An improved Vision Transformer-based method for classifying surface defects in hot-rolled strip steel

Xinglong Feng\textsuperscript{1a*}, Xianwen Gao\textsuperscript{1b*}, Ling Luo\textsuperscript{2c*}

\textsuperscript{1}College of information Science and Engineering Northeastern University, Shenyang 110819, China
\textsuperscript{2}Moviebook Technology Co., Ltd., Beijing 110027, China
\textsuperscript{a*}1610238@stu.neu.edu.cn, \textsuperscript{b*}gaoxianwen@ise.neu.edu.cn
\textsuperscript{c*}ling_luo@moviebook.cn

Abstract—A new Vision Transformer(ViT) model is proposed for the classification of surface defects in hot rolled strip, optimizing the poor learning ability of the original Vision Transformer model on smaller datasets. Firstly, each module of ViT and its characteristics are analyzed; Secondly, inspired by the deep learning model VGGNet, the multilayer fully connected layer in VGGNet is introduced into the ViT model to increase its learning capability; Finally, by performing on the X-SDD hot-rolled steel strip surface defect dataset. The effect of the improved algorithm is verified by comparison experiments on the X-SDD hot-rolled strip steel surface defect dataset. The test results show that the improved algorithm achieves better results than the original model in terms of accuracy, recall, F1 score, etc. Among them, the accuracy of the improved algorithm on the test set is 5.64% higher than ViT-Base and 2.64% higher than ViT-Huge; the accuracy is 4.68% and 1.36% higher than both of them, respectively.

1. Introduction
Currently, artificial intelligence is increasingly used in industrial inspection. Algorithms using artificial intelligence have been gradually replacing manual work in the field of hot rolled strip defect detection due to their more automated advantages. Hot rolled strips are strips and plates produced by hot rolling, which have important applications in many fields such as automobiles and home appliances. The detection of surface defects in hot rolled strip is an important part of hot rolled strip production and is closely related to the surface quality of the strip.

Different kinds of strip surface defects often require different methods of treatment. For example, red iron sheet defects need to be avoided by checking the heating temperature to ensure the proper rolling temperature, while severe slag inclusions defects have to be avoided by removing a portion of the strip with defects. If the strip surface defects are not treated correctly in time, it may lead to the problem of broken strips in the subsequent processing stage, affecting the next production; or the surface quality of the finished strip cannot meet the customer has to reduce the price of sales, thus bringing losses to the enterprise. Therefore, the classification of strip steel surface defects is of great importance. The traditional method of classifying surface defects of hot-rolled strip often relies on manual work, which can meet the needs of production to a certain extent, but there are problems such as low efficiency, long hours of work boring and easy to cause visual fatigue of staff, has been gradually replaced by the strip defect detection system, a typical strip defect detection system as shown in Figure 1.

\*Content from this work may be used under the terms of the CreativeCommons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
The steel strip defect detection system in Figure 1 has been widely used in actual production [1,2], and the defect detection system is mainly composed of detection devices, servers, and consoles. Among them, the inspection device includes industrial camera, industrial light source, protective cover, etc. The inspection devices are symmetrically distributed on the upper and lower surfaces of the strip, and each inspection device generally contains 5-7 industrial cameras. As the strip passes through the inspection device, the industrial camera's shield opens and the camera takes high-speed pictures of the strip surface at a rate of more than 20 frames per second and passes the picture data to the server via fiber optic or network cable, where the algorithm within the server collates and analyzes the picture data and later displays the results to the operator via the console. In this process, the algorithm within the server is the key; in this regard, a lot of research has been conducted by domestic and foreign scholars, and certain results have been achieved.

