Defect Classification Algorithms for Display Manufacturing Based on the Convolutional Neural Network Mixture-of-Experts Model

Yunlong Li, Ting Wang, Jie Wang, Yanming Zhang, Yuyu Liu, Xingqun Jiang
BOE Technology Group Co., Ltd., Beijing, China

Abstract

In the field of industrial display manufacturing, a large number of proprietary defect identification models for detailed scenarios have generated due to the differences in production lines, environments and defect types. The sparse network structure was designed to classify defects by studying the method of mixing these pre-expert models: the model fused the features extracted by the pre-expert models through the gating unit, and the Mixture-of-Experts(MoE) layers were used to extract and transform the semantic features, and finally the prediction results were obtained through the sparse decoding structure. In order to prevent overfitting caused by a small amount of data, this paper improved the convergence performance of the model and reduced the computational cost by reusing parts of the pre-expert models to improve the generalization of feature extraction. This paper provided more possibilities for the defect classification tasks of the MoE model based on Convolutional Neural Network(CNN).

Author Keywords

Mixture-of-Experts Model; Convolutional Neural Network; Defect Classification Algorithms; Self-attention feature encoding layer.

1. Introduction

In the field of industrial display manufacturing, a variety of manufacturing defects occur due to factors such as equipment, environment, and labor, which often need to be detected and screened to ensure the quality of the final product. At the same time, due to the differences in product types, production environments and photo environments etc., there are a large number of display defect identification needs for various image styles, defect types and background types, that is, a large number of proprietary models for detailed application scenarios have been generated, and we call these deep learning models that can achieve good performance in specific scenarios as pre-expert models. For other scenarios, the pre-expert models often cannot play a stable and good role, so the algorithm designing and model training need to be redo, which leads to an increase in computing resources and development costs.

In this paper, a Mixture-of-Experts(MoE) [1][2][3][4] classification model based on Convolutional Neural Network(CNN) was proposed, which fused the features of multiple pre-expert models to increase the reusability of pre-expert models and reduce the cost of model development for new scenarios. This paper iteratively updated the first convolutional layer of the pre-expert models through transfer learning [5][6], which improved the adaptability of domain migration while retaining the features of the source domain. In the feature encoding unit, this paper adopted a MoE layer to reduce compute costs and increase the model size [1][2][3], and in the feature decoding unit, a sparse decoding structure similar to the MoE layer was adopted to increase the performance of feature regression [7].

2. Method

The framework of the CNN-based MoE classification model proposed in this paper was shown in Figure 1, which is mainly composed of multiple pre-expert models, a gating unit for feature fusion and separation, a feature encoding unit based on MoE layer, and a sparse feature decoding unit. The model first extracted features from the input images through multiple pre-expert models, then fused and separated the extracted features by feature representation structure. The model performed feature encoding semantic transformation [8][9][10] through MoE layers on the separated enhanced features, and finally obtains the classification results of the images through sparse decoding units. The main parts of the model were described below:

Gating Unit for Feature Fusion and Separation: In this paper, a gating unit was used to fuse and separate the features extracted from multiple pre-expert models, which is composed of feature flattening [10], feature representation structure based on basis vectors [11], feature gating fusion structure and feature separation structure. The multi-scale output feature image extracted from multiple pre-expert models is flattened according to the length and width dimensions, and the position embedding value [10] is superimposed in the feature flattening stage, that is, the set of multi-scale feature maps is converted into the feature vector. The feature representation structure based on basis vectors refers to the VAE model [12], which represents the feature image as follows:

$$z = \mu + \varepsilon \times e^{\sigma^2} \tag{1}$$

where e^{σ^2} represents the standard Gaussian distribution, μ represents the mean of the feature space, and ε represents the variance of the feature space, that is, the feature z is represented as the form of the weight coefficient in the feature space [12]. The feature representation structure is composed of a mean regression linear layer and a variance regression linear layer, and the feature vector is represented as the set of feature means and variances. The feature gating fusion structure carry out weighted fusion of represented features extracted from multiple pre-expert models, which aims to get the fused basis representation of the features, that is, the mean and variance. The method of fusion is to assign the weight to each pre-expert branch, and these weights are activated by the tanh activation function [13]. The calculation formula is as follows:

$$f = \sum_{k=1}^{n} \tanh(w_k) \times z_k \tag{2}$$

where f represents the fused basis representation of the features, n represents the number of pre-expert branch, z_k represents the basis representation of the k pre-expert model, and w_k represents the fusion weight of the k pre-expert branch. The feature separation structure differentiates features on the basis representation of fused features by enhancing the dominance of each pre-expert branch, which is mainly to prepare for the subsequent differentiated feature encoding. The feature separation structure superimposes the fusion features representation and the basis representation results of corresponding pre-expert branch, and the calculation formula is as follows:

$$F_k = f + z_k, k = 1, \dots, n \tag{3}$$

where F_k represents the separated enhanced feature for the k pre-expert branch.

