

final

by Yuhang Lan

Submission date: 07-Sep-2022 07:47PM (UTC+0100)

Submission ID: 185786458

File name: 32910_Yuhang_Lan_final_430701_1071315743.docx (1.11M)

Word count: 13496

Character count: 74744



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

A research of semantic segmentation algorithm for drug bottle defect detection based on
convolutional neural network

by

Lan Yuhang

Supervisor: Terry Walcott

Project submitted as part of the requirements for the award of MSc-DAI Data and
Artificial Intelligence

September 2022

DECLARATION

I, ...Lan Yuhang..... declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.



Signed :

Date : 2022.9.5.....

FORM OF CONSENT

I _____ Lan Yuhang _____ hereby consent that my Project, submitted in candidature for the M.Sc. Applied Computing, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.



Signed _____

Date _____2022.9.5_____

Where necessary replace M.Sc. Applied Computing with your named award.

Abstract

The pharmaceutical packaging industry is an irreplaceable and important industry. However, due to technical limitations, some finished products have surface defects. Some of these defects are caused by scratches on the original workpiece, some are related to the chemical composition used in moulding, some are related to the ageing of the equipment, etc. Surface defects not only affect the appearance of the product, but in severe cases even the quality of the product, and when a large number of products are downgraded to defective, this can significantly reduce the economic efficiency of the factory. Their classification and identification is the key to detecting surface defects in plastic bottles and caps and analysing the causes of damage. Traditional image defect detection algorithms are not as accurate as they could be due to the low contrast between background and defect features. In recent years, with the rapid development of deep learning technology, good results have been achieved in various fields of image classification. Therefore, the use of deep learning technology to process images of steel surface defects is of great significance in both theory and practice. The main research work in this paper is as follows.

(1) An end-to-end two-stage segmentation classification network is designed for the implementation of surface defect detection on the cap of plastic medicine bottles. For the model, an attention mechanism is introduced on the framework of the residual structure, a serial connection and feature fusion is used for the attention module and the network convolution module, a residual connection is added to the network within and between modules to solve the gradient propagation problem, and a BN layer is added before the module output to reduce the effect of overfitting. A BN layer is added before the output of the module to reduce the effect of overfitting. In this paper, a dataset containing six types of defective images of pharmaceutical bottle caps is constructed and the experimental results are compared to illustrate the improvement of the network for defect detection and classification by comparing the training accuracy curve, test set evaluation and the basis of classification decision.

(2) The loss value is recalculated using the improved focalloss cross-entropy loss function algorithm, and the learning rate is dynamically adjusted using the loss value as

a guiding parameter. The cross-entropy loss function is improved by adding adjustment coefficients to reduce the loss of simple and easy-to-classify images, so that the number of training rounds is used as an important parameter to adjust the learning rate during the training process, accelerating the decline of the learning rate when the number of training rounds is increased and slowing down the decline of the learning rate when the number of training rounds is too low. The experimental results show that this method not only reduces the training time of the network, but also improves the convergence effect of the network.

(4) A multi-scale convolutional kernel and a two-stage segmentation classification defect detection network model backbone were used to achieve feature fusion in the backbone network. The improved FRB attention module was inserted into the horizontal connection of Res34 to increase the shallow network perceptual field and improve the accuracy of the network model for surface defect image recognition.

(5) To address the problem of unbalanced samples in the defective dataset, a training strategy of throwing in positive and negative samples 1:1 was adopted. The results were visualised and the designed module was proved to have improved the overall network structure, better convergence in the validation set and higher accuracy in the test set, which verified the correctness of the improvement idea of this paper.

In this paper, the Res-SegDec-Net-FRB can obtain 99.3% accuracy in the self-built bottle cap dataset with full labels, and 84.14% accuracy with a small number of pixel-level labels

Keywords—Deep learning; Defect detection; end-to-end two-stage networks; attention; ResNet

CONTENTS

CHAPTER 1 INTRODUCTION	1
CHAPTER 2 LITERATURE REVIEW	5
CHAPTER 2.1 Current state of research in the field of image recognition	5
CHAPTER 2.2 Current research in the field of defect detection.....	7
CHAPTER 2.3 Challenges of the problem.....	10
CHAPTER 3 RESEARCH AND/OR DEVELOPMENT METHODOLOGY	13
CHAPTER 3.1 Research content.....	13
CHAPTER 3.2 Research methods.....	14
CHAPTER 3.3 ResNet	18
CHAPTER 3.4 Dilated Convolution	19
CHAPTER 3.5 Attention.....	22
CHAPTER 3.6 Analysis of plastic pill bottle cap dataset and defect characteristics	26
CHAPTER 3.7 Experimental Environment	28
CHAPTER 4 DESIGN AND IMPLEMENTATION	29
CHAPTER 4.1 Model.....	29
CHAPTER 4.2 Segment networks.....	29
CHAPTER 4.3 Training strategy.....	31
CHAPTER 4.4 Label weights.....	31
CHAPTER 4.4 Loss Function.....	32
CHAPTER 5. RESULTS AND DISCUSSION	34
CHAPTER 5.1 Evaluation indicators	34
CHAPTER 5.2 Experimental results in full pixel-label	34
CHAPTER 5.2 Experimental results in 30% pixel-label	37
CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS.....	39
CHAPTER 7. REFLECTION	41
REFERENCE	42

CHAPTER 1 INTRODUCTION

With increasing attention being paid to issues such as food safety and medical health, the pharmaceutical and health industry is booming. The development of pharmaceutical packaging is an important part of the entire pharmaceutical industry chain. With the development of technology and the continuous improvement of people's living standards, medical health and food safety issues are attracting more and more attention. The quality of pharmaceutical bottles is related to the storage, distribution, transport and use of medicines, and directly affects their effectiveness and safety. The production process of pharmaceutical bottles contains several links and processes, each of which is subject to strict quality control. The surface inspection of the bottles before leaving the factory is a very important part of the process, i.e. the inspection of defects such as gas lines, cracks and spots and the measurement of the diameter of the bottle, so that no defects are allowed to exist.

As a special commodity, the safety and effectiveness of medicines need to be effectively protected. Pharmaceutical packaging plays a vital role in the transport, storage, sale and use of medicines. As a sign of the innovative development of the pharmaceutical packaging industry, the demand for pharmaceutical plastic bottles is increasing year by year, and its good development prospects have driven the continuous innovation of related industrial technologies. With the idea that pharmaceutical packaging containers are part of the medicine gradually being established, advanced companies are introducing pharmaceutical injection blow moulding lines to produce high quality high density polyethylene. The presence of poor encapsulation, contamination and scratches in the encapsulation process of pharmaceutical plastic bottles will not only affect the operation of the next process, the quality of the medicine, but will even affect the safety of the lives of those who use the medicine. Therefore, it is necessary to detect defects in the encapsulation and to eliminate them in a timely manner.

According to Freedonia's latest market research report, the global pharmaceutical packaging market demand is expected to grow at an annual rate of 6%. As the main protagonist of pharmaceutical packaging, injection moulded vials account for the largest

share of the pharmaceutical packaging market demand. The large market size and economic benefits also place higher demands on the level of quality and safety regulations for pharmaceutical production, of which the certification of pharmaceutical containers to meet cleanliness standards is a major difficulty. Many enterprises may not be able to meet the relevant requirements even if they increase their investment to improve the relevant hardware facilities more than in the past. Although the production of many injection moulded pharmaceutical bottles is now fully automated, the inspection of their surface quality is still mainly manual, especially in some developing countries.

In summary, the detection of defects can be broadly divided into three methods [1]: the first is the traditional manual detection method. The first is the traditional manual inspection method, which requires workers to operate in a dark room, i.e. to visually inspect the vials for defects, with the disadvantages of slow inspection speed, tedious operation, poor reliability and high leakage rate, and easy fatigue of the inspectors [2]. This method is gradually being banned due to its low efficiency, low accuracy and susceptibility to the subjective mood of the inspection workers; the second is the contact method, which means that the inspection is carried out through direct contact between the mechanical device and the object to be inspected. Although this method can basically meet the needs of the upper production, but the flexibility is poor, equipment manufacturing cost is high, is being eliminated; the third method is currently more popular, that is, the use of machine vision for detection and identification. Machine vision-based defect detection methods are faster, less costly, more accurate and less intrusive than manual ones [3]. In this context, the study of a machine vision injection moulded medicine bottle inspection system with is of great significance in providing support and reference for enterprises to carry out innovation, enhancing the intelligence of relevant enterprises and accelerating the development of the industrial internet and industrial data to the cloud.

As industrial defects can be considered as "anomalies" in the appearance of industrial products, some industrial defect detection methods also adopt the idea of anomaly detection. However, the definition of anomaly detection differs from that of industrial

defect detection. Specifically, the concept of anomaly detection is broader and more abstract, with image anomaly detection focusing on whether the input image is an anomalous instance, whereas industrial defect detection is more concerned with the task of detection at the pixel level. At the pixel level, the difference between anomalies and normal patterns is more subtle and significantly more difficult to detect. Therefore, direct anomaly detection methods are not sufficient for the task of industrial defect detection. In recent years, deep learning methods have come to dominate the field due to their superior performance in processing industrial images with complex backgrounds and weak defects.

Compared to the shortcomings of traditional manual inspection methods, computer vision intelligent inspection systems have the advantages of high efficiency, stability and speed, which completely eliminate the influence of human factors on the slow speed and low accuracy of defect detection, and are therefore favoured by the packaging quality inspection industry. This paper is based on the combination of computer and neural network for the identification of defects in the packaging of pharmaceutical bottles. It is based on the above problems in the packaging of pharmaceutical bottles and attempts to achieve the automatic detection of defects in the packaging of pharmaceutical bottles, and also lays a solid foundation for future research.

I have worked for a technology company dedicated to providing solutions for companies and was involved in the development of an algorithm for a cap defect detection machine for pharmaceutical bottles, and this project is based on this experience. In response to the shortcomings and inadequacies of manual inspection methods, the thesis proposes a solution for the detection of pharmaceutical packaging defects based on chemistry vision. We believe that carrying out research work on this topic has the following theoretical significance and practical guidance value.