2. Related Work

In recent years, scholars have introduced convolutional neural networks (CNNs) for strip steel defect classification. Fu G et al [3] proposed a compact and effective convolutional neural network (CNN) model that emphasizes the training of low-level features and combines multiple sensory fields for fast and accurate steel surface defect classification. Y Liu et al [4] used GoogLeNet as the base model and added constant mapping to it to enhance the algorithm to some extent. The network achieves a speed of 125 FPS, which fully meets the real-time requirements of the actual steel strip production line. Konovalenko et al [5] used a deep learning model based on ResNet50 as the base classifier to perform classification experiments on planar images with three types of impairments, and the results showed that the model has excellent recognition ability, high speed and accuracy at the same time. X Feng et al [6] used the RepVGG model as the base classifier and added spatial attention to enhance the classification ability of the original model, and the improved model achieved the best results in terms of accuracy, recall, and other metrics.

The CNN-based model can automatically perform extraction of strip surface defect features, which improves the efficiency compared to manual feature extraction. However, due to the nature of the convolution operator, the resulting feature maps are locally sensitive; that is, CNNs are more adept at extracting locally valid information in images, but lack an overall grasp of the input data itself. If the algorithm needs to improve the CNN's ability to characterize the overall features of the image, it needs to use larger convolutional kernels and deeper convolution to expand the perceptual field of the model. This can lead to a dramatic increase in model complexity and even cause a dimensional catastrophe that makes training unconvergent.

To solve the above-mentioned problems caused by CNN induction bias (invariance of convolutional operations), scholars have used the Transformer model, which is entirely based on the attention mechanism, in the field of image classification in computer vision. The Transformer model [7], proposed by Google in 2017, was earlier used in Natural Language Processing (NLP) and achieved brilliant results in this field, which has now become a mainstream algorithm in the field of natural language processing. Inspired by this, the Vision Transformer (ViT) [8] model first applied...
Transformer to the field of image classification, and when pre-trained on a large amount of data and migrated to multiple medium-sized or small image recognition benchmarks, ViT yielded excellent results compared to state-of-the-art convolutional networks, while requiring relatively less computational resources for training.

Although ViT has some advantages over CNN, the algorithm requires a large amount of data for pre-training, and direct training on a small data set is not effective. In the problem of classifying surface defects of hot rolled steel strips, large-scale data sets for pre-training ViT algorithm do not exist yet, and the features of surface defects of steel strips are very different from those of existing large data sets, which makes it impossible to solve the problem of classifying surface defects of steel strips by pre-training on existing large data sets and then migrating them. Therefore, how to improve the ViT algorithm and enhance its classification capability with a small amount of data is the key to the problem.

3. Our Method

3.1. Vision Transformer

The design idea of Vision Transformer (ViT) is simple: first, the input image is divided into patches of fixed size, then the embedded patches are obtained by linear transformation, next, the embedded patches of the image are added with additional vectors (class token) and location information for classification into the ViT encoder for feature extraction. The structure of ViT is shown in Fig. 2.

![Fig. 2 The basic structure of ViT](image)

The original input image is divided equally into several small pieces, and each piece is straightened into a 1-dimensional vector. In this process, the width and height of each block need to be the same, and the original image needs to be divided into an integer number of blocks. After straightening, each vector is transformed linearly, and then a vector for classification is added, stitched together and position information is added. The collated vectors are fed into the Transformer encoder, and the MLP header, which contains two fully-connected layers, outputs the category to which the image belongs.

3.2. The Improvement of ViT

The original ViT model requires large-scale datasets for pre-training, which is very limited in the field of strip steel defects where large-scale datasets are lacking for pre-training. In order to optimize the effectiveness of the ViT model in the field of strip defects classification, this paper improves the original ViT model. The goal of this paper is to improve the classification accuracy index of the algorithm while reducing the computational complexity of the algorithm; so that the algorithm achieves better performance than before the improvement and lays a good foundation for further improvement.

As reducing the computational complexity tends to make the model have less number of parameters; while improving the classification accuracy of the model wants to increase the nonlinear expression capability of the model, which generally requires increasing the number of parameters of
the model. In order to meet the objectives of both model accuracy and efficiency improvement, the design idea of this paper is to reduce the width of the model while increasing the depth of the model.

