{p}$. If $\bar{p}$ is very small, we apply the reconstruction only for a few locations, which might not be enough to mitigate the attack. In contrast, if $\bar{p}$ has high value, the area of the inpainted region grows and they may lead to large holes that are difficult to reconstruct. Figure 4 shows an example case with high $\bar{p}$ value. One can clearly observe blurring at the reconstructed areas. Overall, the results indicate that the best performance is obtained at $\bar{p} = 3$ and we use this value in the other experiments.

Figure 3(b) depicts the results for different values of w (size of masked image patch). The overall characteristics is similar to Figure 3(a), although the variation in the results is clearly smaller. One can still observe a clear trend of decreasing performance with increasing w. This is caused by the fact that larger areas are harder to reconstruct accurately. However, if the window size is too small, we might leave too much adversarial perturbations to the output. Figure 5 illustrates a few example reconstructions with different window sizes. Based on the results, we selected $w = 3$ (i.e. 7×7 window) to be used for the rest of the experiments.

Finally, Figure 3(c) illustrates the performance with respect to different values of n (number of considered top ranking classes). We assumed that the original correct label should remain within the top n classes after the adversarial attack. This hypothesis is supported by the observations in [23]. Therefore, if n is very small, the above assumption may not hold in all cases. On the other hand, when n increases, we end up masking (and reconstructing) larger parts of the image. Based on the results, we select $n = 5$ as a reasonable compromise. We will use this value for the rest of the experiments.

5.2. Performance Against Well-known Attacks

In this section, we assess the performance of the $CIIDefence$ against five well known attacks: FGSM [9], IGSM [13], DFool [19], PGD [18] and C&W [5]. The experiments with BPDA [2] attack are presented later in Section 5.4. Here we report the classification accuracies obtained using the test set (3000 images) and different classification networks. We also report the results in the cases where either no attack or no defence is applied. Table 5.2 summarises the results for the different combinations.

If no attack is applied, the $CIIDefence$ seems to result in a small decrease in the classification performance,

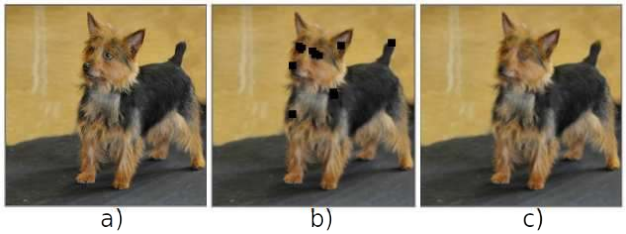

Figure 4. Example of spurious reconstruction caused by large $\bar{p}$ value (number of masked patches per class): a) the original image, b) image where the selected areas are removed; and c) the reconstructed image. The eyes are removed by the reconstruction since multiple masking windows have merged into a large hole covering the entire eye region.

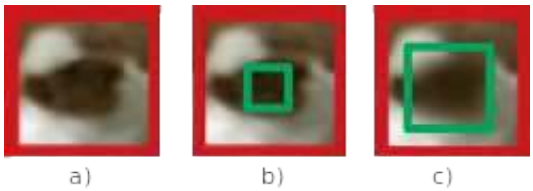

Figure 5. Spurious reconstruction caused by large window size w. a) Image area corresponding to the left eye in Figure 1. b) Result after inpainting the (7×7) block (green box). c) Result after inpainting the (13×13) block (green box). One can observe increased blurriness as the window size is increased.

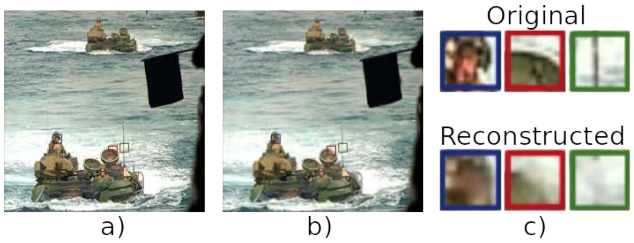

Figure 6. Example of a failure case caused by the class-specific image inpainting: a) the original image; b) fused output image; and c) blurry blocks resulting form the inpainting.

which is less than 2 percent points in all cases. When the adversarial attack is applied, one can observe further performance decrease. However, for the DFool, PGD, and C&W attacks the reduction is roughly one percent points (slightly more with VGG-16). The FGSM and IGSM seem to be the strongest attacks, but even in these cases the performance reduction is only up to 13 percent points, which is considerably better compared to the other state-of-the-art defences (see the next section). By analysing the failure cases, we found out that most of them are due to improper inpainting or denoising that produces blurry output images. Figure 6 illustrates a typical failure case with blurry inpainting result.

