Scientific Research Publishing

Journal of Software Engineering and Applications, 2021, 14, 110-132
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Robustness Augmentation of Deep Learning Model Based on Pixel Change

Yu Zhang[1], **Hexin Cai**[2]

[1]College of Information Science and Technology, Jinan University, Guangzhou, China
[2]College Mathematics and Information, South China Agricultural University, Guangzhou, China
Email: zhaywine@163.com

How to cite this paper: Zhang, Y. and Cai, H.X. (2021) Robustness Augmentation of Deep Learning Model Based on Pixel Change. *Journal of Software Engineering and Applications*, **14**, 110-132.
https://doi.org/10.4236/jsea.2021.144008

Received: March 15, 2021
Accepted: April 25, 2021
Published: April 28, 2021

Abstract

Deep learning has been widely used in many fields. A large number of images can be quickly recognized by the deep learning models to provide information. How to improve the robustness of deep learning applications has become the focus of research. Unfortunately, the recognition ability of the existing deep learning model has been greatly threatened, many images can cause recognition errors in a well-trained model. Although data augmentation is an effective method, the existence of adversarial examples shows that traditional data augmentation methods have no obvious effect on minor pixel changes. After analyzing the impact of pixel changes on model recognition accuracy, a data augmentation method based on a small number of pixel changes is proposed. Our method can optimize the corresponding classification boundary and improve the recognition robustness of the model. Finally, a simple evaluation method to measure the robustness of model recognition is proposed. Our experiments prove the threat of a small number of pixels and the effectiveness of our data augmentation method. Moreover, the data augmentation method has strong generalization ability and can be applied to image recognition in many different fields.

Keywords

Data Augmentation, Deep Learning, Classification Boundary, Small-Distance Samples

1. Introduction

Deep learning [1] [2] [3] technology has achieved gratifying results in many fields [4] [5] [6]. For example, in the fields of smart manufacturing [7] [8] [9] [10] [11], the classifier based on deep learning models can quickly output correct recognition results. For institutions that need to process a large number of images, deep learning greatly improves the efficiency of their research and decision-making. The particularity of smart manufacturing requires deep learning applications has higher requirements for recognition robustness.

In common, the recognition robustness of a deep learning model can be characterized by the recognition accuracy [12]. A model with good recognition robustness usually has high recognition accuracy, and the probability of misclassification is low. A model with good recognition robustness should output the same recognition results for similar images as shown in **Figure 1**. Unfortunately, the current deep learning models are not satisfactory in this regard. Goswami, *et al.* point out that face recognition is susceptible to distortion [13]. A large number of adversarial examples [14] studies have also proved that adding noise to the image can effectively affect the recognition results of the models. This proves that the change of pixels will have a huge impact on the recognition effect of the model. In addition, we found that some deep learning models trained on large data sets are very sensitive to subtle changes in pixels. In our robustness tests on several classic network models, we found that when about 0.12% of the pixels in the image are changed, the misclassified image data far exceeds this ratio. Among them, in the best-performing model, about 3.5% of the images had classification error; in the worst-performing model, about 14% of the images had classification error. The pixel change process has not undergone any careful calculation. This means that malicious people can easily generate a batch of images to influence commercial applications based on these classifiers.

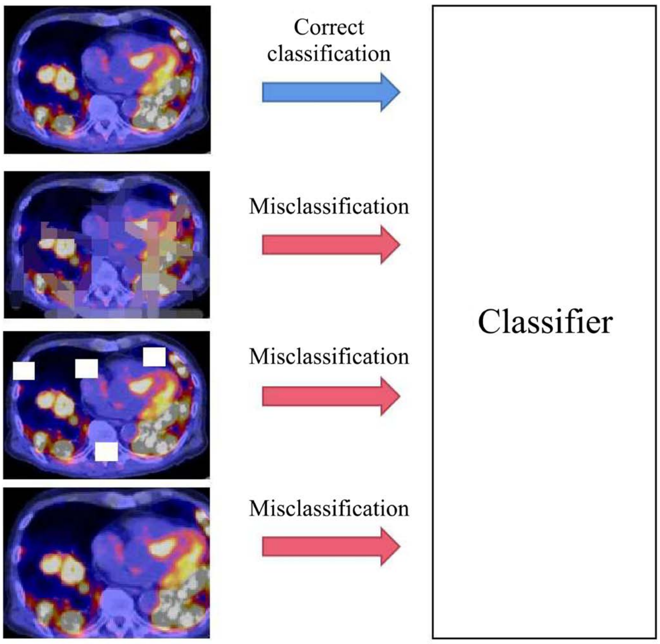


Figure 1. Schematic diagram of model robustness. For this PET image, the model with recognition robustness can normally recognize its original image, but cannot recognize other similar images. For many models, when an image that can be recognized normally is blurred, compressed, occluded, etc., there is a high probability that it will not be recognized.

Robust training is an effective and commonly used method to provide robust optimization methods for deep learning models. By adding different images in the training set, the model can learn the differences between images during training and optimize the final model parameters. Data augmentation methods are usually used in the robust training process to enhance the training effect. However, traditional data augmentation methods [15] [16] often make too many changes to the image, resulting in the model being still sensitive to small changes in pixels. After a detailed investigation of the impact of pixel changes on the robustness of model recognition, we propose a data augmentation method based on a small number of pixel changes to generate small-distance samples with less difference from the original image. Those samples are closer to the original image than the samples produced by other data augmentation methods and can optimize the classification boundary of the model during training and help the model learn the relationship between the main pixels in the image.

Finally, considering the complexity of evaluating the robustness of existing models, we propose a method for evaluating the robustness of deep learning model recognition based on experimental results. This method can easily and effectively evaluate the difference in recognition robustness of different models. Provide researchers with more parameter indicators and model optimization directions.

The main contributions and innovations of this work are summarized as follows:

1) Explored the relationship between classifier robustness and pixel changes. Summarized the factors that can greatly affect the classification accuracy in pixel changes;

2) A method for evaluating the robustness of a classifier based on pixel changes is proposed. This method can evaluate the robustness of the classifier without knowing the internal structure of the classifier. So the user can easily and effectively find the deep learning applications of smart medicine with better recognition robustness;

3) A data augmentation method based on a small number of pixel changes is proposed. This method can effectively improve the robustness of the deep learning applications and can provide defense capabilities for adversarial examples. This can prevent malicious users from attacking the model and reduce the occurrence of accidents.

2. Background

This section mainly introduces the related models and some related concepts, including five classic deep learning models, adversarial examples, and data augmentation methods.

2.1. Deep Learning Models

Deep learning models often have higher training costs because of the deeper hidden layers. In the training process, the training of large data sets often takes a long time. To reduce the cost of training and improve the recognition effect, many users will choose to use pre-trained models. The pre-training models are models that have been well-trained on the large data set. They have good recognition ability for normal images in the data set. Users can simply obtain the corresponding model parameters from the Internet for subsequent training or research.