1. the inspection system will greatly improve the speed, accuracy and reliability of the inspection, while enabling batch inspection.
2. it provides a theoretical basis for the design and development of an application system that can improve the accuracy of pharmaceutical packaging defect detection. the

combination of theoretical knowledge and production practice has wide application value and good market value.

3. summarise the relevant digital image processing methods and seek suitable processing algorithms for detection targets, providing an empirical basis for future research.

4. Summarise the research methods related to industrial vision type products and provide technical prerequisites for future development of related products.

CHAPTER 2 LITERATURE REVIEW

CHAPTER 2.1 Current state of research in the field of image recognition

The American Society of Manufacturing Engineers (SME) Machine Vision Section and the American Robotics Industry Association (Min IA) Automation Vision Section define machine vision as "Machine vision is the use of optics for non-contact perception to automatically acquire and interpret images of a real scene, obtain information and control a machine or process". In other words, machine vision is the study of the science of using cameras and computers to mimic the human eye and brain to perform tasks such as identifying, tracking and measuring targets.

In recent years, with the rapid development of the communications industry, the era of Internet of everything has arrived, and various smartphones and social networking software have emerged, gradually covering people's lives with massive amounts of picture information, with popular applications such as Instagram and Whatsapp circulating tens of millions of pictures every day. Whether it's funny emoticons or beautiful pictures of people, they have become another medium of communication in the new age, spanning national and linguistic boundaries and spreading all over the world. This section explores the development of image recognition and the current state of research in the field of image recognition in the context of massive image data. In recent years, with the development of computer information technology, the increase of computer computing power has led to the rapid development of various types of research based on convolutional neural networks. Since the 1980s, the Neocognitron [4] model was proposed to build a recognition model like the human brain. Then in 1989, LeCun et al [5-6] proposed LeNet-5, which was mainly used for the classification of handwritten digits and first established an early version of backpropagation training. Subsequently, Hecht-Nielsen [7] further refined the theory of backpropagation.

It was the introduction of AlexNet by Hinton et al. in 2012 [8] that really brought deep learning-based image recognition research to the forefront. It was the first model to achieve a top 5 error rate of 15.4% in the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), winning the competition by an absolute margin of 10.9 percentage

points over the second place, turning the field of image recognition upside down and causing many scholars to study deep learning, which was the beginning of the deep learning boom. GoogLeNet [9] and VGGNet [10], two models that followed AlexNet, were another milestone in artificial neural network research, with a deeper network structure proposed by the Visual Geometry Group from the University of Oxford at the 2014 ImageNet-LSVRC challenge. VGGNet was the second runner-up in the 2014 ImageNet-LSVRC challenge, highlighted by its study of network depth. Compared to AlexNet, VGGNet uses several successive 3×3 convolutions instead of the larger convolutional kernels in AlexNet, and for a given perceptual field, the stacking of multiple layers of small convolutional kernels not only increases the network depth, but also results in a significant reduction in the network model parameters. The Inception structure is designed from a different perspective than VGGNet, using multiple size convolutional kernels and then aggregating them to obtain a richer feature set, and using 1×1 convolutional kernels to lift and lower the dimensionality, reducing computational complexity while maintaining recognition accuracy. This reduces the computational complexity while ensuring recognition accuracy. The entire network is overlaid with Inception structures to improve training results. Both network models have their advantages and disadvantages, but recent research has shown that VGGNet has a better generalisation capability than GoogLeNet. In 2015, Kai-Ming He et al. proposed a residual learning framework to simplify network training, namely ResNet [11], which also won the Vision Challenge, and the biggest contribution of this network is that it proposes a residual module, and by repeatedly overlaying the residual module The most important contribution of this network is the introduction of the residual module, which is used to deepen the network by repeatedly overlaying the residual module to obtain more non-linear features to improve the accuracy. The residual module differs from traditional convolutional neural networks in that it uses fast connections to map the shallow output directly onto the deep output, which to a certain extent solves the gradient problem caused by deepening the network and provides a feasible research direction for deep learning tasks with higher semantic features. Although networks such as VGG and

ResNet have achieved good results in terms of model recognition accuracy, due to the gradual deepening of the network, the model computation and parameters gradually become larger. This network structure is built with deeply separable convolution, has fewer model parameters, is computationally fast, and is suitable for deployment in mobile or embedded systems. Although the lightweight convolutional design allows the network model to suffer from some loss of accuracy in tasks related to deep learning. However, the loss is within the acceptable range. With the rapid development of deep learning, MobileNet networks were gradually optimised and improved, and MobileNetV2 [12] and MobileNetV3 [13] networks were proposed. The MobileNetV3 model uses neural architecture search to obtain the optimal model, using modules inherited from the MobileNetV1 and MobileNetV2 networks. Today, convolutional neural networks have been around for 30 years, and have only recently received scholarly attention due to the limitations of computer technology, but in just a few years, research and applications based on convolutional neural networks have gained widespread attention in various fields.

CHAPTER 2.2 Current research in the field of defect detection

Vision-based industrial defect detection is not only of great research value, but also holds great promise for application. However, compared to general target detection tasks, industrial defect detection faces many difficulties, such as lack of defect samples, low visibility of defects, irregular shapes and unknown types, which make it difficult for many existing methods to meet the demands of both high accuracy and high speed, and therefore there are still a lot of problems to be solved on the way to implementation. Since industrial defects can be considered as "anomalies" in the appearance of industrial products, some industrial defect detection methods have also adopted the idea of anomaly detection. However, the definition of anomaly detection differs from that of industrial defect detection. Specifically, the concept of anomaly detection is broader and more abstract, with image anomaly detection focusing on whether the input image is an anomalous instance, whereas industrial defect detection is more concerned with the task of detection at the pixel level. At the pixel level, the difference between anomalies and normal patterns

is more subtle and significantly more difficult to detect. Therefore, direct anomaly detection methods are not sufficient for the task of industrial defect detection. In recent years, deep learning methods have become increasingly dominant in this field due to their excellent performance in processing industrial images with complex backgrounds and weak defects. The literature [14, 15] provides a comprehensive overview and analysis of deep learning-based anomaly detection methods, but lacks a discussion of industrial scenarios. The literature [16, 17] , although reviewed in the context of industrial production, focuses on traditional methods and system control. The literature [18] provides a systematic overview of deep learning-based surface defect detection methods, but mainly focuses on supervised methods. Recently, many new results have emerged based on unsupervised and semi-supervised setups as well.

Vision-based industrial defect detection aims to detect visible defects in a variety of industrial products such as fabrics, chips, pharmaceuticals and even infrastructure materials. These defects, although small, can seriously jeopardise the proper functioning of a product. They can occur at any point in the life cycle of an industrial product, such as during production, transport and use. The concept of a defect is analogous to that of anomaly. Anomalies are data that are outside of the expected pattern[19] and have been defined and classified in a large body of work. Typically, anomalies are classified as point anomalies, contextual anomalies and cluster anomalies based on the presence or absence of contextual relationships between data. Point anomalies, also known as outliers [20], describe independent data whose values deviate from the normal sample; contextual anomalies also describe data points whose values fall within the normal range but do not conform to local contextual patterns; and cluster anomalies describe a collection of related data in which the values of each instance are in the normal range when examined individually, but the correlation characteristics of the collection as a whole do not conform to normal patterns. The relevance characteristics of the set as a whole do not obey the normal pattern. Specifically for image data, the literature [21] classifies image anomalies into low-level texture anomalies and high-level semantic anomalies based on the presence or absence of significant semantics in the anomalies. Similar concepts to anomalies

include novelty and out-of-distribution (OOD) data [22]. In industrial defect detection, normal samples include a wide range of products and defects can be considered as "anomalies" in their appearance. Unlike industrial defects, which are often found in small areas of the image, they are much less significant and have a vague semantic concept. As a result, general image anomaly detection often only needs to distinguish between normal and abnormal samples, whereas industrial defect detection is more concerned with detecting abnormal pixels in an image. In real industrial scenarios, the definition of defects is more subjective, so scholars have tried to find its relevance to anomaly detection. Considering the similarity of these tasks, anomaly detection, novel point detection and OOD detection have also been adopted in some defect detection methods. For example, the pixel values of a defect, considered in isolation, can be likened to novelty points or outliers; when considering the association between pixels, they can be likened to contextual anomalies or cluster anomalies. Although defects are usually unknown patterns, some commonalities can be found in existing defect samples, so summarising the types of defects and contexts can be useful in designing targeted detection methods. Industrial defects can be classified as surface defects or structural defects depending on where they occur and how they manifest themselves. Surface defects are mainly found at localised locations on the surface of products and are usually characterised by texture variations, heterogeneous areas, anti-patterns or incorrect patterns. Examples include surface cracks, colour blocks, sparse weaving of fabrics and incorrect printing of logo text. Defective areas can be likened to outliers or cluster anomalies depending on how different their pixel values are from the surrounding background: outlier defects usually have pixel values that differ significantly from the normal image; cluster anomalies have pixel values that fall within the same range as the surrounding normal area and are therefore more difficult to detect. Structural defects are mainly caused by structural errors in the product as a whole, including deformation, misalignment, defects and contamination. Examples include bends in the wire, missing or misplaced edges of diodes, etc. Accordingly, backgrounds can be divided into textures and objects, depending on whether they include the overall structure of the product. The texture category focuses on

the local surface of the product and is divided into simple, regular and irregular textures in order of complexity. The object category includes the whole product, which is more complex and has a background that interferes with the product. In this case, not only surface defects but also structural defects need to be taken into account. It can be seen that the degree of weakness of different types of defects on different backgrounds varies. Even between different examples of the same defect, there can be significant differences in visibility. Depending on the granularity of the output, industrial vision defect detection tasks generally include classification and localisation. For an image instance to be measured, the classification task first classifies it as a normal or defective sample; when the type of defect is known, the type of defect can be further discriminated, e.g. there can be multiple defects on a fabric such as threads, different colours, etc. The objective of the localisation task is to find the specific area of the defect in the image, which can be divided into detection (detection frame) and segmentation (pixel level) depending on how the defect area is described. In fact, these tasks can sometimes be performed simultaneously, for example, visualisation methods [24-26] can be used in classification models for pixel-level localisation, and segmentation results can also be used to determine the classification of the whole image. Due to the irregular shape and dimensional variation of industrial defects, it is difficult to accurately represent the position of defects in the detection frame and it is easy to introduce a lot of irrelevant background information, which makes the evaluation of defect detection performance difficult. As a result, researchers have focused on defect segmentation methods for practical defect location tasks. In view of this, this paper will focus on the work on classification and segmentation tasks.

