The 18th International Confernce on Automation Technology (Automation 2021),
November 12-14, 2021, Kimmen, Taiwan

# Using Attention Mechanism on Tool Defect Classification

**Ting-Yu Kuo[1], Shih-Ya Chang[2], and Chen-Kuo Chiang[3, *]**

[1]*kmichael123@gmail.com, Chiayi, Taiwan*
[2]*otischang.true@gmail.com, Chiayi, Taiwan*
[3]*cck.adrian@gmail.com, Chiayi, Taiwan*
*Corresponding author: *cck.adrian@gmail.com*

***Abstract:*** In the field of industrial manufacturing, tool processing is a very widely used and indispensable technology, and the continuous use of the tool will cause the surface to be uneven due to external forces, which is collectively referred to as defect. The tool may have different wear patterns, resulting in inconsistent product quality or even machine damage during processing. Therefore, we classify the tool images based on deep learning method, and refer to the attention mechanism proposed by previous work. We finally compare the accuracy between each method if it has been improved. In this paper, we use Residual Network (ResNet), the most widely used backbone model in the field of computer vision, as the baseline, and add the recent popular attention mechanism based on it. Under two different attention methods with the dataset from Kaggle competition, it can be found that without using pre-trained model, the attention mechanism using DACL can increase the accuracy by 1.7%, while using self-attention can increase the accuracy by 4%.

**Keywords:** defect, deep learning, ResNet, attention mechanism, DACL, self attention.

## 1. Introduction

In recent years, manufacturing processing that use cutting tools are widely used, and while rapid manufacturing through machines, the quality of production must also be considered. The degree of tool wear is one of the key factors affecting the quality of production. The ability to accurately and quickly classify wear patterns will effectively improve production quality. In recent years, a considerable number of deep learning methods have been applied in the industrial field with good results. This paper will focus on the images of tool wear and try to use the attention mechanism which popular in recent years to assist the learning of the model to improve the accuracy of the classification of wear types.

## 2. Related Work

### 2.1 ResNet

Convolutional neural network (CNN) is a deep learning model widely used in image classification. The convolutional layers can effectively extract image features and pooling layers are used to reduce dimension. Then VGG family [1] is one of the known CNN model with a deep network structure. With the small convolutional filter, VGG16 increases receptive field size and gets more image information. However, VGG16 is prone to have the problem of gradient vanished. Thus, K. He et al. proposed a deep residual network (ResNet) [2] in 2015, which can effectively avoid gradient vanished.

With the help of the identity mapping in the residual block, ResNet can keep the parameters updated correctly under the “deep” network architecture and avoid gradient vanished. After ResNet won the first place in ILSVRC in 2015, ResNet has been widely applied in various computer vision tasks. In this paper, ResNet18 is used as the baseline and backbone of two attention methods.

### 2.2 DCAL

As mentioned in the paper [3], the difficulty encountered in facial expression recognition is that intra-class variations and inter-class similarities make the model difficult to learn and recognize. Therefore, DCAL, an attention mechanism, is proposed to improve the model learning and attention to the details of features, and achieve state-of-the-art performance.

In the DCAL attention mechanism architecture, it can be divided into two parts: Context Encoder Unit (CE-Unit) and multi-head binary classification. The CE-Unit is set after the back-bone model and is used to extract more important features. After going through three fully-connected layers, a CE- feature is obtained, and this feature is used as the input of multi-head binary classification. In the first stap of  multi-head binary classification is to expand the CE-feature size to 2x the output of backbone model, then do softmax on the pairwise values to obtain the importance ratio in the feature. Finally, multiply this importance ratio by the output of  backbone model, then the attention feature is obtained and can be better reprecentation.

The attention feature is then using in classification in this paper rather than the input of center loss in the previous work.