To explore the impact of small pixel changes on the deep learning model classifier, we selected the following five classic deep learning pre-trained models as the aims of the research as shown in **Table 1**. All five models have been fully trained on the large dataset ImageNet [17]. Without adding noise to the input image, these models all have very good recognition effects[1].

These five deep learning models have good recognition capabilities in image recognition and are widely used in various image recognition fields as basic models. In general, the problems of these 5 models also occur in most image recognition models.

2.2. Adversarial Examples

Szegedy [23] proposed the concept of adversarial example in 2013. The adversarial example is a kind of image that can make the deep learning model output wrong classification results with high confidence. The typical adversarial example does not appear to be significantly different from the normal image. In smart manufacturing, if a malicious user replaces normal manufacturing images with adversarial examples [24], it will cause unfortunate consequences. It is difficult for workers to detect these replaced pictures with the naked eye.

Adversarial examples can significantly affect the recognition ability of deep learning models, which has attracted wide attention from researchers [14] [25] [26]. In general, the adversarial attacks can be categorised into black box attacks [27] and white box attacks [28] based on whether the adversary knows the internal information of the target model. In black-box attacks, the attackers cannot get the relevant internal information of the model. In white-box attacks, the attackers have full access to the model's structure and parameters.

Table 1. Introduction of five pre-trained models.

Model	Description	Recognition rate
ResNet [18]	2015 Imagenet recognition competition champion	96.43%
DesNet [19]	Optimization of Resnet model	94.17%
VGG [20]	2014 Imagenet recognition competition runner-up	93.20%
AlexNet [21]	2012 Imagenet recognition competition champion	80.30%
SqueezeNet [22]	Similar to the effect of Alexnet, but the structure is simplified	80.30%

[1] For the sake of authenticity and the difference of experimental equipment, the recognition accuracy of these five classic deep learning models come from their related papers and competition data. Among them, the competition results of ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) are often referred to, which has very high reliability.

Based on the goal of attackers, the adversarial attacks can be divided into targeted attacks and non-targeted attacks. The targeted attack means that the attacker hopes that the generated image can point to a particular category. In non-targeted attacks, the generated adversarial examples only need to be classified incorrectly. A summary of widely used adversarial attack techniques is provided in **Table 2**.

Most of the disturbance noise of the adversarial example is distributed on the overall pixels of the image, and the attacker uses complex calculations to make these disturbance noises as small as possible. But for the deep learning model, the pixel changes of the image are very obvious, because the machine can easily get the specific value of each pixel. The One-pixel attack proves that even if only one pixel is modified, classification errors can occur in the deep learning model.

As a kind of image that can cause the classification error of the deep learning model, the adversarial example is extremely threatening to the recognition robustness of the model. Compared with the change of a few pixels in the image, the pixel change range of the adversarial example is larger and the calculation is more precise. From the perspective of pixel distribution, both applying noise to the whole image and changing just 1 pixel can make the model identify errors. Exploring the effects of pixel changes including distribution and range on the model can provide more information for improving the robustness of the model and defending adversarial examples.

2.3. Data Augmentation

Data augmentation [15] can effectively improve the robustness of deep learning models [16] [34]. By expanding the data in the training set, a large number of images with training significance can be quickly obtained. These obtained images have correct labels and can supplement the unity of features in the original training set images. As shown in **Table 3**, commonly used data augmentation methods can be divided into the following categories: geometric-based modification methods; color-based modification methods; data augmentation methods based on machine learning methods.

Table 2. Introduction of classic adversarial attacks.

Method	Target	Black/White Box	Noise distribution
FGSM [14]	targeted	white box	Whole image
DeepFool [29]	targeted	white box	Whole image
C&W [30]	targeted	white box	Whole image
BIM [25]	targeted	white box	Whole image
JSMA [31]	targeted	white box	Whole image
ATNs [32]	targeted	white box	Whole image
One Pixel [33]	non-targeted	black box	1 pixel or very few pixels

Table 3. Three types of data augmentation methods. Among them, the geometric-based data augmentation method mostly changes the image globally, which can quickly generate a large number of images. Color-based data augmentation methods change the pixel values in the image. Data augmentation methods based on machine learning can purposely generate images that users expect to generate.

Method	Description
Geometric-based modification methods	Perform geometric transformations on images, including flipping, rotating, cropping, deforming, scaling, etc.
Color-based modification methods	Change the content of the image itself, including noise, blur, color change, erasure, fill, etc.
Machine learning methods	Use machine learning methods to generate images purposefully [35] [36] [37], such as using GAN [38] (generative adversarial networks) to generate confrontation images to expand the training set.

Most of the current data augmentation methods greatly change the image. For deep learning models that have been trained on a large number of images, the images generated by these methods are very different from the original images within the model, and the relationship between pixels has been completely disrupted. It is difficult for the classifier to learn the impact of a few pixel changes on the image through these images. Traditional data augmentation methods have greatly changed the image so that the model cannot learn the correlation between normal pixels.

For the model to learn the effects of pixel changes, a data augmentation method with fewer changes is necessary. Because a small amount of change means that the main pixel information of the image still exists, the classifier can learn the main information of the image through repeated training, thereby optimizing its classification boundary, improving classification accuracy and recognition robustness.

3. The Effect of Few Pixel Changes on Model Recognition Robustness

The recognition robustness of a deep learning model can be measured by its recognition ability. Good recognition robustness means that the model can make the same judgment on similar images. However, the recognition robustness of deep learning models has been seriously threatened. Recognition errors of some images have appeared in deep learning applications in many situations, and there is almost no difference between these images and the images that can be correctly recognized.

Most of the pixels changed by the adversarial attack have undergone complex calculations and are distributed across the entire image, which makes the generated adversarial example and the original image quite different. From this perspective, it seems to be enough to make the model misclassify. However, the existence of One-pixel attacks proves that only modifying one pixel is enough to make a well-trained deep learning model recognize errors.

On the other hand, although the attack ability of adversarial examples is very powerful, it is also extremely vulnerable to various factors that cause the attack to fail. Research on the robustness of adversarial attacks shows that many adversarial examples lose their attack ability after the pixel changes [39] [40], which makes the deep learning model make correct judgments.

Based on this fact, two issues will become the focus of our discussion:

1) Why does the random modification of a very small number of pixels will make a well-trained deep learning model misclassify?

2) Why does the adversarial example generated by careful calculation will lose the ability to attack after modifying a few pixels?

For the first question, the main reason is that the existing deep learning models are not robustly trained. After the model accepts an image with pixel changes, the differences will be magnified inside the model, which will affect the final classification results. The robust training will make the model has a higher tolerance for pixel changes. From the perspective of clustering as shown in **Figure 2**, the robust training model has a better classification boundary, and more different but similar images will be clustered together.