CHAPTER 2.3 Challenges of the problem

Unlike the task of classifying, detecting and segmenting generic targets, industrial defect detection poses a new challenge by taking into account the specific task requirements, data samples and characteristics of the deployment environment. In this paper, we analyse both the data challenges and the task challenges.

Data Challenges: Lack of samples. The problem of sample paucity is manifested in three main ways. Firstly, the amount of data available is limited. Secondly, defects are rare in products due to the strict control of defect rates in the production process, so the number of defective samples is very small compared to the more accessible normal samples. Thirdly, accurate labelling of defects is costly. As a result, there is a clear imbalance in the industrial data, with the majority of samples being normal and only a few being defective, and even fewer being accurately labelled. Unknowability and subjectivity. The pattern of defects is often unpredictable, making it difficult to describe defects directly. Current artificially generalised defects still do not cover all types of defects, resulting in many previous methods failing to identify new types of defects. At the same time, the definition of defects often needs to be contextualised and more subjective, which makes it more difficult to label data

Low visibility. There are three main aspects of low visibility. Firstly, defects are often only a small part of a high-resolution image; secondly, the contrast between the defect and the surrounding background is low and not significant enough; thirdly, defects have a low signal-to-noise ratio and the noise introduced by the imaging can also interfere with the detection of defects. Therefore, the difference between positive and negative samples in industrial defect detection is very small compared to general anomaly detection.

There are many different types of defects. Defects are often irregular. Different types of defects may exist in the same industrial product, and the same defect may be diverse in shape, size, color and other characteristics. In the past, methods were usually only able to detect certain types of defects. Nowadays, methods need to be able to detect a wide range of defects to avoid missing them.

Complex backgrounds. Whereas previous methods have focused on simple, regular product backgrounds, today's products to be tested have more complex structures and textures. A single high-resolution image of a product may contain multiple components and thus vary in style from area to area. There is also a certain amount of variation between different images of the same product to be measured, as the imaging conditions are not always identical. In the case of unmanned inspection tasks, there are even multiple

views. In this regard, the model needs to be able to accommodate the intra-class variation of normal samples and avoid misidentifying normal areas as defects.

Task challenges:

High accuracy. Industrial vision defect detection requires methods that are highly accurate and can detect even subtle defects that are difficult to observe with the human eye. Not only do they need to minimise missed and false detections, but they also need to be able to adjust their performance in a timely manner. Different applications require different levels of missed and false detections. In most industrial scenarios, the preference is to avoid missed detections. For defect segmentation tasks, a high accuracy segmentation area facilitates subsequent defect evaluation and repair.

Low overhead. Current imaging technologies can already provide high resolution images for industrial vision systems. For practical deployments, the computational time and storage resource overheads need to be taken into account. For example, models may be deployed in environments where computing power is limited, such as embedded systems. High-speed, real-time approaches are also a focus of research in order to run in parallel with actual manufacturing or intelligent inspection

CHAPTER 3 RESEARCH AND/OR DEVELOPMENT METHODOLOGY

CHAPTER 3.1 Research content

This paper takes the convolutional neural network model and the image of surface defects on the cap of plastic medicine bottles as the research object, and addresses many problems in the classification and recognition model of surface defects on the cap of plastic medicine bottles, and conducts research from the following aspects.

(1) The training strategy of balancing positive and negative samples is used to ensure a balanced number of positive and negative samples for the problem of unbalanced data set of surface defects on plastic medicine bottle caps . The segmentation part uses a modified ResNet34 network to obtain the segmentation map, makes classification predictions in the classification network, generates a heat map using Grad CAM technique to visualise the recognition results, and analyses the impact on recognition accuracy after fusing the underlying feature information. The internal working principle of the convolutional neural network is analysed, and the image characteristics are analysed to determine the subsequent ideas for network improvement.

(2) The label weights are processed, combined with the characteristics of the surface defects of the cap of plastic medicine bottles, the coarse labels are inflated and deflated, and the weights are set for the pixel points of the labels in them, the closer to the centre, the higher the weights are, reducing the requirements of the network for labels and enhancing its recognition ability in the case of less amount of labels, so as to achieve the purpose of enhancing the robustness of the network.

(3) Most of the original object detection algorithms only use top-level features to do prediction, but we know that the lower-level features have less semantic information, but the target location is accurate; the higher-level features have more semantic information, but the target location is coarse. In this paper, we adopt the multi-scale feature fusion approach, in which the multi-scale features obtained from the segmentation network are input into the classification network, and the multi-scale features are fused to make predictions. This approach fuses the feature maps of different network layers with different resolutions while still retaining the deep semantic information, and the number

of channels remains unchanged, and the corresponding location feature maps are summed to achieve feature fusion.

(4) The FRB attention module is introduced to increase the perceptual field of the shallow network layers, and the effect of the improved RFB module on the recognition accuracy after the introduction of different layers is analysed. Through experimental comparison, the best insertion position of the RFB module is given, and it is also proved that the network model proposed in this paper can improve the classification accuracy.

(5) Drawing on the idea of detecting small target objects in target detection, the cross-entropy loss function is improved by introducing focalloss to reduce the loss results of simple and easy classification samples, so that the computational resources of the model are tilted towards difficult classification samples, thus increasing the attention weight of difficult classification samples. And a dynamic adjustment of the learning rate is proposed for the calculation, which can guarantee the performance of the neural network while reducing the training time to a certain extent, and the method is proved to be effective through experimental comparison.

CHAPTER 3.2 Research methods

A convolutional neural network is a feed-forward neural network and a powerful artificial neural network technique that incorporates complex convolutional computations and has a deep structure [27]. Its neuronal responses preserve the spatial structure of the specific problem and perform translation-invariant classification of the input features according to the depth structure, giving it a powerful representational learning capability. Convolutional neural networks are widely used in applications such as image target detection, image semantic segmentation and image recognition and classification, as they can learn representations directly from the original input image without the need for time-consuming image pre-processing. A convolutional neural network usually consists of a series of layers, which typically include a convolutional layer as well as an activation function, a pooling layer and a fully connected layer. The output of each layer of the network model is processed by a differentiable function on the output of the previous

layer to obtain the output of a new hidden layer. The convolutional layer, which is generally composed of multiple convolutional units, is an important part of a convolutional neural network responsible for feature extraction. The shallow convolutional layers in the network can only extract relatively low-level features such as lines or edges, while the deeper convolutional layers can extract complex features from the low-level features. The two-dimensional convolution algorithm is calculated as shown in Equation (3-1).

$$\begin{aligned}
 Y(m,n) &= X(m,n) * H(m,n) \\
 &= \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} X(i,j) * H(m-i,n-j) \\
 &= \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} X(m-i,n-j) * H(i,j) \quad (3-1)
 \end{aligned}$$

The activation function, which usually acts as a mapping between the feature inputs and the feature outputs in a convolutional neural network. As the activation function is a non-linear mapping, it is irreplaceable for convolutional networks to learn and understand multivariate and complex functions, allowing them to be applied to a wide range of non-linear models, thus increasing the scope of application of the network model. Commonly used activation functions include the Sigmoid function, the ReLU function and the Softmax function. The Sigmoid function can map any input to an output between 0 and 1. In the extreme case, if the input tends to negative infinity, the output tends to 0; if the input tends to positive infinity, the output tends to 1. Therefore, it is often used as a threshold function in neural networks, and its mathematical expression is shown in Equation (3-2).

$$f\left(\frac{1}{1+e^{-x}}\right) \quad (3-2)$$

The modified linear unit ReLU simply determines whether the input is greater than 0. When the input passing through the ReLU function is less than 0, then the output is directly 0 and the neuron is inactive. The ReLU function is expressed in equation (3-3).

$$f(x) = \max(0, x) \quad (3-3)$$

The Softmax function is used to map the different outputs of many neurons into the

interval between 0 and 1, which is equivalent to the probability that the current output belongs to each classification, thus achieving the purpose of multi-classification. The Softmax function is usually applied to the last layer of the network model, and its functional equation is shown in Equation (3-4).

$$s_i = \frac{e^i}{\sum e^j} \quad (3-4)$$

where e^i and e^j denote the output of the neuron. The pooling layer, whose main role is to scale the input feature map and extract the main features in the convolutional neural network. The pooling layer is usually applied after the convolutional layer to reduce the size of the feature parameter matrix and hence the number of parameters and floating point computational overhead in the neural network model. A pooling operation with a step size of 2 is shown in the figure3-1.

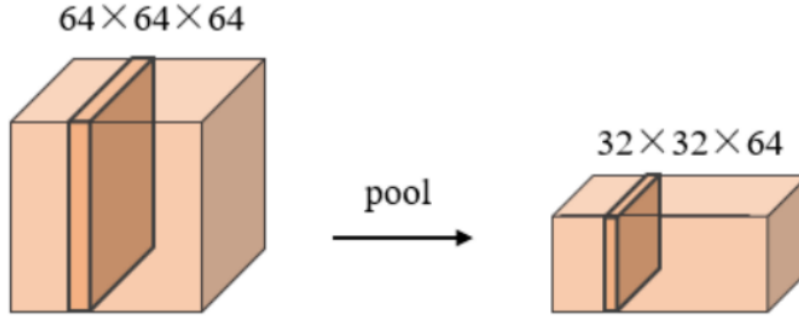


Fig.3-1. Pooling

The pooling operation can extract higher-level features to represent the input image, and then identify differences in relative position changes such as rotation, symmetry and scaling of the target image, effectively improving the recognition accuracy of the neural network. The general pooling operations include maximum pooling, mean pooling and global mean pooling, where maximum pooling and mean pooling are calculated as shown in Equations (3-5) and (3-6) respectively.