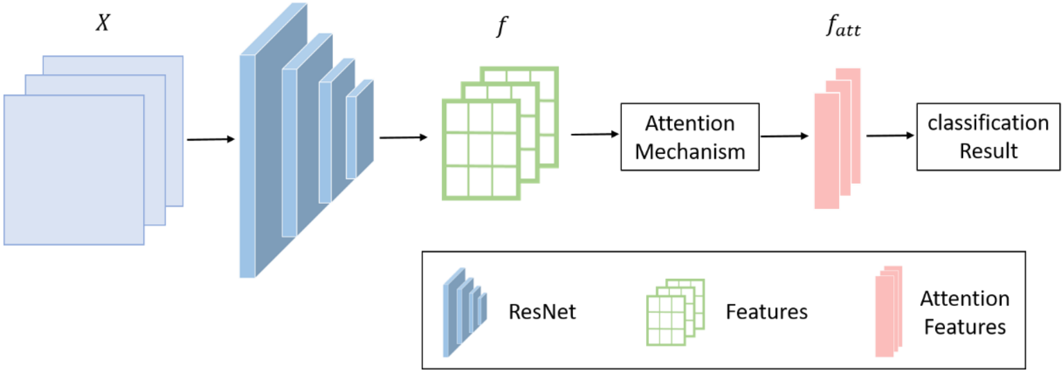

Fig.1. The main architecture of the model.

## 2.3 Self attention

In order to solve the Natural Language Processing (NLP) task, the paper [4] is proposed, which uses the mechanism of self-attention. The main idea of self-attention is to do operation on features such as transformation and multiplication, so that the model can focus on important parts and thus obtain state-of-the-art performance. After then, other deep learning tasks, such as segmentation, object detection, image generation and even classification, people have begun to propose methods of using self-attention and can also achieve better performance.

In this paper, we refer to the self-attention mechanism on the image proposed by [5], which let each component focus on other components and model thereby enhanced the learning of features. Then, the method is compared with the baseline whether the accuracy is improved.

## 3. Method

In this paper, two classification models is proposed that use different attention mechanisms, and the main architecture is shown in Fig.1. First, the RGB image $X$ is used as the input of the model. After the feature extraction of the backbone model ResNet, feature $f$ is obtained. This feature $f$ will be used as the input of the attention mechanism. After attention operation, the output will be an attention feature $f_{att}$ , which can make the model do better classification. The following will introduce two attention mechanisms we used.

### 3.1 DACL Attention

Inspired by [3], we use their attention network as the attention mechanism. This architecture consists of two modules: Context Encoder Unit (CE-Unit) and Multi-head Binary Classification.

CE-Unit is composed of linear layers and hyperbolic tangent. The purpose of CE-Unit is dimensionality reduction of the spatial feature map of facial expressions and eliminate irrelevant information, thereby encode the important latent feature ($e_i$). The structure of CE-Unit is shown in Fig.2.

First, the spatial feature map $f$ obtained from ResNet is transformed into a feature vector $f'$ by average pooling, and then passed through a series of linear layers. The linear layer can extract important features that represent the characteristics of the data. We stacked three linear layers in CE-Unit to find the latent features $e_i$of facial expressions. The number of layers and the parameters of the CE-Unit can be adjusted according to different tasks.

It is said that not every dimension in the deep features obtained by neural network is important to the classification result. To get the importance of each dimension in the deep feature, we use the latent features $e_i$ obtained from CE-Unit to calculate the attention weight. The structure of Multi-head Binary Classification is shown as Fig.2 orange part.

In order to obtain the importance of each dimension in the d-dimensional deep feature, the feature vector $e_i$ calculated by CE-Unit is first upsampling to the dimension of d*2 through a linear layer. 2 means that there are two outputs for each dimension, which are $P_{i,j}^{ex}$ and $P_{i,j}^{in}$. These two represent the exclusion score and the inclusion score in the $j^{th}$ dimension, respectively. Exclusion represents unnecessary information in this dimension, and inclusion represents important information. Next, the two of $P_{i,j}^{ex}$ and $P_{i,j}^{in}$ in each dimension will be normalized with the softmax function, and take the inclusion score normalized by softmax used as the attention weight $\alpha_{i,j}$ . The formula is as follows:

$$a_{i,j} = \frac{e^{P_{i,j}^{in}}}{e^{P_{i,j}^{in}} + e^{P_{i,j}^{ex}}} \tag{1}$$

Finally, a d-dimensional attention weight $\alpha_i$ will be obtained, and then $\alpha_i$ and the deep feature $X_i$ of facial expressions obtained by ResNet will be element-wise multiplication. The important dimensions in the deep feature $f_{att}$ are strengthened.

### 3.2 Self-attention

The second attention mechanism we use is refer to [5].