5.3. Comparison with State-of-the-art Defences

Most of the existing defences are primarily focusing on the performance with datasets containing a small number of classes and small image. Such method would be unsuitable for the experiments using ImageNet dataset. Therefore, we include only the following recent works, proposed originally for larger-scale images, in our comparison: feature squeezing [33]; randomized resizing and padding [32]; quilting and TVM [10]; PD and wavelet denoising [23]; and high level guided denoisers [14]. We use the implementations from the original authors, except for the robust activation maps in the pixel deflection that has no publicly available program code. For this part, we used our own implementation of the method.

The experiments were performed using the test set (3000 images) and the ResNet-101 classifier. We report the performance against FGSM [9], IGSM [13], DFool [19],and C&W [5] attacks. We also report the results when no attack is applied (clean images). For comparability, we report the result in terms of destruction rate that depicts the fraction of adversarial images that are correctly classified after applying the defence [23]. For instance, the destruction rate of 1 indicates that all the adversarial images are correctly classified due to the defence. We note that, ensemble training could further improve the efficacy of adversarial defences, but it was not considered to avoid classifier re-training.

Table 2 presents the results for different attack and defence combinations. Firstly, it can be observed that all defences decrease the original classification performance with clean images. However, the reduction seems to be smallest using the proposed $CIIDefence$. We believe that this is obtained by inpainting only small image blocks considering the local and global image characteristics. Furthermore, when the attacks are applied, the proposed defence provides clearly the best protection. The difference is most significant with FGSM and IGSM attacks.

5.4. Performance Against BPDA Attack

Modern defences usually include a non-differentiable operation that aims at preventing the attacker from obtaining the full gradient information using back propagation algorithm. Such approach is referred as gradient masking and it is very effective against many well known attacks. Similar operation is done in $CIIDefence$ by fusing the inpainted areas and the denoised image.

However, Athalye et al. [2] proposed an attack known as BPDA where the non-differentiable components are approximated with differentiable ones for obtaining approximations of the true gradients. It was shown in [2] and [1] that BPDA was able to breach several well-known defences in almost all the cases. BPDA is considered as one of the most effective attacks proposed so far. Therefore, we wanted to carefully analyse the performance of the proposed defence in the case of BPDA attack.

We apply the BPDA against the $CIIDefence$ by replacing the non-differentiable inpainting, denoising, and fusion layers with the identity function (for backward pass only). Such an approximation can be performed because the original and fused images are similar [2]. Moreover, we apply PGD along with the BPDA by setting the maximum number of iterations to 100.

The obtained results indicated that BPDA was able to circumvent $CIIDefence$ in 20% of the cases, when the maximum allowable intensity variation was set to 3/255. The results in [1] and [2] showed that the similar BPDA attack was able to completely breach the state-of-the-art defences presented in [32], [10], [23] and [14] (i.e. rendering the corresponding classification accuracy to 0%). Therefore, our results with the $CIIDefence$ are substantially better com-

Table 1. Performance of the proposed defence in against the well-known attack methods and without any attack.

Attack	VGG-16 No Defence[2]	VGG-16 Proposed	Inception-v3 No Defence[2]	Inception-v3 Proposed	ResNet-101 No Defence[2]	ResNet-101 Proposed
Original[1]	100%	98.7%	100%	98.1%	100%	98.6%
FGSM	18.5%	85.2%	19.2%	86.3%	18.9%	85.7%
IGSM	12.3%	92.9%	13.1%	93.6%	14.7%	93.8%
DFool	13.4%	96.3%	15.3%	97.1%	16.2%	97.6%
PGD	0%	98.1%	0%	97.6%	0%	97.8%
C&W	0%	96.5%	3.4%	97.9%	2.7%	98.2%

[1]When no attack is applied.
[2]When no defence is applied.

pared to the previous state-of-the-art defences.

Upon inspection, it was found that the failure cases arise when BPDA was able to substantially decrease the score and the ranking of the true class label. In such case, our reconstruction approach is not able to recognise and inpaint the relevant areas as the maximum number of considered classes was limited to n. Nevertheless, our total number of failure cases was significantly lower as compared to other defences.