$$y_{m,n}^d = \max(x_i), x_i \in R_{m,n}^d \quad (3-5)$$

$$y_{m,n}^d = \frac{1}{m*n} \sum x_i, x_i \in R_{m,n}^d \quad (3-6)$$

The fully connected layer, usually at the end of the entire convolutional neural network, is used to output the predicted probabilities of each class using the available high-level image features, i.e. to act as a 'classifier'. Operations such as the convolutional and pooling layers are used to extract advanced features from the input image, while the fully connected layer is used to predict the probability of the output class using the activation function of the non-linear mapping. As the convolutional modules extract features from the images, the increasing number of convolutional modules will contribute to the performance of the network model to some extent [28]. However, when the number of convolutional modules reaches a certain level, the accuracy of the network model on the validation set decreases as the complexity of the model increases. Early Stopping [29], Dropout [30] and Batch Normalization (BN) can effectively reduce the effect of overfitting. The BN process consists of several steps: firstly, the mean of the batch data is calculated; secondly, the variance of the batch data is calculated; thirdly, the normalisation process is performed; and finally, the scale transformation and offset are completed. The process is shown in Equation (3-7), Equation (3-8), Equation (3-9) and Equation (3-10) respectively

$$\mu_\beta = \frac{1}{m} \sum_{i=1}^m x_i \quad (3-7)$$

$$\sigma_\beta^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_\beta)^2 \quad (3-8)$$

$$x_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}} \quad (3-9)$$

$$y_i = \gamma x_i + \beta \quad (3-10)$$

where β and γ are the learnable reconstruction parameters, μ_β is the data mean, σ_β^2 is the variance of the data set, x_i is the normalised data, and the output y_i is the value after adding the scaling γ and the translation variable β . There are various loss functions

used in the network training process, and the commonly used loss function is the multi-category cross-entropy loss function, which is calculated as shown in Equation (3-11)。

$$Loss = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n y_i \log(p_i) \quad (3-11)$$

where m is the number of samples, n is the number of label types, y_i is the label if the category is $i = 1$, otherwise 0, and p_i is the predicted output probability of the network model.

CHAPTER 3.3 ResNet

The problem proposed by resnet to solve: as the depth of the network increases, the gradient disappears or the gradient explodes will become more and more obvious. Assuming that the error coefficient of each layer is a factor less than 1, then the back propagation process is multiplied by a factor less than 1 every time it is propagated forward, then the deeper the network is, the more the gradient will tend to zero; conversely, assuming that the error coefficient of each layer is a factor greater than 1, then the deeper the network is, the more the gradient will tend to infinity. Then the deeper the network is, the more the gradient will tend to infinity.

As the defective features of the plastic medicine bottle cap surface defect image have low contrast with the normal area features, the reliance on low-level feature information is greater, and some categories of defective point targets are smaller, while the traditional CNN network hierarchy has the semantic characteristics from low to high, more low-level feature information is lost, the recognition ability of different stages is different, and the performance of semantic consistency is also different. The feature map generated by the low-level network has high resolution and contains a lot of details of small defective images, such as position and size, etc. However, due to less convolutional computation, there is a lot of noise and the perceptual field is small, so the semantic consistency is poor. The feature maps generated by the deep network have a large amount of high-level semantic information after multi-layer convolutional computation, and have a large perceptual field and good semantic consistency, but the feature maps have low resolution and lose a lot of detailed information.

The application of ResNet34 to manual parsing requires the use of spatial information

at the low level and semantic information at the high level. The semantic consistency of the higher stages is used to guide the lower stages to a higher prediction accuracy. As the task of recognising surface defects on the cap of a plastic medicine bottle focuses more on the underlying background texture information and the characteristics of the part of the classification with small target defect points, and as the deeper the neural network is, the lower the level of resolution of the image and the loss of information each time it is downsampled, this paper is based on the structure of ResNet34 and the specific network structure is connected as shown in Figure 3.2.

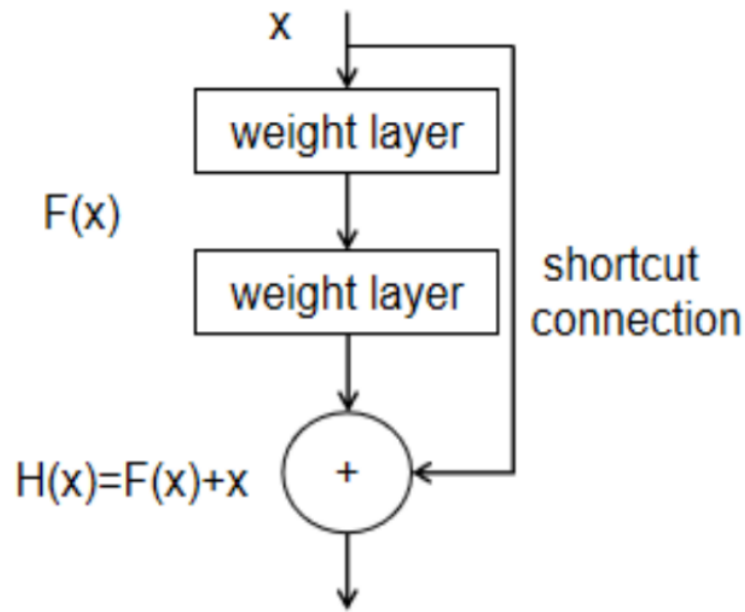


Fig.3-2. Res Block

The residual linking enhances feature contrast and mitigates feature decay. Without loss of significant information, this structure not only reduces the effect of noise and improves the recognition rate of the characterised features, but also makes model optimisation easier. In this paper, the Batch Normalization operation is chosen to deal with the problem of overfitting, i.e. to standardise the image feature values between layers.

CHAPTER 3.4 Dilated Convolution

Dilation convolution[31], is the process of spacing the data in the convolution kernel

by introducing a new parameter "dilation rate" into the convolution calculation. A comparison between normal and dilation convolution is shown in Figure 3-3(a)(b) (for a 3×3 convolution with a dilation rate equal to 2)。

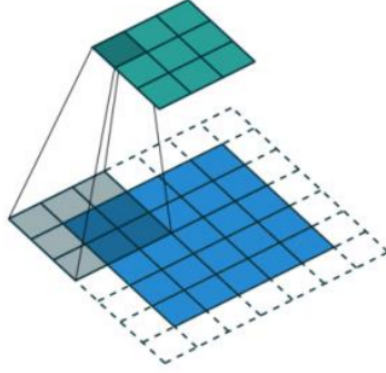


Fig.3-3. (a)

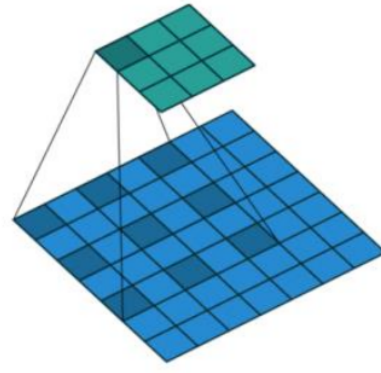


Fig.3-3. (b)

A 7×7 convolutional layer is equivalent to the superposition of three 3×3 convolutional layers, which makes it easier to learn to generate a well-performing feature space with fewer parameters, and therefore the mainstream convolutional neural network models are almost always structured with 3×3 and 5×5 convolutional layers. However, at the same time, there are a series of problems with up-sampling and down-sampling in deep convolutional neural networks. On the one hand, the loss of internal data structure and spatial hierarchy information due to pooling, and on the other hand, the loss of data after down-sampling of small targets (under certain conditions) makes it impossible to reconstruct the information. This problem is particularly acute in semantic segmentation, and is well avoided by the design of dilated convolution. Figure 3-4(a) below shows a normal convolution with a convolution kernel of 3×3 and a dilation rate of 1; (b) shows a null convolution with a dilation rate of 2 and a convolution kernel of 3×3 ; and (c) shows a null convolution with a dilation rate of 4 and a convolution kernel of 3×3 . (a) The figure has a perceptual field of 3×3 , with each point in the perceptual field having a convolution weight of 1 and each point participating in the convolution; (b) the figure has a perceptual field of 7×7 , but only 9 red points have a convolution weight of 1 and the remaining points do not participate in the convolution; (c) the figure has a perceptual field of 15×15 , and

similarly only 9 red points have a convolution weight of 1 and the remaining points do not participate in the convolution. The field of perception is calculated as F_{i+1} in Eq. (3-12) when the dilation rate is $i+1$. From the figure, it can be seen that the field of perception can reach 15×15 after convolving three layers of voids with dilation rates of 1, 2 and 4.

$$\begin{aligned} M_s(F) &= \sigma(f^{7*7}([AvgPool(F), MaxPool(F)])) \\ &= \sigma(f^{7*7}([F_{avg}^S, F_{avg}^S])) \end{aligned} \quad (3-11)$$

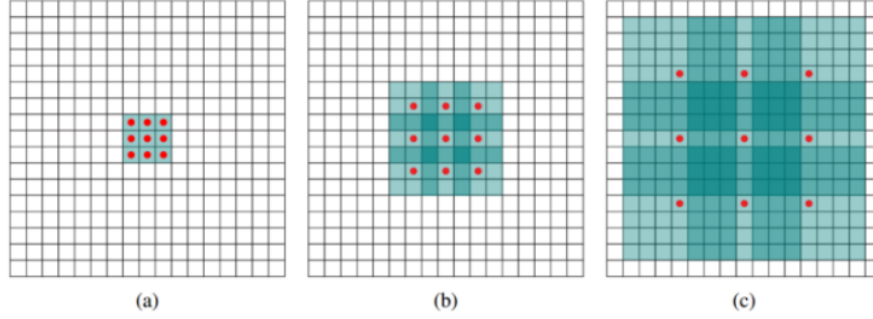


Fig.3-4 Dilation convolution with different expansion rates

If the pooling layer is simply removed, the perceptual field will be too small, and the expansion convolution can be used to further increase the perceptual field while keeping the $h \times w$ constant. The gridding effect: treating the image as a chessboard will result in loss of detail information, using multiple layers of expanded convolution will result in some pixels not being used and loss of detail information, comparing the expanded convolution with different expansion coefficients, the expansion coefficients 1, 2 and 3 are all 13×13 , but obviously the perceptual field is much larger and no detail information is lost.

In summary, what does null convolution do? Expanding the receptive field: In deep nets, downsampling (pooling or $s2/conv$) is always done in order to increase the receptive field and reduce the computational effort, which increases the receptive field but reduces the spatial resolution. In order not to lose resolution and still expand the field of perception, one can use null convolution. This is very useful in detection and segmentation tasks. On the one hand, a larger field of perception allows for the detection of large targets for segmentation, and on the other hand, a higher resolution allows for precise target localisation.