First, for the original ViT model, we reduce the number of nodes in the encoder. The comparison of the modified results with the original ViT-related parameters is shown in Table 1. From the table, it can be seen that the hidden layer dimension \( D \) and MLP dimension are reduced significantly. Since the hidden layer dimension \( D \) and MLP dimension exist in each encoder layer, these two parameters are highly related to the number of parameters of the final model, so the number of parameters of the changed model will be effectively reduced.

| Model       | Hidden Layer Dimension \( D \) | MLP Dimension | Encoder layer |
|-------------|---------------------------------|---------------|---------------|
| ViT-Base    | 768                             | 3072          | 12            |
| ViT-Large   | 1024                            | 4096          | 24            |
| ViT-Huge    | 1280                            | 5120          | 24            |
| Ours method | 144                             | 100           | 12            |

In order to improve the model nonlinear expression capability, inspired by VGGNet[9] and AlexNet[10], two fully connected layers of the same size are added to the fully connected layer of ViT in this paper, so that the original fully connected structure of ViT with three layers becomes four layers. Compared with ViT-Base, the comparison diagram before and after the improvement is shown in Fig. 3.

(a) Before improvement  
(b) After improvement

Fig. 3 Multi-layer perceptron layer before and after improvement

The fully connected layer containing 3072 neurons in Figure 4(a) is replaced with a fully connected layer containing two 100-neuron layers in Figure 4(b). Although the number of participants in the two fully connected layers containing 100 neurons is approximately three times that of the single fully connected layer with 3072 neurons, the improved model reduces the hidden layer dimension from 768 to 144, and thus the number of participants remains relatively small in aggregate. Since the classification of strip surface defects requires a high level of abstraction of features for effective differentiation; therefore, theoretically a deeper fully connected layer in this paper would be more useful for classifying strip surface defects, which will be verified by experiments below.

4. The Experimental results

4.1 The dataset

The dataset used for the experiments in this paper is from the X-SDD hot-rolled strip surface defect dataset from the literature [6], which contains seven categories of defect data from the hot-rolled strip site. The dataset contains a total of 1360 defect images, of which 70%, i.e., 952 images, are selected as training samples and the remaining 408 images are used as test samples. The pattern of defects is shown in Figure 4, and the defects represented by (a)-(g) are: oxide scale of plate system, red iron sheet, surface scratches, slag inclusions, finishing rollprinting, iron sheet ash and oxide scale of temperature system, respectively. As can be seen from the figure, the data set selected in this paper has
a relatively obvious degree of differentiation, and defects can be accurately classified by human eyes; in practice, trained quality inspectors can classify these defects with nearly 100% accuracy. Therefore, this dataset can be used as a benchmark dataset to validate the improved ViT algorithm proposed in this paper.

Fig.4 X-SDD: Surface Defects Dataset for Hot Rolled Steel Strips

4.2 Experimental environment
The experiments in this paper were conducted in an Intel Core i3-4160 CPU (3.60G HZ) environment with 8G RAM and Win10 operating system. Random horizontal and vertical flips were used to enhance the dataset during the training phase, and the images were normalized using mean and variance. The input data size was uniformly adjusted to 224 pixel × 224 pixel, the batch size was set to 10, the learning rate was set to 0.0001, and the model was optimized using the Adam optimizer. 100 rounds of experiments were trained on the jupyter environment of Anaconda.