For the second question, the adversarial examples will be correctly recognized because the model is well-trained. A model well-trained means the model has a good ability to recognize. The pixel changes can affect the disturbance noise's attack ability so that a well-trained model will output the correct result. From the perspective of clustering, as shown in **Figure 3**, random pixel changes can make the adversarial example return to the correct classification interval.

As shown in **Figure 4**, good robustness training can get a better classification boundary, thereby reducing the impact of random pixel changes and improving the robustness against attacks.

Therefore, improving the recognition robustness of a deep learning model means to optimize the classification boundary for the model so that it can limit the changes caused by a small number of pixel changes within the classification boundary. At the same time, due to the optimization of the classification boundary, for adversarial samples that also rely on pixel changes, the model can also have a certain defensive ability, so that some adversarial samples cannot cross the classification boundary.

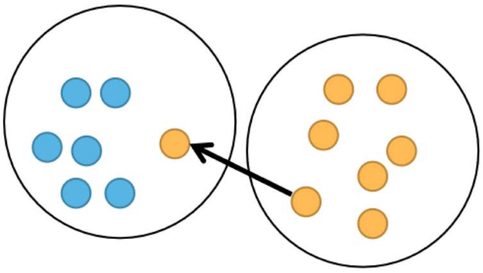


Figure 2. Schematic diagram of the impact of normal image pixel changes on classification. The change of the representative pixel here causes the recognition change within the deep learning model, and the difference is magnified so that the changed image crosses the implicit classification boundary, resulting in a wrong classification.

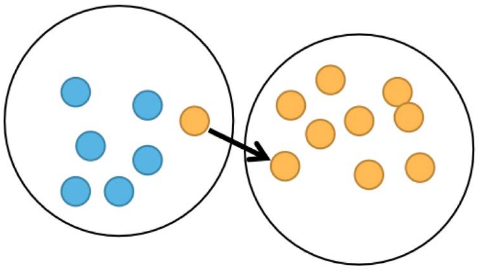


Figure 3. Schematic diagram of the impact of the change of the adversarial sample pixel on the classification. The change of the representative pixel here causes the internal recognition change of the deep learning model, and the original attack ability of the adversarial example is affected so that the changed image is restored to the correct classification.

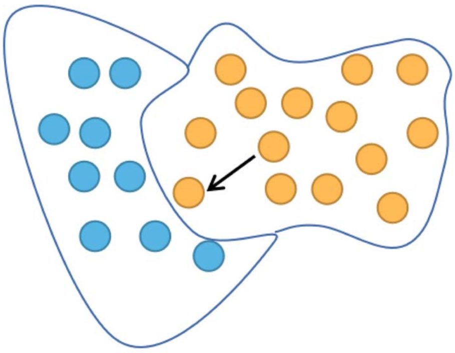


Figure 4. Model classification diagram with good recognition robustness. The model has a good implicit classification boundary, which can keep the image within the classification boundary after changes. In this way, the classification of the changed image will not change.

4. Data Augmentation Method

To improve the recognition robustness of existing deep learning models, we propose a data augmentation method with smaller differences from the original image. During the training process, images similar to the training set images will be added to enhance the recognition accuracy and recognition robustness of the final model. In the previous article, the implicit classification boundary in the deep learning model was mentioned. Our data augmentation method can effectively optimize the corresponding classification boundary. This process can be illustrated by a binary classification problem.

For a normal image **x** and its corresponding classification label **y**, assuming the corresponding classification boundary is a straight line. The classification straight line $\mathcal{F}$ obtained by the deep learning training process can be defined as:

$$\mathcal{F} : \mathbf{w}^{\mathrm{T}} x + b = 0 \tag{1}$$

where **w** is the relevant classification boundary parameter obtained after deep learning training. When the value of $f\left(x_0\right)=w^{\mathrm{T}}x_0+b>0$, we consider the label of the sample x_0 to be y, and when the value of $f\left(x_0\right)=w^{\mathrm{T}}x_0+b<0$, we consider the label of the sample x_0 to be the remaining labels. The prediction process can be summarized as:

$$\text{label}=\begin{cases}\mathbf{y} & f\left(x\right)>0\\ \text{other} & f\left(x\right)<0\end{cases} \tag{2}$$

From the formula for the distance from a point to a straight line, the distance **r** of the sample x_0 from the classification boundary $\mathcal{F}$ is:

$$\mathbf{r}=-\frac{f\left(x_0\right)}{\left\|\mathbf{w}\right\|_2^2}\mathbf{w} \tag{3}$$

When a small number of pixels on the image change, although the difference is small, the modified sample crosses the classification boundary $\mathcal{F}$, causing the model to output incorrect recognition results as shown in Figure 5.

Our data augmentation method can produce such images with fewer differences, these images are very similar to the original image, only a few pixels are different. Retraining with these small-distance samples can optimize the corresponding classification boundary and increase the distance between the sample and the classification boundary.

Figure 6 shows the basic effect of our program. After training with small-distance samples, the relevant parameters in the model have changed:

$$\mathcal{F}':\mathbf{w}_{new}^{\mathrm{T}}x+b_{new}=0 \tag{4}$$

$$\mathbf{r}'=-\frac{f\left(x_0\right)_{new}}{\left\|\mathbf{w}_{new}\right\|_2^2}\mathbf{w}_{new} \tag{5}$$

$$\mathbf{r}'>\mathbf{r} \tag{6}$$

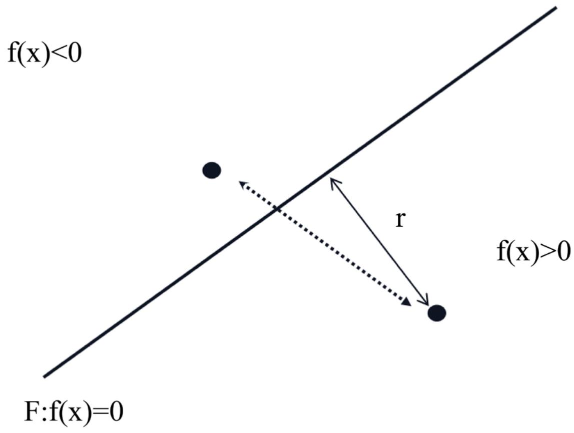


Figure 5. The classification boundary of the two categories. The distance between the sample and the classification boundary is **r**. When the sample is changed so that the predicted distribution of the sample changes by more than the distance **r**, the predicted result of the sample changes.