ISSN 0097-996X/25/5602-1032-$1.00 © 2025

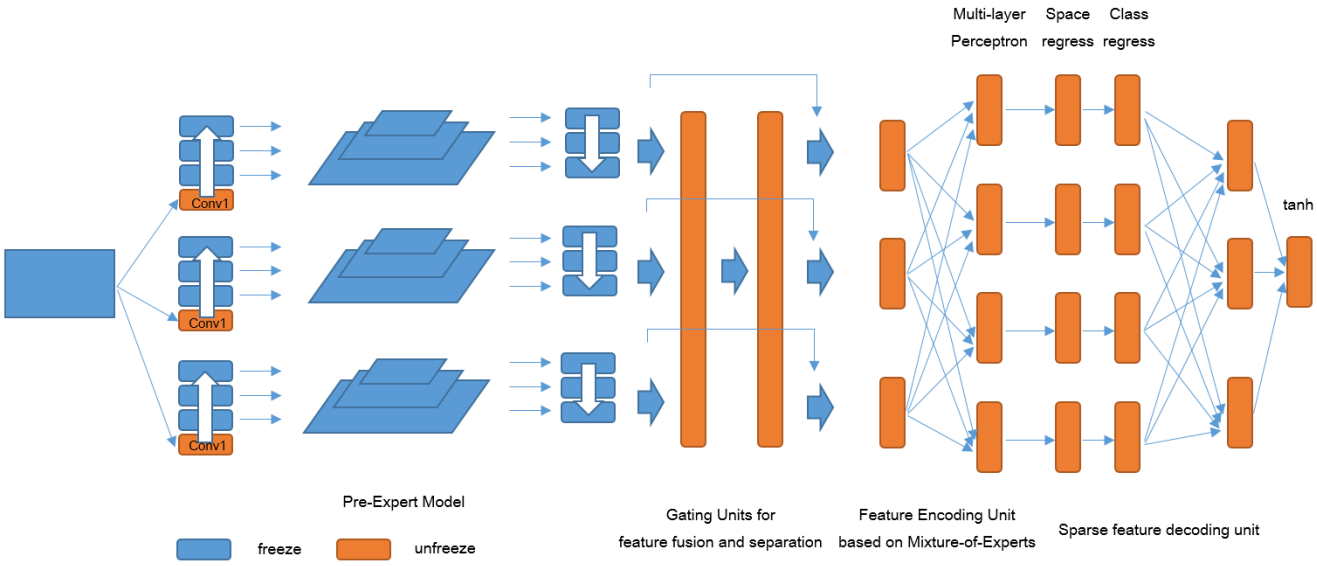


Figure 1. Framework of the Mixture-of-Experts Model based on Convolutional Neural Networks.

Feature Encoding Unit based on Mixture-of-Experts: In the feature encoding unit, this paper used MoE layer for feature transformation. Firstly, different gating weights were calculated by the noisy top-k gating function from MoE layer [4], which are the coefficients of the separated enhanced features input to MoE perception layers. The formula is as follows:

$$G_{kj} = \begin{cases} softmax\left(topk\left(F_k \times w_{kj} + R_{nosie}\right)\right), \\ k = 1, \dots, n, j = 1, \dots, L \end{cases} \quad (4)$$

where G_{kj} represents the gating weight from the k encoder branch to the j MoE perceptron branch, and w_{kj} and R_{nosie} represent the weight and noise which are network parameters commonly used by the MoE layer to help the model be trained [1]. Then, all separated enhanced features were selectively input based on gating weights to different MoE perceptron branches, each of which shares network parameters [1]. Finally, the encoder features from each MoE perceptron branch were derived from the gated set of feature transformations with separated enhanced features.