5.5. Ablation Studies

In this section, analyse the contributions of the denoising and reconstruction components of our method. Furthermore, we compare them with JPEG [6] compression and image quilting technique [10], which could be considered as corresponding baseline methods, respectively. The analysis is performed using the training set of 500 images and the results reported in Table 3 in terms of top-1 classification accuracy. It can be observed that the best performance is achieved when using the proposed inpainting based reconstruction approach in combination with the wavelet denoising [23]. The comparison with quilting [10] method suggests that the proposed class-specific inpainting of small image areas is better option over full image reconstruction.

6. Conclusion and Future work

In this paper, we proposed a Class-specific Image Inpainting Defence ($CIIDefence$) that was shown to withstand several well-known powerful adversarial attacks including the BPDA [2]. The defence consisted of a class specific image reconstruction and an image denoising stages, which were later fused to form the transformed image. The proposed combination implements a non-differentiable layer that is not easily approximated with a simple differentiable alternative as seen in the BPDA experiments.

The image reconstruction is applied to small and carefully selected image areas most influential to the classification outcome. The selection procedure is guided by the class activation maps. Moreover, the $CIIDefence$ does not require retraining or modifications to the final classifier. The experiments, indicate that the $CIIDefence$ has minimal effect on the initial classification accuracy, while being excellent protection against the state-of-the-art adversarial attacks. According to the results, the $CIIDefence$ clearly outperform several recently proposed defence methods.

The possible future work includes improving the image denoiser and the inpainting techniques. In addition, one could increase the number of considered classes to further improve the performance agains the BPDA attack.

Table 2. Comparison with other well known defences in terms of destruction rate.

Defence	FGSM	IGSM	DFool	C&W	No Attack
[33]	0.449	0.687	0.798	0.927	0.938
[32]	0.467	0.712	0.956	0.972	0.941
[10]	0.642	0.895	0.866	0.871	0.789
[14]	0.514	0.742	0.948	0.978	0.946
[23]	0.638	0.803	0.819	0.856	0.963
Proposed	0.822	0.926	0.971	0.982	0.986

Acknowledgement

Funding for this research was provided by the Academy of Finland project number 324346.

Table 3. The results of the ablation of our method study.

Denoiser	JPEG	JPEG	None	[23]	None	[23][1]
Inpainting	No	Yes	No	No	Yes	Yes[1]
Quilting	No	No	Yes	No	No	No[1]
Original	93.6%	95.8%	74.2%	95.4%	98.2%	99.2%
FGSM	35.8%	81.6%	68.4%	39.2%	78.2%	87.6%
IGSM	34.4%	89.9%	67.2%	30.4%	83.6%	93.8%
DFool	61.2%	91.6%	65.8%	58.6%	88.8%	97.8%
C&W	83.2%	96.4%	68.6%	94.2%	97.4%	98.4%

[1]Used in the proposed $CIIDefence$ method.

References

[1] Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018. 7

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018. 1, 2, 6, 7, 8

[3] Kerckhoffs Auguste. La cryptographie militaire. *Journal des sciences militaires*, 9:538, 1883. 5

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 5

[5] Nicholas Carlini and David Wagner. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017. 2, 3, 6, 7

[6] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017. 2, 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5

[8] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017. 1

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. 2, 6, 7

[10] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 2, 3, 4, 7, 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision*, pages 262–270, 2015. 4

[13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 2, 6, 7

[14] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Xiaolin Hu. Defense against adversarial attacks using high-level representation guided denoiser. pages 1778–1787, 2018. 2, 7, 8

[15] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4

[16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 1

[17] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015. 3

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2018. 2, 6

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2, 6, 7

[20] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 4

[21] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017. 1

[22] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. 1

[23] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018. 3, 4, 5, 6, 7, 8

[24] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017. 5

[25] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. 2018. 3

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, pages 618–626, 2017. 4

[27] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: adversarial eliminating with gan. *arXiv preprint arXiv:1707.05474*, 2017. 3

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[30] 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. 1

[31] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1

[32] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 2, 5, 7, 8

[33] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 2, 5, 7, 8

[34] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. pages 5505–5514, 2018. 5

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 4

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 4

[37] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886. IEEE, 2012. 4