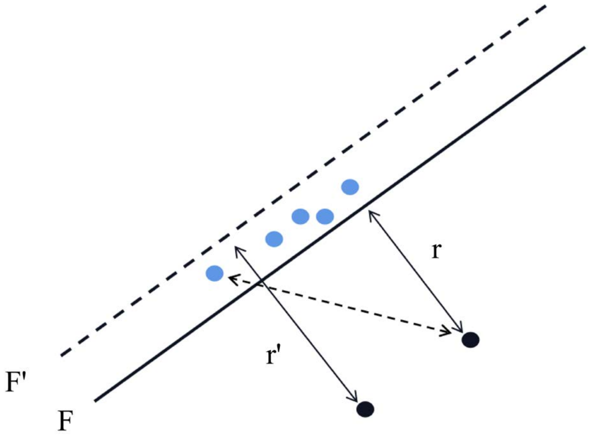


Figure 6. The classification boundary after robustness training. The blue sample points are small-distance samples obtained by modifying the original sample with a few pixels change. The new classification boundary can be obtained by retraining the blue sample points.

A new classification boundary $\mathcal{F}'$ can be obtained. The classification boundary $\mathcal{F}'$ can correctly distinguish the original sample and the small distance sample. And the distance $\mathbf{r}'$ from the original sample to the new classification boundary $\mathcal{F}$ is increased, so that the new model can accept more modifications to the original image.

The image produced by the traditional data augmentation method will have a large change at the pixel level (Gaussian noise, stretching, inversion, and other operations can change the pixels on each pixel in the original image). These changes are amplified within the model and ultimately affect the subsequent classification. The images produced by the traditional data augmentation method have similarities for humans, but they do not have similarities within the deep learning model. Our data augmentation method can fill this gap by the small-distance samples. The generated samples not only has similarity for humans but also similar to the normal image in the deep learning model. This method can strengthen the learning ability of the deep learning model for normal image classification and reduce the recognition errors caused by pixel changes.

The situation of multiple classifications is similar to that of two classifications. There is a classification boundary between a certain type of image and different types of images, but due to the complexity of deep learning, it is difficult for us to know the specific location of the boundary. Through our small pixel data augmentation method, an image similar to the original image can be input into the model, so that the model has a more accurate classification boundary. This boundary can distinguish the influence of a small number of pixel changes, and still classify the image with pixel changes as the original image, thereby improving the recognition robustness of the model.

The process of our data augmentation program is as follows in algorithm 7. The main process of this algorithm can be summarized as follows: randomly generate the pixel positions in the image that need to be changed, and then randomly modify the pixels at the corresponding positions on the image. This change will not affect the pixels outside the modified position, and the resulting image is visually similar to the original image.

Algorithm 1: Data augmentation method based on few pixel changes

```
Input: Training data set Q,the number of
       changed images Num, number of pixel
       changes n, pixel change range R, pixel
       change position P
Output: The images with few pixel changes
position list=[];
num=0 ;
for $i$ in range($n$) do
    x=random int(p) ;
    y=random int(p) ;
    position list $\leftarrow$ (x,y) ;
end
for image in $Q$ do
    if num=Num then
        Quit ;
    end
    for (x,y) in position list do
        image[0,x,y]=random int(R) ;
    end
    Num+1 ;

end
```

The purpose of this algorithm is to randomly modify a few pixels in each image. The **Num** is used to control the number of modified pixels. The previous experiments have proved that modifying 0.12% of pixels can cause 10% of images to be classified incorrectly. The value of Num should be distinguished according to different training needs. We suggest that if there is no special requirement, the Num can be chosen as 64 for the 224×224 image. The **R** can control the magnitude of pixel change, and The **P** can limit the range of pixel modification.

After the small-distance samples corresponding to each category in the training set are generated, a new model can be obtained through small-scale retraining. Our experiments prove the feasibility and effect of this method. Figure 7 briefly summarizes this process. At the same time, due to the randomness of the samples, this method usually needs to consume computing resources to generate enough small-distance samples.

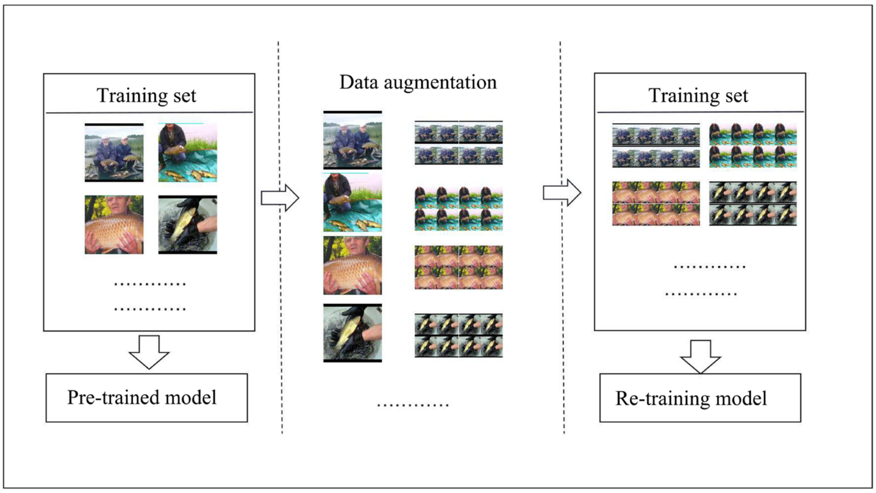


Figure 7. Schematic diagram of data augmentation and training. In the data augmentation stage, the images in the training set are randomly taken out, and corresponding small-distance samples are generated as needed. These small-distance samples are used as training images to retrain the model.

5. Robustness Evaluation Standard

For a deep learning model, we simply define its robustness evaluation direction as a model with good recognition robustness, the number of misclassified images should be positively correlated with the number, amplitude, and range of pixel changes. The fewer the pixel changes, the fewer the number of misclassified images. In the case of image classification error, the more pixel changes involved in cause the image classification error, the higher the recognition robustness of the model.

Based on this definition, to better analyze the recognition errors caused by pixel changes, we propose the Pixel-index (can be referred to as Pi) to measure the robustness of the deep learning model. The calculation formula of Pi is:

$$\text{Pi} = \frac{w}{p \times r \times a} \tag{7}$$

The parameters used are summarized as follows in Table 4.

As an example, for 10,000 images with a size of 224 × 224, when the number of modified pixels is 64, the value range of modified pixels is 0 to 255, and the position range is 200 × 200, 1000 images are incorrectly recognized. Then the corresponding pi value is calculated as follows:

$$w = 1000/10000 = 10\% \tag{8}$$

$$p = 64/(224 \times 224) = 0.1276\% \tag{9}$$

$$r = 255/255 = 100\% \tag{10}$$

$$a = (200 \times 200)/(224 \times 224) = 79.72\% \tag{11}$$

$$\text{Pi} = \frac{10\%}{0.1276\% \times 100\% \times 79.72\%} = 98.30 \tag{12}$$

Table 4. Pixel-index parameter list.