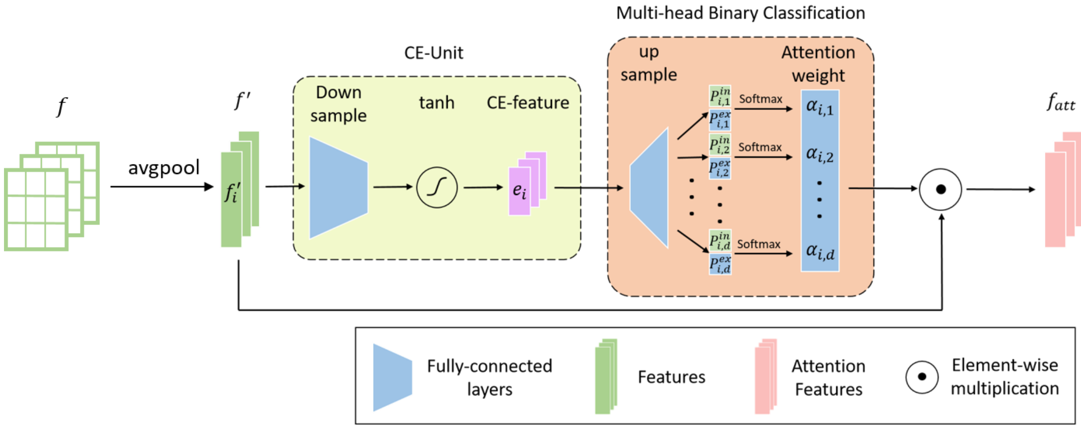

Fig.2. The architecture of DACL attention.

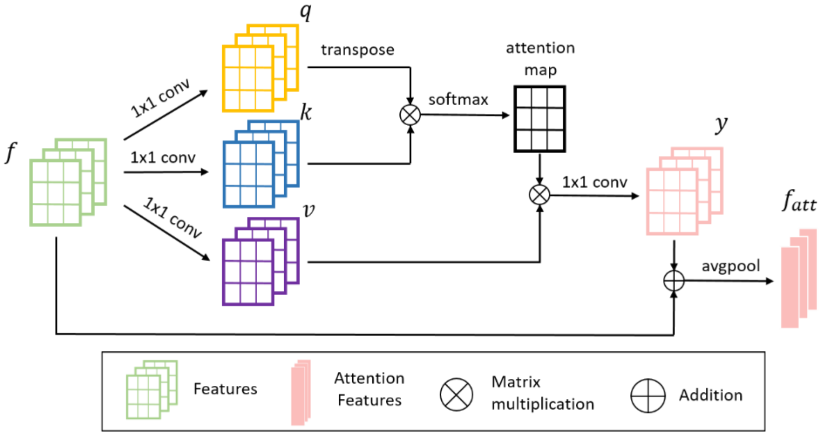

Fig.3. The architecture of self-attention.

First, the feature map $f$ passes through three different 1*1 convolutional layers to obtain three new feature maps $q$, $k$ and $v$. After transposing $q$ to $q^T$, $q^T$ are multiply by $k$ and obtain the special feature maps. Then the softmax operation is performed on each row of the special feature maps and obtain an attention map. The formula is as follows:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \text{ where } s_{ij} = q^T k \tag{2}$$

$\beta_{j,i}$ indicates the extent to which the model attends to the $i^{th}$ location when synthesizing the $j^{th}$ region and $N$ is the number of feature locations of features from the backbone model. The output of the attention layer is $\mathrm{o} = (o_1, o_2, \ldots, o_j, \ldots, o_N)$ and $conv$ is a 1*1 convolution layer, where:

$$o_j = conv\left(\sum_{i=1}^{N} \beta_{j,i}\, v\right) \tag{3}$$

In addition, we further multiply the output of the attention layer by a scale parameter and add back the input feature map. Therefore, the output is given by,

$$y_i = \gamma o_i + f_i \tag{4}$$

where $\gamma$ is a learnable scalar and it is initialized as 0. Finally, the output $y$ is transformed to $f_{att}$ by using global average pooling. It's expected that $f_{att}$ can do better prediction through the output layer.

## 4. Experiment

### 4.1 Dataset

We evaluate the performance of the attention network on the dataset of Kaggle competition named “Severstal: Steel Defect Detection”, a dataset which wanted to classify different category of steel defect and it's similar to the tool defect. In this dataset, we drop the multi label images and balance the number of images in each category. Thus, we totally use 160 images of each class for training and 30 images of each class for testing to solve the four-class classification task.