Capturing multiscale contextual information: The cavity convolution has a parameter to set the dilation rate, which means that the convolution kernel is filled with dilation rate-1 zeros, so that when different dilation rates are set, the perceptual field will be different, i.e. multiscale information is captured. Multiscale information is quite important in visual tasks. As can be seen here, the null convolution can expand the perceptual field arbitrarily and without introducing additional parameters, but if the resolution is increased, the overall computational effort of the algorithm will definitely increase.

CHAPTER 3.5 Attention

Attention Mechanism (AM) [32], an important image feature processing method in convolutional neural networks, has been studied and practically applied in the fields of machine translation, classification recognition, target detection and speech processing.

The principle of the attention mechanism is to learn the weight distribution of the original input based on the relevant feature map, and then use the resulting weight parameters to weight and sum the original feature map to more effectively perform visual tasks such as classification, segmentation, detection and analysis.

The effects and applications of attentional mechanisms in neural networks have been thoroughly investigated in the literature [33-34]. The main function of attentional mechanisms in computer vision is to allow the network system to focus its attentional resources on more feature-rich image regions. In recent years, the effective combination of network models and visual attention mechanisms has become increasingly valuable in deep learning research, where the combined network models not only have the ability to focus on useful image regions, but also help researchers understand the internal implementation principles of neural networks. In particular, most of these combined models focus on the use of masks to form attentional mechanisms. By using a new layer of weighting parameters, the mask identifies key valuable features in the input image, which in turn allows the neural network to focus on the effective regions of interest in the input image during the training process, thus creating attention.

There are several reasons why attentional mechanisms are becoming more and more widely used in neural networks: firstly, neural network models have previously been

regarded as black-box models, whereas attentional mechanisms can explain the principles well and greatly improve the transparency of the machine learning process; secondly, classical network models combined with attentional mechanisms are currently an effective approach to solving practical applications, for example, in machine translation, emotion analysis, semantic segmentation, dialogue systems, problem communication and multi-classification problems. Secondly, the combination of classical network models with attention mechanisms is an effective approach for solving practical applications, such as machine translation, sentiment analysis, semantic segmentation, dialogue systems, problem communication, multi-classification problems, etc. Furthermore, attention mechanisms can help to solve the problem of performance degradation due to input length in recurrent neural networks.

The primary design goal of the attention mechanism is to enable the convolutional neural network to focus its attention on valuable areas of the input image, reduce the waste of attention resources on invalid areas, and improve the accuracy of the network vision task by better learning the interconnectedness of the input image content features. In addition, the attention mechanism can reduce the computational overhead of the network model for high-dimensional image input by optimising the structure of the input data set, and reduce the reliance on computer hardware for the network model to process complex image input.

Attention tends to be more focused on the region or channel of the image input, which can be subdivided into three main attention domains: channel domain, spatial domain and hybrid domain. In the channel domain, the global average pooling or global maximum pooling is performed on the spatial features of each channel to extract the representative features of each channel, and then each representative feature is given a weight parameter of key information relevance. This is represented by Squeeze-andExcitation Networks [35], i.e. SENet. In the channel attention module, two pooling methods, global average pooling and global maximum pooling, are used to aggregate the spatial features of different information, and the computational process can be represented by Equation (3-13) respectively

$$\begin{aligned}
M_s(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\
&= \sigma(W_1(W_0(F_{avg}^C)) + W_1(W_0(F_{max}^C)))
\end{aligned} \tag{3-11}$$

Where F denotes the input features, AvgPool and MaxPool denote the averaging and max pooling operations, MLP denotes the multilayer perceptron and the Sigmoid function, W_0 and W_1 denote the network weights of MLP, F_{avg}^C and F_{max}^C denote the pooled feature information. The input features can be represented as $H \times W \times C$. After global average pooling and global maximum pooling in the spatial domain, two $1 \times 1 \times C$ channel features are obtained, followed by squeeze-and-excitation operations. The new features are multiplied with the original input features to obtain the new features after processing by the channel attention module.

The spatial domain is transformed to remove unimportant spatial information from the input features, while retaining the key information. Currently, the pooling layer in a convolutional neural network can use maximum pooling or average pooling to compress the image information and retain the key information.

In addition, feature normalisation can be used to remove some of the spatial information to reduce the amount of computation and improve accuracy. A representative example is the Spatial Attention Module [36]. In the Spatial Attention Module, the computational process can be represented by Equation (3-12)

$$\begin{aligned}
M_s(F) &= \sigma(f^{7 \times 7}([AvgPool(F), MaxPool(F)])) \\
&= \sigma(f^{7 \times 7}([F_{avg}^S, F_{avg}^S]))
\end{aligned} \tag{3-12}$$

In the spatial attention module, the input features of size $H \times W \times C$ are pooled by the average pooling operation and the maximum pooling operation in the channel dimension to obtain a spatial feature representation of $H \times W \times 1$. In the spatial attention module, the input features of size $H \times W \times C$ are pooled in the channel dimension to obtain a spatial feature representation of $H \times W \times 1$ by averaging and max pooling operations respectively, and then the two feature descriptions are fused in the channel dimension.

The spatial attention features are then multiplied with the original input features to obtain the new features after processing by the spatial attention module.

Mixed-domain attention is a combination of spatial and channel-domain attention.

Spatial domain attention is mainly a non-differentiated process that ignores the differences in information contained in different features in the channel domain, which can lead to significant limitations in the extraction of advanced features from images. In contrast, the channel domain attention ignores the spatially important or unimportant feature information in each channel and directly performs a global pooling operation, which is too generalised and lacks the retention of locally important features.

Therefore, by combining the characteristics of the two design approaches, These include the Convolutional Block Attention Module [37] and the Bottleneck Attention Module [38], namely CBAM and BAM.

The CBAM consists of a channel attention module and a spatial attention module, which are connected in the way shown in Figure 3-5.

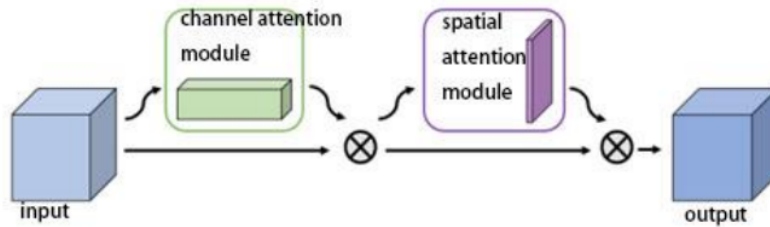


Fig.3-5 CBAM structure

In addition to this in order to obtain a more hierarchically rich semantic feature, I also tried to introduce an attention mechanism called FRB. Due to the constant pooling of layers inside the CNN, many pixel points of information are constantly lost, and for defects, these details are important. The introduced FRB achieves this loss by retaining multi-layer semantic features to compensate for this loss. The FRB has two main branches as shown in Figure 3. The size of the input feature map is assumed to be $W \times H \times C$, where W , H and C are the width, height and number of channels respectively. For the left branch, a $1 \times 1 \times (C/2)$ convolution layer (i.e. the bottleneck layer) is added to reduce the number of channels of the input feature map from C to $(C/2)$. The bottleneck layer is specifically designed to reduce the computational complexity. Next, the feature map is sampled from $W \times H \times (C/2)$ to $2W \times 2H \times (C/2)$ using 2×2 deconvolution with stride 2. The spatial

resolution is reconstructed by the inverse convolution layer, and the information is recovered after pooling the layers. Third, a 3×3 stride 2 convolution is used to output a $W \times H \times (C/2)$ feature map. Finally, a bottleneck layer is added to recover the number of channels to c . Thus, the left branch consists of one anti-folded and three folded layers, each followed by a Relu layer. The purpose of the left branch is to recover texture information through the anti-fold product layer. By upsampling and downsampling the features separately, and finally adding the input tensor, the three are fused to obtain the output. The output feature map of FRB is the same scale as the input, but more informative and semantically richer, and its use as a plug-and-play module is also very flexible and versatile. The process is shown in Figure 3-6.

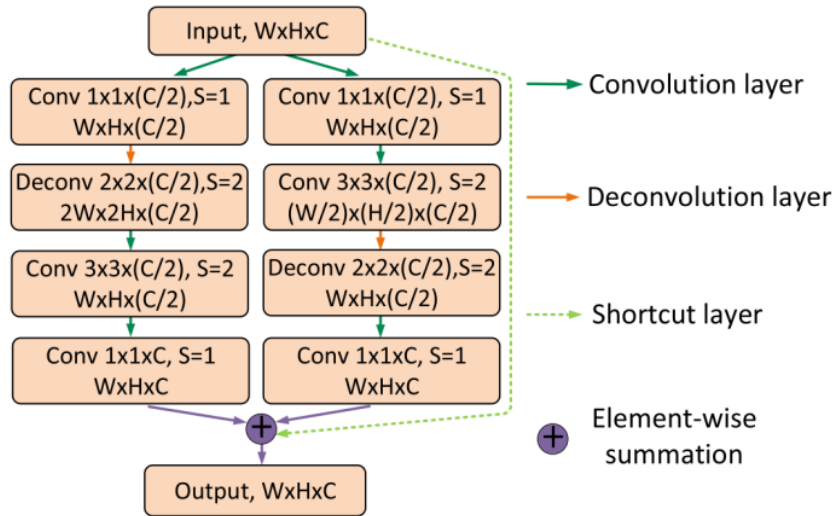


Fig.3-6 FRB structure

CHAPTER 3.6 Analysis of plastic pill bottle cap dataset and defect characteristics

This paper focuses on the task of identifying surface defects on the caps of plastic medicine bottles based on deep learning. In this type of image processing task, the choice and quality of the dataset largely affects the training results of the model, and a representative and clean image can greatly improve the recognition effect of the model. Due to the difficulty of obtaining and collating images of defects on the surface of plastic bottle caps, there are no authoritative public datasets available, so two datasets were

selected for this paper. and a private dataset consisting of six classes and 2200 images. During the training process, feature information of the images in the training set is learnt, loss values are calculated, network weights are updated by back-propagation, predictions are made on the validation set using each model, the prediction accuracy of the validation set is recorded, the parameters corresponding to the best model are selected and used to adjust the model parameters, and finally the model is tested using the test set after convergence. Figure 8 below shows some of the images of the steel surface defects

Defect recognition on the surface of plastic bottle caps has certain similarities to other types of image recognition, but also has certain unique features, because the defect features are not obvious, the contrast with the surrounding normal area features is too low, some categories of defects are small, and the colour is not conspicuous, and sometimes even the naked eye can be confused, when judging defects usually need to combine a lot of background feature information, and between different categories of defects are often recognized as the same category because of similar shapes, and similar defects are easily recognized as different categories because of the difference in defect size, and when the defect is small, need to focus on the acquisition of small target feature information, different categories with similar defect shapes and the same category with a large difference in defect size. image as shown in Figure 3-7.