Parameter	Description
w	Percentage of images with incorrect recognition
p	Proportion of modified pixels
r	Percentage of pixel modification value range
a	Percentage of pixel modified location range

Pi focuses on the pixel changes of the image rather than the internal structure of the model, and can faithfully reflect the sensitivity of the model to changes in the image pixels. The pixel change of the image includes the number of changed pixels, the changed pixels value range, and the changed pixels position range. The smaller the PI means that the deep learning model is less sensitive to changes in pixels, and its recognition robustness is higher. Because the previous experiment has proved that the number of pixels modified, the size is positively correlated with the wrong recognized image number. A small PI means that more pixels need to be modified for the same number of wrong images. Pi is also suitable for adversarial examples. Because the attack ability of the adversarial example also comes from the changes to the pixels, but the number of pixels and the location range of the changes will be different from random changes.

Comparing the pi of different models requires the calculation of the same batch of images. Table 5 shows the pi values of several deep learning models on the same 10,000 images. It can be seen that the deep learning model with smaller Pi has better recognition robustness and higher recognition accuracy. The pi calculation does not need to know the internal structure of the model in advance, only needs the data set test. The user can simply use his (her) own test data to quickly obtain the robustness comparison of different models.

6. Experiment

The experiment environment is a Linux server with 4 GPUs, the system is 64-bit Ubuntu 18.04.4LTS, the memory is 256 G, the processor is Intel Xeon(R) CPU E5-2683 V3 @ 2.00 GHz, the GPU model is GTX 1080ti, and the video memory is 12 GB. The five pre-training models are provided by Torchvision. Torchvision is a well-known toolset for processing the image and video in Python and has a very wide range of applications in deep learning. The images used in the experiment are all from the ImageNet dataset, and it can be ensured that each image participating in the experiment can be correctly recognized by these five pre-training models before pixels change.

6.1. Robustness Experiments for Deep Learning Models

6.1.1. Experiment 1: The Impact of Pixel Changes on Deep Learning Models

This experiment explored the impact of pixel changes on the accuracy of deep learning models. We selected 10,000 images in the ImageNet dataset, and these images can be correctly classified by those models selected in the experiment. We randomly modified a different number of pixels on each image, and the final experimental results are as follows.

Table 6 shows that when the position and the range of pixel changes are not limited, a very small amount of pixel changes can cause a large number of images misclassified. For Imagenet images with a size of 224 × 224, 4 pixels only account for 0.0080% of the total pixels of the image, but it has been able to cause about 1% of the images to have classification errors in Resnet. And when the number of pixel modifications is increased to 64 (0.12%), about 5% of images will be classified incorrectly in Resnet. And these changes did not only occur on Resnet, the other 4 models also showed sensitivity to pixel changes. Figure 8 shows the changing trend of accuracy and wrong images. This trend means that when more pixels are changed, more images will have recognition errors.

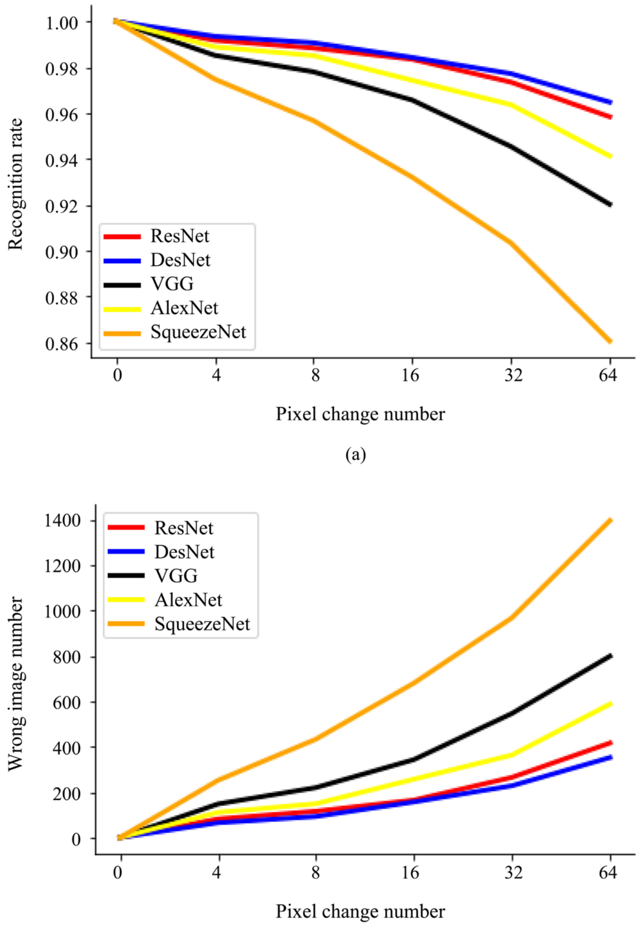
In this experiment, the number of pixel changes is limited to 64 pixels or less. For normal images, the number of changed pixels is very small. At the same time, these pixel changes are not carefully calculated, but random changes. Although the changes have been very subtle, it is difficult to ignore the image data with classification errors, which shows that the existing deep learning models are very sensitive to pixel changes.

Table 5. Robustness comparison of Pi-based classic models.

Model	Pi	wrong images	Recognition rate
Desnet	27.58	352	94.17%
Resnet	32.60	416	96.43%
Alexnet	46.00	587	93.20%
VGG	62.53	798	80.30%
Squeezenet	109.32	1395	80.30%

Table 6. The effect of pixel changes on the recognition accuracy of the deep learning models. The figure shows the number and proportion of pixel changes. In the experiments on five classic deep learning models, it can be observed that Desnet performs best (the best robustness), and Squeezenet performs the worst (the worst robustness).

Pixel changes		4 0.0080%	8 0.0159%	16 0.0319%	32 0.0638%	64 0.1276%
ResNet	Misclassification	82	116	165	265	416
	Recognition rate	99.18%	98.84%	98.35%	97.35%	95.84%
DesNet	Misclassification	65	93	157	228	352
	Recognition rate	99.35%	99.07%	98.43%	97.72%	96.48%
VGG	Misclassification	148	220	343	546	798
	Recognition rate	98.52%	97.8%	96.57%	94.54%	92.02%
AlexNet	Misclassification	111	149	257	363	587
	Recognition rate	98.89%	98.51%	97.43%	96.37%	94.13%
SqueezeNet	Misclassification	252	433	680	967	1395
	Recognition rate	97.48%	95.67%	93.2%	90.33%	86.05%



(a)

(b)

Figure 8. The changes in recognition accuracy and number of false images. When more pixels are changed, more images will have recognition errors. (a) The recognition accuracy of deep learning models; (b) The wrong images number of deep learning models.