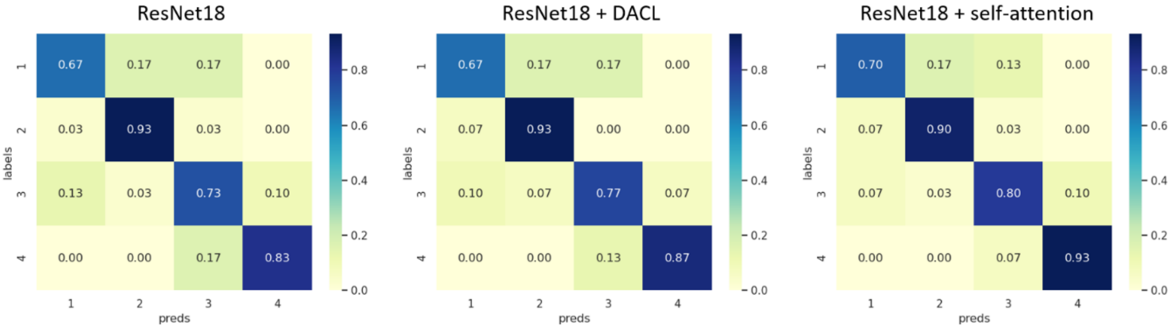

Fig.4. The confusion matrix of each method.

Table 1. Comparison of performance

| Method | Accuracy (%) |
|---|---|
| ResNet18 | 79.3 |
| ResNet18 + DACL | 81.0 |
| ResNet18 + self-attention | 83.3 |

### 4.2 Ablation studies

In the experiment, we use ResNet18 as the backbone model and remove output layer and average pooling layer of ResNet18. All input images are resized to 256 length and width. After the calculation of the model, outputs of the backbone will be feature maps $f$ as shown in Fig.1. Then go through two attention mechanisms to obtain the attention features $f_{att}$. Finally, the attention features $f_{att}$ will be calculated the classifycation result through the output layer. We compare two attention mechanisms with the original ResNet18 as shown in Table 1.

As we can see, with no attention mechanism added in the network, the original ResNet18 can achieve 80.2% accuracy. If adding DACL attention, the accuracy can improve about 1.7% while adding self-attention can improve 4% accuracy.

### 4.3 Confusion matrix

The confusion matrices shown as Fig.4 displays the predicted distribution of each class. As we can see, samples of class 1 has more than 10% of being classified into class 2 or class 3 through each method. It's seems that the features of class 1 are harder to be learned by the model or there are more noisy images in class 1. Though the predicted score on class 2 of self-attention method is lower than other method, the predicted scores on other classes have been improved obviously.

### 5. Conclusion

In this paper, we refer to the attention mechanism in previous works and select two methods to implement on the classification task to improved accuracy on a tool defected small dataset. Maybe in the case of less data, ResNet model can't extract the features very well. After adding the attention mechanism, it can help the model focusing more on useful features and then improve the performance for about 4% accuracy. In the future, we will try to test on others larger datasets to compare the performance of each method and find the useful network architecture which can do better on tool defect classification tasks.

### 6. References

[1] K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint (2014).

[2] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, *Deep Residual Learning for Image Recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition (2016).

[3] Farzaneh, Amir Hossein and Qi, Xiaojun, *Facial Expression Recogni-tion in the Wild via Deep Attentive Center Loss*, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021).

[4] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia, *Attention is All You Need*, Advances in neural information processing systems (2017).

[5] Zhang, Han and Goodfellow, Ian and Metaxas, Dimitris and Odena, Augustus, *Self-Attention Generative Adversarial Networks*, Proceedings of Machine Learning Research (2019).

[6] Wang, Fei and Jiang, Mengqing and Qian, Chen and Yang, Shuo and Li, Cheng and Zhang, Honggang and Wang, Xiaogang and Tang, Xiaoou, *Residual attention network for image classification*, Proceedings of the IEEE conference on computer vision and pattern recognition (2017).

[7] Moldovan, O.G.; Dzitac, S.; Moga, I.; Vesselenyi, T.; Dzitac, I. *Tool-Wear Analysis Using Image Processing of the Tool Flank*. Symmetry 2017, 9, 296.

[8] Wu X, Liu Y, Zhou X, Mou A. *Automatic Identification of Tool Wear Based on Convolutional Neural Network in Face Milling Process.* Sensors (basel), (2019).