Sparse Feature Decoding Unit: In the feature regression stage of the model, this paper proposed a sparse feature decoding unit based on the gating mechanism. The sparse feature decoding unit is composed of multiple feature decoders and a gating structure, each feature decoder is composed of the spatial regression layer and the categorical regression layer. The spatial regression layer decoded the spatial dimension of the encoded features to 1, and the categorical regression layer decoded the features after spatial regression layer into category scores. Then, the output of multiple feature decoders was gated fusion by different gating weights from MoE layer. Finally, the tanh-activated gating structure fused different category scores from multiple gated fusion features into classification output. The formula is as follows:

$$\begin{cases} pred = \sum_{k=1}^{n} \tanh(w_k) \times z_{c\_k} \\ z_{decoder\_j} = class_j\left(space_j\left(z_{moe\_j}\right)\right), j = 1, \dots, L \end{cases} \quad (5)$$

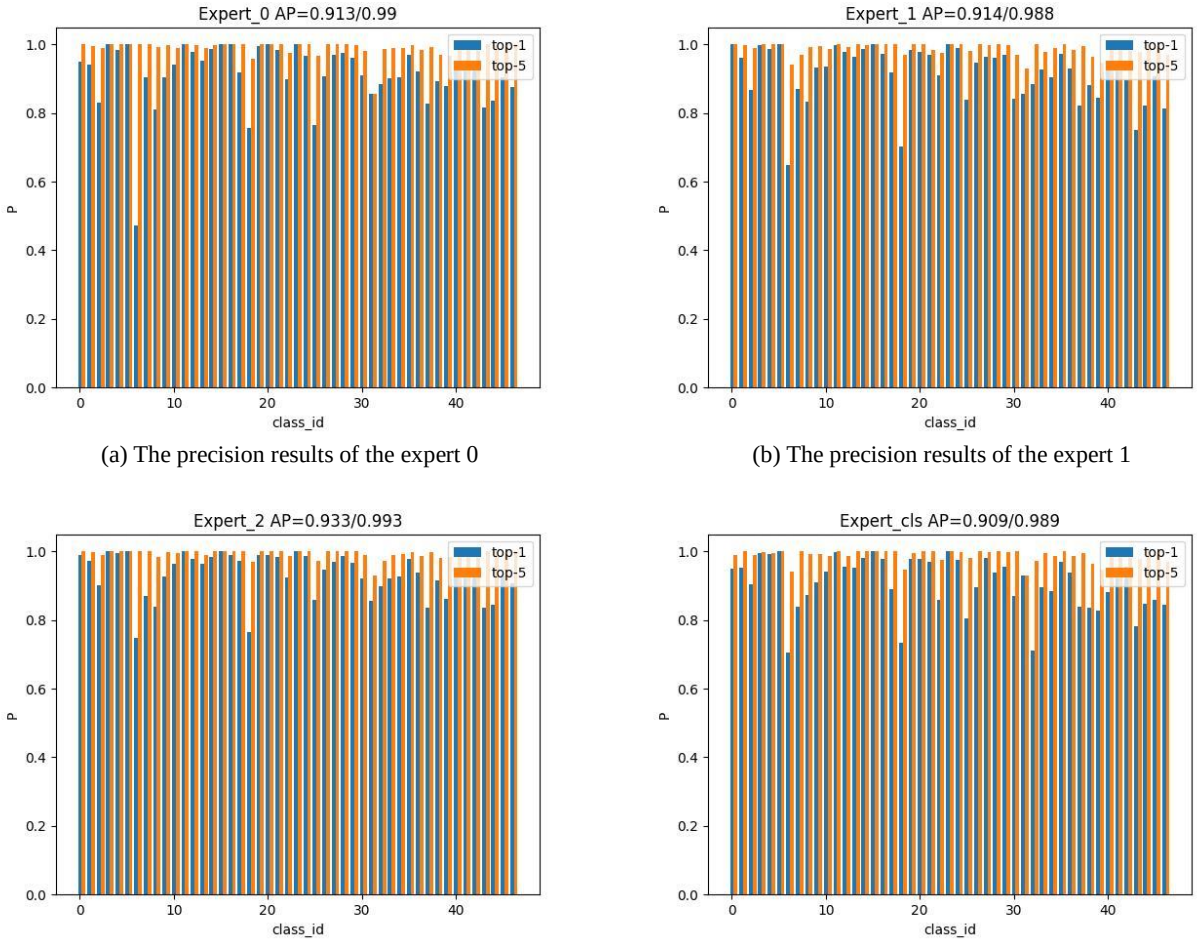
where pred represents the output category scores, w_k represents the output weights, z_{c_k} represents the output of gated fusion by G_{kj} and $z_{decoder\_j}$. $z_{decoder\_j}$ was shown in the second equation which represents the decoder features of the j decoding branch, $space_j$ and $class_j$ represent the spatial and categorical regression layers of each decoding branch, and z_{moe_j} represents the encoder features of the j MoE layer.