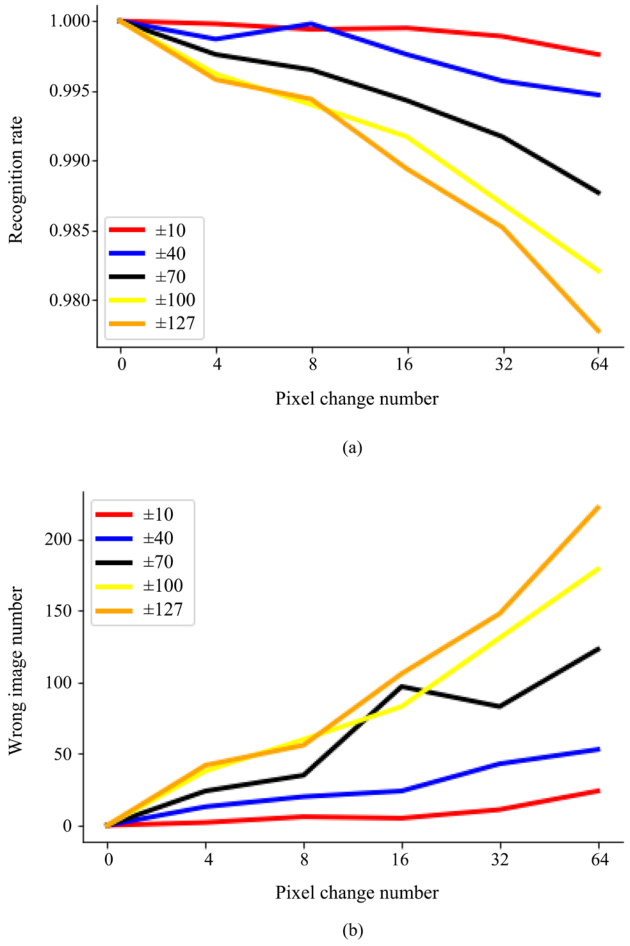
6.1.2. Experiment 2: The Impact of Pixel Changes Size on Deep Learning Models

In this experiment, we explored the effect of the magnitude of change in pixel changes. The selected images are also 10,000 images from ImageNet that can be correctly identified. But when making pixel changes, we limit the amplitude of each change, so that the change in pixel value is centered on the original value, and random changes of different amplitudes are made. The final experimental results are as follows.

Table 7 summarizes the impact of different pixel change range on the deep learning models and Figure 9 shows the trend of model accuracy and the number of error images. It can be seen that when the pixel changes are small, the number of misclassified images is also small. The number of images that are misclassified at the same time will increase as the magnitude of pixel changes increases. There are obvious differences in the number of error images caused by different changes. This shows that for a deep learning model, the magnitude of pixel change is an important factor affecting its classification accuracy.

Table 7. The effect of pixel changes size on the recognition accuracy of the deep learning model. Each experiment from top to bottom limits the size of the pixel changes. Centering on the value of the pixel itself, increase or decrease a random value each time. The intervals of change are [−10, +10], [−40, +40], [−70, +70], [−100, +100], [−127, +127]. When the value of the pixel exceeds the normal range, it will be limited to the boundary value 0 or 255.

	Pixel changes	4 0.0080%	8 0.0159%	16 0.0319%	32 0.0638%	64 0.1276%
10	Misclassification	2	6	5	11	24
	Recognition rate	99.98%	99.94%	99.95%	99.89%	99.76%
40	Misclassification	13	20	24	43	53
	Recognition rate	99.87%	99.80%	99.76%	99.57%	99.47%
70	Misclassification	24	35	97	83	123
	Recognition rate	99.76%	99.65%	99.43%	99.17%	98.77%
100	Misclassification	38	60	83	131	179
	Recognition rate	99.62%	99.40%	99.17%	98.69%	98.21%
127	Misclassification	42	56	106	148	222
	Recognition rate	99.58%	99.44%	98.94%	98.52%	97.78%



(a)

(b)

Figure 9. The changes in recognition accuracy and number of false images. It can be seen that there are obvious changes, and there is a big difference between the five curves. (a) The recognition accuracy of deep learning models; (b) The wrong images number of deep learning models.

6.1.3. Experiment 3: The Impact of Pixel Changes Area on Deep Learning Models

In this experiment, we explored the effect of the position of the change in pixel changes. The selected images are also 10,000 images from ImageNet that can be correctly identified. But when making pixel changes, we limited the position range of pixel changes, gradually narrowing from the full image to the center position. In general, the main information of an image is located in the center of the image. We hope that this experiment can explore the impact of pixels at different locations on the results of deep learning classification. The final experimental results are as follows.

Table 8 summarizes the impact of different pixel change position on the deep learning models and Figure 10 shows the trend of model accuracy and the number of error images. Different from the previous experimental results, in this experiment, it can be seen that different positions have little effect on the classification results of the classifier. In Figure 10, it can be seen that when the number of changed pixels is less than 64, the number of images with classification errors is relatively close for different location ranges. This means that when the number of pixel changes is small, the location of the change does not play a significant role.

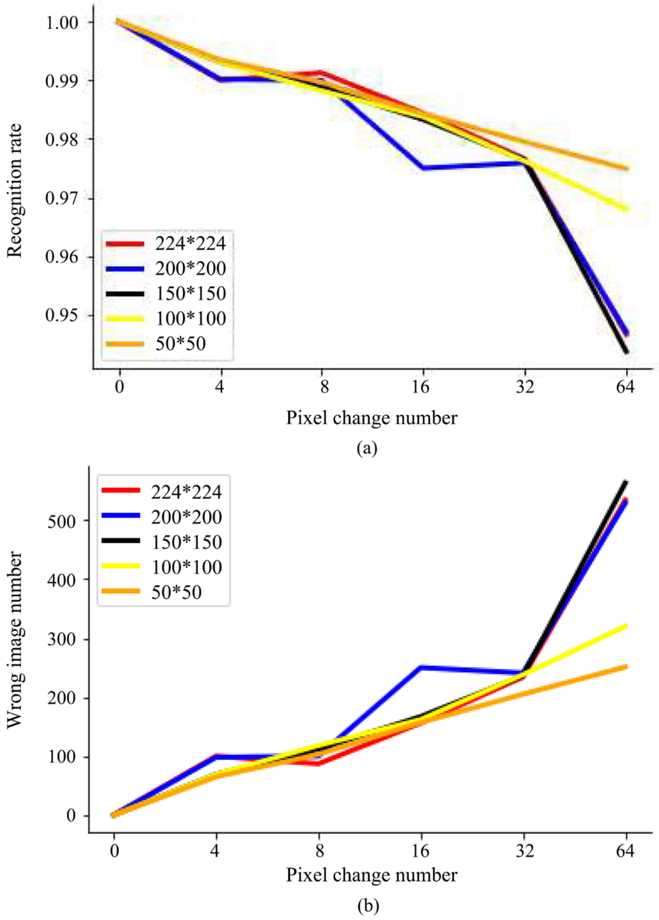
6.1.4. Experiment 4: The Impact of Pixel Changes on Adversarial Examples

The first three experiments proved that pixel changes can cause classification errors of normal images. This experiment will explore whether changes in few pixels can change the recognition results of abnormal images. The experimental results after pixel changes are as follows.