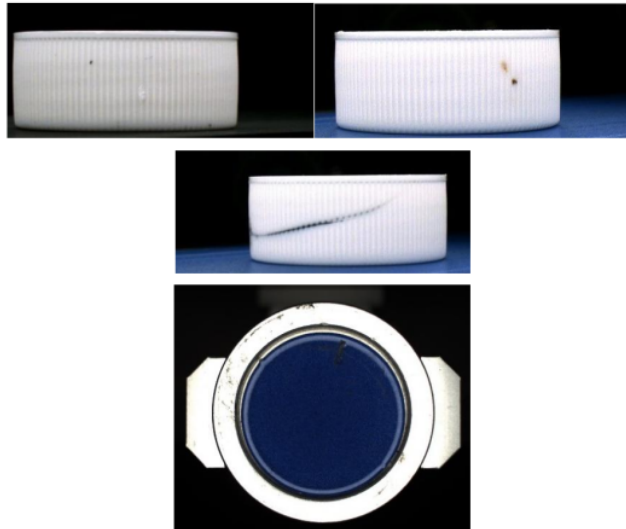


Fig.3-7 Plastic bottle cap defect dataset sample

In order to gain a more intuitive understanding of the different classes of defective feature information and to obtain a visual interpretation visually, a gradient-weighted class activation mapping method [39] (Grad-CAM) is used for visualization.

Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron, discriminating the image location for specific attention decisions.

CHAPTER 3.7 Experimental Environment

At present, scholars in the field of artificial intelligence continue to explore deep learning with enthusiasm, and various open source frameworks have emerged. Pytorch has a large number of users and an active community, and one of the most important features of the Pytorch framework is that it uses a dynamic computational graph structure, which allows developers to adjust the network structure flexibly without having to rebuild the entire network from scratch. The Pytorch API interface is simple, consistent and easily extensible, and does not require complex and repetitive code calls during application. In addition, Pytorch's official documentation is thorough and easy to understand, and its large community of users provides a great platform for debugging and solving specific problems. The experiments were conducted on the Pytorch deep learning framework using the ColabPro platform provided by Google and accelerated training using a GPU, P40.

CHAPTER 4 DESIGN AND IMPLEMENTATION

CHAPTER 4.1 Model

In the current mainstream research work, image classification is mainly divided into two aspects: on the one hand, classification research based on convolutional deep networks to extract global features of images, and on the other hand, classification research combining attention mechanism to recognize the input image features according to their importance. In this paper, we propose a convolutional network model combined with an attention mechanism to classify the cap of a plastic medicine bottle, while exploring the significant features of the image categories and taking into account the subtle differences in the distinction between the categories.

The input is a $512 \times 512 \times 3$ -channel image, and the image is first segmented by a segmentation network with a backbone of Res34 to obtain a $64 \times 64 \times 1$ mask and a $64 \times 64 \times 1024$ volumetric vector to be fed into the classification network. In order to prevent the pre-training classification network from returning worthless gradients, a coefficient of 0 is added to the gradients returned by the classification network during training, which is equivalent to adding a diode one-way channel. After obtaining the mask map, the 1025 channel features are fed into the classification network, and after a series of transformations, a $1 \times 1 \times 66$ one-dimensional vector is finally obtained and sent to the fully connected layer for scoring to determine whether the image is a defective sample. The overall structure is shown in Figure 4-1.

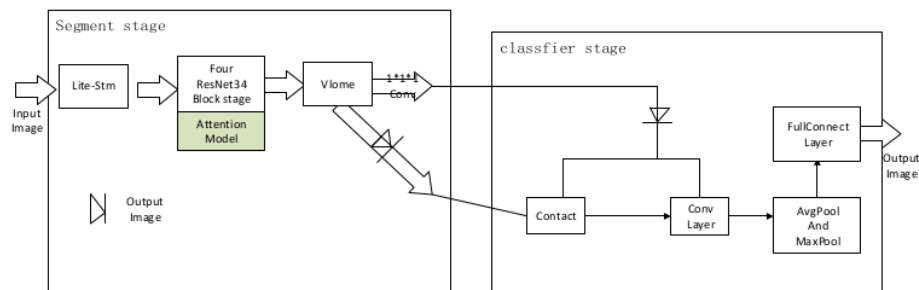


Fig.4-1 Total Model structure

CHAPTER 4.2 Segment networks

In order to avoid bottlenecks in the representation, it is desirable that the image size

does not decay sharply when fed into the convolutional network. If the information in the flow-through layers (especially the convolutional layers) is overly compressed, a large amount of information will be lost and the training of the model will be difficult [40].

Therefore, in order to avoid "bottlenecks" in the design of the front-end Stem structure of this model, the Stem structure is built as shown in Figure 4-2. Two 3×3 convolutional layers with stride=2 and three 1×1 convolutional kernels are used to minimise the computational effort. After convolution of the Stem structure, the width and height of the feature map is reduced and the number of channels is increased from $512 \times 512 \times 3$ to $128 \times 128 \times 64$.

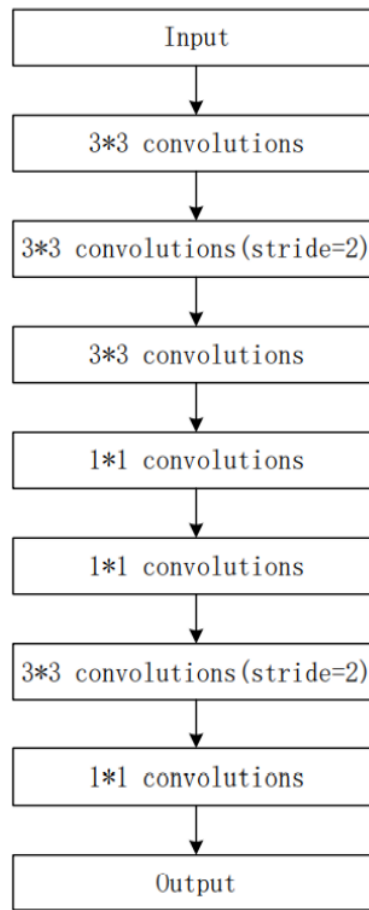


Fig.4-2 Lite-Stm in ResNet34

Then the $128 \times 128 \times 64$ features are obtained by the three-layer residual block, the

64*64*128 features by the four-layer residual block, the 64*64*256 features by the six-layer residual block, and the 64*64*512 feature volume by the three-layer residual block, and then the 1*1*64 mask map is obtained by the 1*1*1 convolution kernel, where the volume and mask The gradient passed after the graph is not affected by the classification network.

CHAPTER 4.3 Training strategy

In this project, we tested both all-labelled and mixed-labelled cases. All-labelled refers to an image with both positive and negative labels indicating whether the sample has defects and pixel-level segmentation labels indicating defects, weakly-labelled refers to only positive and negative labels, unlabelled refers to no labels at all, and mixed-labelled refers to positive and negative labels but only a small number of pixel-level segmentation labels. To ensure a balanced number of positive and negative samples, a training strategy of balancing positive and negative samples is used to ensure a balanced number of positive and negative samples are thrown into the network

When pixel-level labelled images are unavailable, segmentation learning can be optionally disabled by setting the weights of its loss function to zero. For defect-free training samples, the segmentation output for all pixels should always be zero so that the network can train the segmentation sub-network in supervised mode, even in the case of defect-free images.

During supervised training, at the beginning of training, the segmentation network has not yet converged to produce a relatively correct output, and the output mask map is an invalid output, so at this point, the classification network is also not producing a meaningful output. So the gradients back propagated from the classification layer to the segmentation network at this point may have a negative impact on the segmentation part. Adding a weight of 0 to these gradients at the beginning of training prevents these incorrect gradients, and this weight can be slowly reset back to 1 as the training process progresses.

CHAPTER 4.4 Label weights

Since some defects are too small and it is difficult to type fine and accurate labels, and on the other hand, labels that are too fine are likely to lose features during downsampling, I apply a split loss weight mask obtained by a distance transformation algorithm to the labels in training. The effect is to give greater weights, i.e. more attention, to the centre of the labelled region, and smaller weights, less attention, to the outer parts. In other words, this approach is similar to an artificially set attention mechanism for labelling, and this approach reduces the effect of blurring arising from defective edges on the recognition effect, which is formulated roughly as (4-1).

$$L_{seq}(pix) = \Omega\left(\frac{Dis(pix)}{Dis(pix_{max})}\right) * \hat{L}(pix) \quad (4-1)$$

$\hat{L}(pix)$ is the original loss, $Dis(pix)$ is the distance between pixel point and centre point.

CHAPTER 4.4 Loss Function

In this paper, although the deep learning model is an end-to-end model due to the existence of two stages of the network of segmentation and classification, so the two loss functions of segmentation and classification L_{seq} and L_{cls} are generated, and the final loss function can be derived from the merger of these two loss functions, Totalloss show in Eq. (4-2), (4-3). (4-4).

$$L_{total} = \lambda * \gamma * L_{seq} + (1 - \lambda) * \delta * L_{cls} \quad (4-2)$$

$$\lambda = 1 - \frac{n}{n_{ep}} \quad (4-3)$$

$$\gamma = \begin{cases} 1 & \text{Negative image} \\ 1 & \text{positive image with pixel-level lable} \\ 0 & \text{positive image with no pixel-level lable} \end{cases} \quad (4-4)$$

Meanwhile the segmentation and classification network Loss uses Focalloss as the benchmark loss to reduce the loss results of simple and easy classification samples, so that the model computational resources are tilted towards difficult classification samples, thus increasing the attention weight of difficult classification samples. It also proposes a

way to dynamically adjust the learning rate calculation, which can guarantee the performance of the neural network while reducing the training time to a certain extent.