4.3 Experimental results and analysis
In this paper, several metrics are selected to evaluate the experimental results, such as January, Macro-recall, Macro-precision, and Macro-F1. Among them, Macro-recall, Macro-precision, and Macro-F1 are obtained by considering the multiclassification problem as multiple two The recalls, accuracies and F1 scores are averaged according to Equations (1)-(3).

\[
Recall = \frac{TP}{TP + FN} \tag{1}
\]

\[
Precision = \frac{TP}{TP + FP} \tag{2}
\]

\[
F_1 = \frac{2PR}{P + R} \tag{3}
\]

where TP denotes true cases, i.e., the total number of positive samples predicted by the classifier as positive TN denotes true negative cases, i.e., the total number of negative cases predicted by the classifier; FP denotes false positive cases, i.e., the total number of negative cases predicted by the classifier; and FN denotes true negative cases, i.e., the total number of positive cases predicted by the classifier. Considering each class as a binary classification problem with the sum of the other classes, we can obtain TP, TN, FP, and FN for each class, which leads to Eqs. (4)-(8).

\[
n_{\text{correct}} = TP_0 + TP_1 + \ldots + TP_{N-1} \tag{4}
\]

\[
Accuracy = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{5}
\]

\[
Macro - Recall = \sum_{i=1}^{N} Recall_i \tag{6}
\]

\[
Macro - Precision = \sum_{i=1}^{N} Precision_i \tag{7}
\]
where $N$ represents the defect type and $n_{\text{total}}$ represents the total number of samples. The experimental results are shown in Table 2.

| Model         | Accuracy | Macro-Recall | Macro-Precision | Macro-F1 |
|---------------|----------|--------------|-----------------|----------|
| ViT-Base      | 71.57%   | 72.81%       | 70.84%          | 69.84%   |
| ViT-Large     | 72.06%   | 67.77%       | 69.22%          | 68.04%   |
| ViT-Huge      | 74.75%   | 74.92%       | 74.16%          | 73.36%   |
| **Our Method** | **77.21%** | **74.38%**  | **75.52%**      | **74.12%** |

As can be seen from Table 2, compared to the three basic structures given in the literature [8], this paper, with simple modifications to the model, makes the model have higher Accuracy, Macro-Precision, and Macro-F1. The accuracy is 77.21%, which is 25.64% higher than ViT-Base, 25.64% higher than ViT-Large 5.15%, and also 2.46% higher than ViT-Huge; this indicates that the improved model improves significantly in the accuracy index. The improved model is only 0.46% smaller than ViT-Huge in the Macro-Recall metric, and is higher than the original ViT algorithm in all other metrics, which indicates that the improved model is balanced in all aspects. Among ViT-Base, ViT-Large, and ViT-Huge, ViT-Huge performs best in all four metrics, which indicates that a sufficiently large number of parameters can make the model achieve better results under the original ViT structure.

5. Conclusion
This paper pioneers the use of ViT method to classify the surface defects of hot rolled strip on X-SDD strip surface defect dataset, and improves the original ViT algorithm to propose a new ViT incorporating multiple fully connected layers. The performance of the improved model is substantially improved compared with the original ViT-Base. Nevertheless, the accuracy of the ViT algorithm in classifying surface defects in strip steel needs to be further improved, and further work will be carried out subsequently.

References
[1] Kostenetskiy P, Alkapov R, Vetoshkin N, et al. Real-time system for automatic cold strip surface defect detection[J]. FME Transactions, 2019, 47(4): 765-774.
[2] Mazur I, Koinov T. Quality Control system for a hot-rolled metal surface[J]. Frattura ed Integrità Strutturale, 2016, 10(37): 287-296.
[3] Fu G, Sun P, Zhu W, et al. A deep-learning-based approach for fast and robust steel surface defects classification[J]. Optics and Lasers in Engineering, 2019, 121: 397-405.
[4] Yi L, Li G, Jiang M. An end-to-end steel strip surface defects recognition system based on convolutional neural networks[J]. steel research international, 2017, 88(2): 1600068.
[5] Konovalenko I, Maruschak P, Brezinová J, et al. Steel Surface Defect Classification Using Deep Residual Neural Network[J]. Metals, 2020, 10(6): 846.
[6] Feng X, Gao X, Luo L. X-SDD: A New Benchmark for Hot Rolled Steel Strip Surface Defects Detection[J]. Symmetry, 2021, 13(4): 706.
[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
[8] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[10] 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.