We chose a very typical FGSM method (Fast Gradient Sign Method) [14] to generate adversarial examples. The 10,000 images participating in this experiment are all generated by this method and will be recognized as a different category from the original image under the classifier. It can be seen from Table 9 that when a small number of pixel changes are made to this type of image, a considerable proportion of the image will lose the attack ability and return to the correct classification.

On the one hand, this experiment shows that the adversarial examples are also sensitive to pixel changes. On the other hand, it also shows that for the deep learning model, pixel changes will greatly change the calculation process of each internal layer, amplify the changes, and change the final Recognition results.

Whether it is a normal image or an adversarial example, the classification results have changed after a small number of pixel changes. This shows that the existing deep learning models are very sensitive to pixel changes. For the pixel change, the magnitude of the pixel change can affect the classification result of the deep learning classifier more than the position of the pixel change. For adversarial examples, the recognition result of such images is originally the wrong classification result. A small number of pixel changes can restore the classification results of such images to normal classification.



(a)

(b)

Figure 10. The changes in recognition accuracy and number of false images. It can be seen that when the abscissa is 8 and 32, the ordinates of the five curves are very similar. (a) The recognition accuracy of deep learning models; (b) The wrong images number of deep learning models.

Table 8. The effect of pixel changes on the recognition accuracy of the deep learning model. Each experiment from top to bottom limits the position of the pixel changes. Taking the size of the image as the starting point, each experiment will reduce the range of pixel changes and maintain center symmetry. Taking the size of the image as the starting point, each experiment will reduce the range of pixel changes and maintain center symmetry. For example, 220 × 220 represents a rectangle centered on the center of the image from a 224 × 224 image, and its length and width are both 220.

Pixel changes		4 0.0080%	8 0.0159%	16 0.0319%	32 0.0638%	64 0.1276%
224 × 224	Misclassification	100	87	156	234	534
	Recognition rate	99.00%	99.13%	98.44%	97.66%	94.66%
200 × 200	Misclassification	98	100	250	241	529
	Recognition rate	99.02%	99.00%	97.50%	97.59%	94.71%
150 × 150	Misclassification	69	112	167	236	563
	Recognition rate	99.31%	98.88%	98.33%	97.64%	94.37%
100 × 100	Misclassification	69	118	162	238	320
	Recognition rate	99.31%	98.82%	98.38%	97.62%	96.80%
50 × 50	Misclassification	65	103	157	205	251
	Recognition rate	99.35%	98.97%	98.43%	97.95%	97.49%

Table 9. The effect of pixel changes on the recognition accuracy of the deep learning model.

Pixel changes		4 0.0080%	8 0.0159%	16 0.0319%	32 0.0638%	64 0.1276%
FGSM	Misclassification	6697	4576	4361	2698	2260
	Recognition rate	33.03%	54.24%	56.39%	73.02%	77.40%

6.2. Experiments for Our Data Augmentation Method

To verify the effectiveness of our data augmentation method, we conducted experiments on the resnet101. We made random pixel changes to 1000 images in the training set. And small-scale retraining was carried out based on the ResNet101 pre-training model. According to the number of modified pixels, five new ResNet101 models were trained. The modified value of the pixel is limited between 0 and 255, and there is no limit to the position of the modified pixel. After that, random pixel changes were performed on these 1000 images of the same type and the accuracy of the model was tested.

Specifically, in addition to the training set, the other parameter settings are consistent with the training process for each training. In the parameter setting, we do not fix the parameters of the last layer (fully connected layer), and modify the number of categories to the number of image categories participating in the training, the learning step is 0.001, the number of iterations is 20, and the test sample is 1000. The final experimental results are as follows.

Table 10 shows the model's ability to recognize random pixel changes after using our data augmentation method.

It can be seen from the experiment that when the changed pixels are small, the retrained model will have a better effect. The main reason why we can achieve such an effect is that the model we used has been trained on a large data set and has a very high recognition ability. Our data augmentation method performs very small-scale retraining without changing the original recognition accuracy and optimizes the implicit classification boundary of the corresponding category images in the training set.

The improvement of our data augmentation method for recognition robustness is also reflected in the ability to recognize adversarial examples. Training by modifying a few pixels can make the model defensive against pixel changes. The defense capability is also effective for adversarial examples because adversarial examples can modify pixels to achieve misclassification. From the perspective of pixels, a few pixel changes and adversarial attacks all modified the pixel values in the image. We conducted experiments on 10,000 adversarial examples generated by the FGSM method, and input them into the ResNet101-4 model we trained. The recognition results are as shown in Table 11.

It can be seen from this experiment that the high threat of adversarial samples has been reduced. The only 24.9% of the images still maintain the attack capability, and most of the images are restored to the correct classification. This shows that our data augmentation method can effectively improve the model's robustness to FGSM attacks.

Table 10. Experimental results of five new training models on the same batch of images. 1000 images were selected for each training, and the original model was trained on a small scale after modifying a different number of pixel values. Resnet101-4 represents that the training image of the model is generated from images with 4 pixels changed.

Pixel changes		4 0.0080%	8 0.0159%	16 0.0319%	32 0.0638%	64 0.1276%
ResNet101-4	Misclassification	22	21	24	21	27
	Recognition rate	97.80%	97.90%	97.60%	97.90%	97.30%
ResNet101-8	Misclassification	22	22	25	28	31
	Recognition rate	97.80%	97.80%	97.50%	97.20%	96.90%
ResNet101-16	Misclassification	29	30	28	29	37
	Recognition rate	97.90%	97.00%	97.20%	97.10%	96.30%
ResNet101-32	Misclassification	27	26	29	32	35
	Recognition rate	97.30%	97.40%	97.10%	96.80%	96.50%
ResNet101-64	Misclassification	47	48	50	52	55
	Recognition rate	95.30%	95.20%	95.00%	94.80%	94.50%

Table 11. The new model's defense ability tests against adversarial examples.

Model	Misclassification	Recognition rate
ResNet101-4	2490	75.10%
ResNet	10,000	0%

7. Discussion

Our experiments prove the effectiveness of our data augmentation method. As mentioned above, training the model with small-distance samples can optimize the implicit classification boundary of the model, making it more tolerant of pixel changes. This is why the retrained model also has the ability to recognize adversarial samples.

8. Conclusion

This paper mainly explores the recognition robustness of five commonly used classic deep learning models. These models are very sensitive to subtle changes in pixels. To solve this problem, this paper proposes a new data augmentation method and a new robustness evaluation standard Pixel-index. The data augmentation method in this paper can modify a small number of pixels in an image randomly, help the model learn the feature distribution of normal images. Users can quickly obtain the recognition robustness evaluation of the model without knowing the internal information structure of the model by Pixel-index. In general, this research can improve the recognition ability of deep learning models, and provide a good augmentation method and comparison method for multiple types of recognition applications.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) Deep Learning, Volume 1. MIT Press, Cambridge.