Focal Loss was introduced to address the problem of imbalance between the number of hard and easy samples (note that there is a difference between imbalance between positive and negative samples) and can be used in a wide range of applications.

A single-stage target classifier will usually have a large number of classification targets, with only a very small number of positive samples, and a very unbalanced number of positive and negative samples. To solve the problem of unbalanced positive and negative samples, we usually add a parameter in front of the cross-entropy loss to balance the weights of positive and negative samples but this does not solve the whole problem. In total, samples can be classified into the following four categories based on positive, negative, hard and easy: positive hard, positive easy, negative hard and negative easy. Although the parameters introduced balance the positive and negative samples, they do not help the imbalance between the hard and easy samples. These hard samples have a low loss, but due to the extreme imbalance in numbers, the number of easy split samples is relatively too large and ultimately dominates the total loss. The idea of Focal Loss is simply to reduce the loss of the high-confidence (p) samples by some amount, with the final Focal Loss taking the following form in Eq. (4-6)

$$L_{seq}, L_{cls} = -y(1-p)^{\gamma} \log(p) - (1-y)p^{\gamma} \log(1-p) \quad (4-5)$$

CHAPTER 5. RESULTS AND DISCUSSION

CHAPTER 5.1 Evaluation indicators

The evaluation of the experimental results is based on mAP, precision and accuracy indicators, which are calculated as in equations (5-1), (5-2) and (5-3).

$$mAP = \frac{1}{k} \sum_{i=1}^k \frac{TP}{TP + FN} \quad (5-1)$$

$$Precision = \frac{TP}{TP + FP} \quad (5-2)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5-3)$$

where TP, FP, TN, FN stand for true positive, false positive, true negative, false negative respectively

Accuracy is the proportion of all positive forecasts that are correct. For example, if you forecast rain 10 times and it only rains 3 times, your forecast accuracy is 0.3 because $TP = 3$, $FP = 7$, $TN = 3$, $FN = 7$.

Accuracy, which is the proportion of correct forecasts out of the total number of forecasts. The higher the value of the evaluation metric, the more accurate the model prediction

Auc is the area enclosed by the roc curve and the axis below

The experiments were conducted using PyTorch to build the network model, using GPU acceleration and a Tesla-V100 graphics card, with a batch size of 32 and 70 training rounds. The optimiser was SGD, and the initial learning rate was 0.001.

CHAPTER 5.2 Experimental results in full pixel-label

In order to evaluate the effectiveness of the proposed network model, a comparative ablation experiment was conducted by controlling the combination of attention and backbone modules and some training strategies respectively. The backbone network is represented by the VGG structure and the ResNet structure, while the attention is selected from the more representative attention models CBAM and FRB attention for obtaining high and low level semantics, which can be used in combination with classical network models to achieve enhancement of useful image features and improve the performance of

the network in visual tasks. In this paper, the attention network is obtained by adding the attention model CBAM to the basic structure of the residual network ResNet34 using a cascade approach, thus enhancing the recognition and classification performance of the network model. The six models are combined in a cascade, and the attention network is trained on the bottle cap defect detection dataset with all pixel-level labels input to obtain the mAP and loss and AUC training plots for the training and validation sets as shown in Figures 5-1 to 5-8 respectively.

Details of the final results of the experiment are shown in Table 1

Table 1. Model performance comparison under different attention and backbone networks

model	mAP
VGG-SegDec-Net	91.85%
Res-SegDec-Net	92.55%
VGG-SegDec-Net-CBAM	92.42%
VGG-SegDec-Net-FRB	97.00%
Res-SegDec-Net-CBAM	97.29%
Res-SegDec-Net-FRB	99.3%

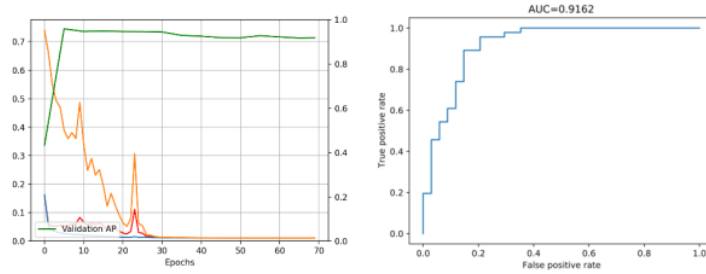


Fig.5-1,2. VGG-SegDec-Net

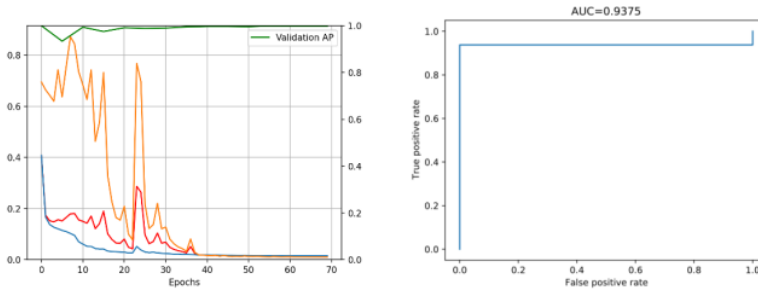


Fig.5-3,4. Res-SegDec-Net

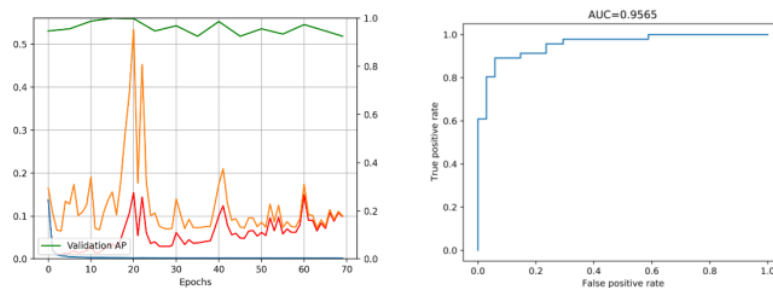


Fig.5-5,6. VGG-SegDec-Net-CBAM

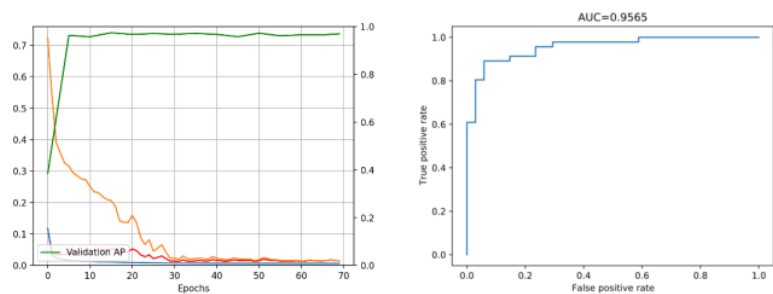


Fig.5-7,8. VGG-SegDec-Net-FRB

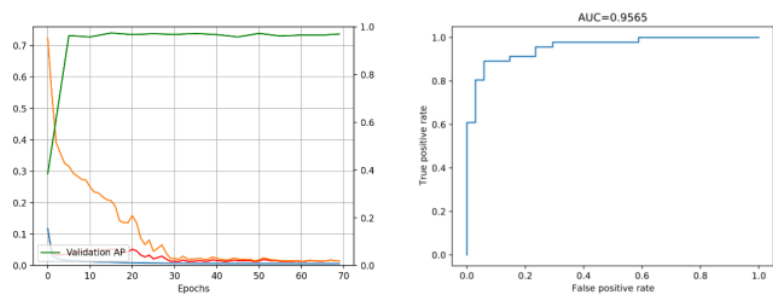


Fig.5-9,10. Res-SegDec-Net-CBAM

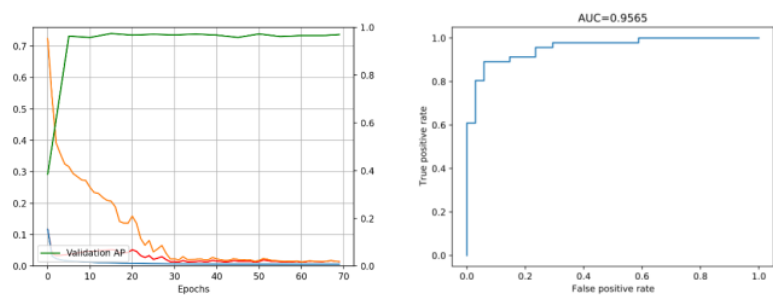


Fig.5-11,12. Res-SegDec-Net-FRB

Details of the final results of the experiments are shown in Table 1 At the end of each training round the data from the validation set was used to verify the strengths and weaknesses of the model, and if the validation results from the validation set were superior compared to the best model. A sample of the experimental output is shown in the figure 5-13.

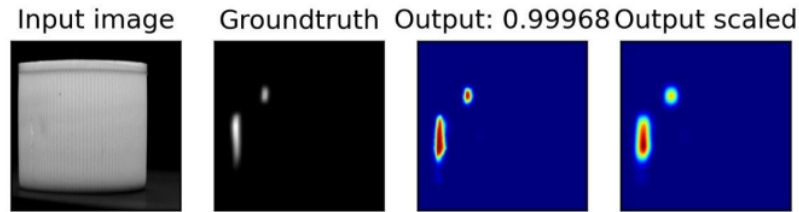


Fig.5-13,4. Output Sample

If the validation result is superior to the previous result, the model is saved for that round. Training is stopped when the set epoch value is reached, and the final

The model parameters with the best performance on the validation set are obtained, and the test set is used to calculate the model accuracy on the copy of the model for that round.

The test set accuracy is obtained. The experiments use the test set accuracy as the performance evaluation criterion. Test set accuracy represents the number of samples in the test set for which the model predicted correctly divided by the total number of samples in the test data set. Provided that the improved focalloss function is also used, the 99.3% accuracy was achieved on the test set, a 6.75% improvement over the unused accuracy of 92.55%.The change in accuracy and loss value of the training and validation sets during the convergence process can be seen that the model with different backbone and attention mechanism for network training can obtain the best accuracy model as Res-SegDec-Net-FRB, which is the Res34-based backbone segmentation classification model with the introduction of attention mechanism.