[2] Lecun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436. https://doi.org/10.1038/nature14539

[3] Deng, L. and Yu, D. (2014) Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, **7**, 197-387. https://doi.org/10.1561/2000000039

[4] Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015) Deep Face Recognition. *Proceedings of the British Machine Vision Conference*, Swansea, 7-10 September 2015, 41.1-41.12. https://doi.org/10.5244/C.29.41

[5] Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S.Z. and Hospedales, T. (2015) When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Boston, 7-12 June 2015, 142-150. https://doi.org/10.1109/ICCVW.2015.58

[6] Alotaibi, M. and Mahmood, A. (2017) Improved Gait Recognition Based on Specialized Deep Convolutional Neural Network. *Computer Vision and Image Understanding*, **164**, 103-110. https://doi.org/10.1016/j.cviu.2017.10.004

[7] Wang, J., Ma, Y., Zhang, L., Gao, R.X. and Wu, D. (2018) Deep Learning for Smart Manufacturing: Methods and Applications. *Journal of Manufacturing Systems*, **48**, 144-156. https://doi.org/10.1016/j.jmsy.2018.01.003

[8] Essien, A. and Giannetti, C. (2020) A Deep Learning Model for Smart Manufacturing Using Convolutional LSTM Neural Network Autoencoders. *IEEE Transactions on Industrial Informatics*, **16**, 6069-6078. https://doi.org/10.1109/TII.2020.2967556

[9] Lee, J., Azamfar, M., Singh, J. and Siahpour, S. (2020) Integration of Digital Twin and Deep Learning in Cyber-Physical Systems: Towards Smart Manufacturing. *IET Collaborative Intelligent Manufacturing*, **2**, 34-36. https://doi.org/10.1049/iet-cim.2020.0009

[10] Ahuett-Garza, H. and Kurfess, T. (2018) A Brief Discussion on the Trends of Habilitating Technologies for Industry 4.0 and Smart Manufacturing. *Manufacturing Letters*, **15**, 60-63. https://doi.org/10.1016/j.mfglet.2018.02.011

[11] Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D. and Tzovaras, D. (2021) Machine Learning and Deep Learning in Smart Manufacturing: The Smart Grid Paradigm. *Computer Science Review*, **40**, Article ID: 100341. https://doi.org/10.1016/j.cosrev.2020.100341

[12] Ulicny, M., Lundstrom, J. and Byttner, S. (2016) Robustness of Deep Convolutional Neural Networks for Image Recognition. In: *International Symposium on Intelligent Computing Systems*, Springer, Berlin, 16-30. https://doi.org/10.1007/978-3-319-30447-2_2

[13] Goswami, G., Ratha, N., Agarwal, A., Singh, R. and Vatsa, M. (2018) Unravelling Robustness of Deep Learning Based Face Recognition against Adversarial Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32, 6829-6936.

[14] Goodfellow, I.J., Shlens, J. and Szegedy, C. (2014) Explaining and Harnessing Adversarial Examples.

[15] Van Dyk, D.A. and Meng, X.L. (2001) The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, **10**, 1-50. https://doi.org/10.1198/10618600152418584

[16] Miko lajczyk, A. and Grochowski, M. (2018) Data Augmentation for Improving Deep Learning in Image Classification Problem. 2018 *IEEE International Interdisciplinary PhD Workshop* (*IIPhDW*), Swinoujscie, 9-12 May 2018, 117-122. https://doi.org/10.1109/IIPHDW.2018.8388338

[17] Deng, J., Dong, W., Socher, R., *et al*. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[18] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[19] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[20] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[21] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[22] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K. (2016) Squeezenet: Alexnet-Level Accuracy with 50x Fewer Parameters and < 0.5 mb Model Size.

[23] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2013) Intriguing Properties of Neural Networks.

[24] Ozbulak, U., Van Messem, A. and De Neve, W. (2019) Impact of Adversarial Examples on Deep Learning Models for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, 300-308. https://doi.org/10.1007/978-3-030-32245-8_34

[25] Kurakin, A., Goodfellow, I. and Bengio, S. (2016) Adversarial Examples in the Physical World.

[26] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A. (2019) Adversarial Examples Are Not Bugs, They Are Features. 33*rd Annual Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 125-136.

[27] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B. and Swami, A. (2017) Practical Black-Box Attacks against Machine Learning. *Proceedings of the* 2017 *ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi, 2-6 April 2017, 506-519. https://doi.org/10.1145/3052973.3053009

[28] Ilyas, A., Engstrom, L., Athalye, A. and Lin, J. (2018) Black-Box Adversarial Attacks with Limited Queries and Information.

[29] Hosseini, H., Chen, Y., Kannan, S., Zhang, B. and Poovendran, R. (2017) Blocking Transferability of Adversarial Examples in Black-Box Learning Systems.

[30] Carlini, N. and Wagner, D. (2016) Towards Evaluating the Robustness of Neural Networks. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017. https://doi.org/10.1109/SP.2017.49

[31] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B. and Swami, A. (2016) The Limitations of Deep Learning in Adversarial Settings. 2016 *IEEE European*

Symposium on Security and Privacy (*EuroS&P*), Genoa, 7-11 September 2020, 372-387. https://doi.org/10.1109/EuroSP.2016.36

[32] Baluja, S. and Fischer, I. (2017) Adversarial Transformation Networks: Learning to Generate Adversarial Examples.

[33] Su, J., Vargas, D.V. and Sakurai, K. (2019) One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, **23**, 828-841. https://doi.org/10.1109/TEVC.2019.2890858

[34] Wong, S.C., Gatt, A., Stamatescu, V. and McDonnell, M.D. (2016) Understanding Data Augmentation for Classification: When to Warp? 2016 *IEEE International Conference on Digital Image Computing*: *Techniques and Applications* (*DICTA*), Gold Coast, 30 November-2 December 2016, 1-6. https://doi.org/10.1109/DICTA.2016.7797091

[35] Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. and Le, Q.V. (2018) Autoaugment: Learning Augmentation Policies from Data. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00020

[36] Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y. (2020) Random Erasing Data Augmentation. *The Thirty-Fourth AAAI Conference on Artificial Intelligence* (*AAAI*-20), New York, 7-12 February 2020, 13001-13008. https://doi.org/10.1609/aaai.v34i07.7000

[37] Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. and Le, Q.V. (2019) Autoaugment: Learning Augmentation Strategies from Data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 113-123. https://doi.org/10.1109/CVPR.2019.00020

[38] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020) Generative Adversarial Networks. *Communications of the ACM*, **63**, 139-144. https://doi.org/10.1145/3422622

[39] Jan, S.T., Messou, J., Lin, Y.C., Huang, J.B. and Wang, G. (2019) Connecting the Digital and Physical World: Improving the Robustness of Adversarial Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, 962-969. https://doi.org/10.1609/aaai.v33i01.3301962

[40] Luo, B., Liu, Y., Wei, L. and Xu, Q. (2018) Towards Imperceptible and Robust Adversarial Example Attacks against Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, 7-12 February 2020.