CHAPTER 5.2 Experimental results in 30% pixel-label

Since it is difficult to obtain very standard and detailed labels in real-life situations, many researchers have worked on developing unsupervised, weakly supervised and mixed-supervised network models. In this paper, we experiment with three models with

few pixel labels, only 30% of the pixel-level segmentation labels, and all positive and negative labels.

Table 2 The case of some models with 30% pixel-level labels

model	mAP
VGG-SegDec-Net	64.53%
Res-SegDec-Net-CBAM	77.37%
Res-SegDec-Net-FRB	84.14%

The specific indicators for each of these classes in Res-SegDec-Net-FRB are shown in Table 3 below:

Table3. 30% Mixed supervised conditions of Res-SegDec-Net-FRB for each type of identification

Class	FP	FN	AP
1	2	0	99.92%
2	180	41	21.60%
3	10	19	85.00%
4	0	5	99.96%
5	3	13	98.37%
6	0	0	100%
Sum/Avg	195	78	84.14%

As can be seen from the table, with only thirty percent of pixel labels, the proposed model still achieves an accuracy of 86.50% on the test set, which is an improvement of about 6.76% and 19.61% respectively compared to the comparative experimental results of other model algorithms. The experiments demonstrate the superiority of the model proposed in this paper in the defect recognition task with less segmented labels.

The results of the table experiments show that the mAP metric of the Res-SegDec-Net-FRB model proposed in this paper is better than the comparison model and has a greater improvement than the benchmark SegDecNet with backbone replacement.

This chapter focuses on the idea of building a network model based on the attention mechanism, and constructs a parallel network with an attention mechanism and a convolutional module to achieve an effective fusion of global convolutional features and

local features with more image discriminative information, so that the weight of background invalid areas in the image is reduced, and the robustness of the network is compared for different label amounts.

The network's robustness is compared for different label amounts, and the learning of key local areas is enhanced by the label weights. Finally, the proposed network is compared with a number of experimental results to illustrate the effect of the proposed network on the classification accuracy with different indicators

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

With the rapid development of the pharmaceutical industry, drug safety regulations are becoming stricter and stricter, but due to technical limitations, defective products inevitably occur during the production of pharmaceutical bottles and account for a not insignificant proportion, i.e. the surface of plastic pharmaceutical bottle caps is prone to defect formation. Defects on the plastic surface include excessive oxidation, stains, cracks, dents, scratches and spots. Surface defects not only affect the appearance of the product and reduce the economic efficiency of the factory, but also seriously affect the quality of the product when the defects are too large. In this paper, we analyse the accuracy of different convolutional neural networks on the surface defects of plastic bottles and build a suitable network model for the recognition of surface defects on plastic bottles.

This paper presents a series of studies and experiments on the problem of cap defect detection using deep learning techniques based on the construction of a cap defect detection dataset for pharmaceutical bottles, based on a corporate project that I have experienced.

The main results of this paper are as follows:

(1) For the network model of plastic bottle cap defect detection, this paper introduces the attention mechanism on the framework of residual structure, adopts the structure of serial connection and feature fusion for the attention module and the network convolution module, adds the residual connection to solve the gradient propagation problem within and between modules, and adds BN layer before the module output to reduce the effect of overfitting. In this paper, a dataset containing six types of defective images of

pharmaceutical bottle caps was constructed to compare the experimental results and to illustrate the improvement of the network for defect detection and classification by comparing the training accuracy curve, test set evaluation and the basis of classification decision.

(2) The improved focalloss cross-entropy loss function algorithm is used to recalculate the loss value and dynamically adjust the learning rate change using the loss value as a guiding parameter. The cross-entropy loss function is improved by adding adjustment coefficients to reduce the loss of simple and easy-to-classify images, so that the number of training rounds is used as an important parameter to adjust the learning rate during the training process, accelerating the decline of the learning rate when the number of training rounds is increased and slowing down the decline of the learning rate when the number of training rounds is too low. The experimental results show that this method not only reduces the training time of the network, but also improves the convergence effect of the network.

(3) Using a multi-scale convolutional kernel and a two-stage segmentation classification defect detection network model backbone to achieve feature fusion of the backbone network, the improved FRB attention module is inserted into the horizontal connection of Res34 to increase the perceptual field of the shallow network and improve the accuracy of the network model for surface defect image recognition.

(4) To address the problem of unbalanced samples in the defective dataset, a training strategy of throwing in positive and negative samples 1:1 was adopted. The results were visualised by using Grad-CAM technique, which proved that the fusion of the underlying feature information had improved the recognition accuracy, the validation set had better convergence and the test set had higher accuracy.

CHAPTER 7. REFLECTION

The experimental results of this study are satisfactory, but as the problem of simple defect detection is essentially a dichotomous problem, as long as the defect is identified, and as the experimental data in this paper was only collected with an industrial camera capturing the picture, no other external factors such as ambient lighting, product colour etc. were taken into account.

There is another aspect of the problem that exists, as various medicine bottle packaging styles, defects, vary, the data in this paper are all from the same plastic medicine bottle caps, from this perspective the generalisation of the conclusions may not be well validated, perhaps it is difficult to get very satisfactory results after replacing the data set, it is known that in deep learning network models affect the segmentation classification effect results of the most important perhaps It is known that it is not the network itself, but perhaps more the adaptability of the network structure to different kinds of datasets, and more experiments are needed to investigate and prove this issue.

There is still much room for future research on this topic, which can be summarised in the following two points.

(1) The images of the defective pill bottle dataset constructed in this paper contain six images of different angles and styles, and more images of pill bottle cap types can be collected in the subsequent research work for further study. In addition, other defective targets such as steel, fabric, etc. can be classified and generalized.

(2) In terms of the optimization of the focalloss algorithm, there is still room for further exploration of the two-stage loss approach designed in this paper. In the subsequent research work, we can combine the switching time nodes of the two loses, as well as the gradient direction and learning rate after switching. In addition, it is possible to explore the optimisation of the network model by combining the implementation of other optimisers.

REFERENCE

- [1] Li Yong . Research on online inspection system of glass bottles based on machine vision [D]. Master's thesis, Shandong University, 2012.
- [2] ZHOU Xianen,WANG Yaonan,ZHU Qing,WU Chengzhong,PENG Yu. Research on bottle mouth defect detection method based on machine vision[J]. Journal of Electronic Measurement and Instrumentation,2016,30(05):702-713.
- [3] Li Dan,Bai Guojun,Jin Yuanyuan,Tong Yan. Machine vision-based packaging bag defect detection algorithm based on machine vision[J]. Laser and Optoelectronics Advances in Science,2019,56(09):188-194.
- [4] Fukushima K . Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4):193-202.
- [5] Lecun Y, Boser B, Denker J, et al. Backpropagation applied to handwritten zip code Recognition[J]. Neural Computation, 1989, 1(4):541-551.
- [6] Lecun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in Neural Information Processing Systems, 1990, 2(2):396-404. Hecht-Nielsen R. Theory of the backpropagation neural network[M].
- [7] Neural networks for perception. Academic Press, 1992:65-93
- [8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25:1097-1105.
- [9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. Proceeding of the International Conference on Learning Representations, 2015:473-491.
- [10] He K, Zhang X, Ren S, et al. Deep residual Learning for image recognition[C]. Proceeding of the International Conference on Computer Vision and Pattern Recognition, 2016:770-778.
- [11] Howard A G, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural

- networks for mobile vision applications[C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017:432-445.
- [12] Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018:4510-4520
 - [13] Howard A, Sandler M, Chu G, et al. Searching for mobileNetV3[C]. Proceeding of the 17th IEEE International Conference on Computer Vision(ICCV), 2019:1314-1324.
 - [14] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: a review. 2020. ArXiv:2007.02500
 - [15] Chalapathy R, Chawla S. Deep learning for anomaly detection: a survey. 2019. ArXiv:1901.03407
 - [16] Czimmermann T, Ciuti G, Milazzo M, et al. Visual-based defect detection and classification approaches for industrial applications—a survey. *Sensors*, 2020, 20: 1459
 - [17] Zhou D H, Wei M H, Si X S. A survey on anomaly detection, life prediction and maintenance decision for industrial processes. *Acta Automatica Sin*, 2013, 39: 711–722
 - [18] Tao X, Hou W, Xu D. A survey of surface defect detection methods based on deep learning. *Acta Autom Sin*, 2021, 47: 1017–1034
 - [19] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv*, 2009, 41: 1–58
 - [20] Hawkins D M. Identification of Outliers. London: Chapman and Hall, 1980
 - [21] Ruff L, Kauffmann J R, Vandermeulen R A, et al. A unifying review of deep and shallow anomaly detection. *Proc IEEE*, 2021, 109: 756–795
 - [22] Bulusu S, Kailkhura B, Li B, et al. Anomalous instance detection in deep learning: a survey. In: *Proceedings of the 42nd IEEE Symposium on Security and Privacy* Princeton, 2020
 - [23] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative

- localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2921–2929
- [24] B. Y. GOODFELLOW I, COURVILLE A. Machine learning basics “Deep learning,” vol. 1, pp. 98-64, 2016.
 - [25] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 618–626 13
 - [26] Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise. 2017. ArXiv:1706.03825
 - [27] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:936-944.
 - [28] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]. Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2015: 2017-2025.
 - [29] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
 - [30] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning. PMLR, 2015: 448- 456.
 - [31] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]. Proceeding of the International Conference on Learning Representations, 2016:1-13.
 - [32] Lee J B, Rossi R A, Kim S, et al. Attention models in graphs: A survey[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(6): 1-25.
 - [33] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C]. 28th Annual Conference on Neural Information Processing Systems. 2014: 2204-2212.
 - [34] Zhu X, Cheng D, Zhang Z, et al. An empirical study of spatial attention

- mechanisms in deep networks[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6688-6697
- [35] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
 - [36] Laskar Z, Kannala J. Context aware query image representation for particular object retrieval[C]. Scandinavian Conference on Image Analysis. Springer, Cham, 2017: 88-99.
 - [37] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
 - [38] Park J, Woo S, Lee J Y, et al. BAM: Bottleneck attention module[C]. British Machine Vision Conference (BMVC). British Machine Vision Association (BMVA), 2018
 - [39] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.
 - [40] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016:2818-2826.

FINAL GRADE

/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