The Mixture-of-Experts Model based on Convolutional Neural Networks: To sum up, the specific process of the MoE classification model based on CNN proposed in this paper is as follows: The input images passed through the feature extraction structures of multiple pre-expert models respectively, and the output multi-scale feature map of each pre-expert model was obtained. The separated enhanced feature $\{F_k, k = 1, \dots, n\}$ was obtained after passing through the gating unit for feature fusion and separation, which was represented by the basis of the feature space. Then, in this paper, the multiple encoding features were obtained by semantic transformation of the separated enhanced features by the MoE layer, and the image classification task of information fusion was finally completed by using the sparse decoding unit in the feature regression stage.

3. Experiment

3.1 Datasets

The dataset used for the experiment in this paper came from the industrial display manufacturing, and was taken from the production lines of two different processes: A total of 11,852 defect samples from 3 process links of production line A contained 42 categories; A total of 1411 defect samples from 1 process link of production line B contained 5 categories. In the experiment, these defect samples were cut and scaled to 224*224, and 90% of each category of defect was randomly selected as the training set while the rest 10% as the testing set, and the classification precision of top-1 and top-5 was used as the evaluation metrics.



(a) The precision results of the expert 0

(b) The precision results of the expert 1

(c) The precision results of the expert 2

(d) The precision results of YOLOv8

Figure 2. Classification precision results for each category in ablation experiment on the impact of pre-expert model.

3.2 Pre-expert model

The pre-expert model used in this paper came from the defect detection and segmentation tasks in the field of industrial display manufacturing. This paper used 6 pre-expert models came from 6 different defect detection and segmentation tasks, including 3 YOLOv8 defect segmentation models related to production line A and 3 YOLOv8 defect detection models unrelated to this experiment. All of pre-expert models have got more than 80% average precision in the corresponding pre-task.

3.3 Ablation experiments.

In the ablation experiments, parts of the network parameters were frozen in the training stage while the rest of the parameters were trained. The unfrozen parts of the network parameters were the parameters of first convolutional layer in each pre-expert model. The model mentioned in each of the following experiments was trained over 300 epochs, which loss decreased and converged.

3.3.1 The impact of the pre-expert model

Firstly, the selection of the pre-expert model of gated fusion was compared and analyzed: three models related to production line A were used as the pre-expert models for the expert 0, three models unrelated to this experiment were used as the pre-expert models for the expert 1, and all 6 models mentioned above were used as the pre-expert models for expert 2. The top-1 and top-5 precision results of each class of the final trained MoE model on the experimental test data were shown in the Figure 2.

The experimental results showed that the average performance of the mixed multiple pre-expert models proposed in this paper is better than that of the traditional classification model. The average precision on the mixed dataset by the classification models of the expert 0 and the expert 1 were at the medium level and had similar performance, but the top-1 classification precision of some categories which were related to the pre-expert model is poor. It showed that the proposed model can have good performance under both correlated and uncorrelated pre-expert model settings, but the limited pre-expert model setting may lead to overfitting in the categories with few samples in the case of data imbalance. The average precision on the mixed dataset by the classification models of the expert 2 was the highest, and there

was no overfitting as that of the classification models of expert 0, that is, the more comprehensive the pre-expert model, the more robust the performance of the proposed MoE model. The results were consistent with the performance of the transfer learning method, that is, when the amount of data in the original domain is not large enough, the transfer learning is affected by the differences between the original domain and the target domain [6]. When the pre-expert model is limited and the defect samples are unbalanced, the MoE model may be overfitting in the target domain, and when the pre-expert model increases, the MoE model solves this problem by the knowledge retained from a broader source domain. To sum up, the results in ablation experiment on the impact of pre-expert model showed that the MoE model on the basis of mixing multiple different pre-expert models proposed in this paper can show robust performance on the mixed dataset of multiple scenarios.

3.3.2 The impact of sparse decoding units

Secondly, the number of sparse decoding branch was compared, and the experimental results were shown in Figure 3.

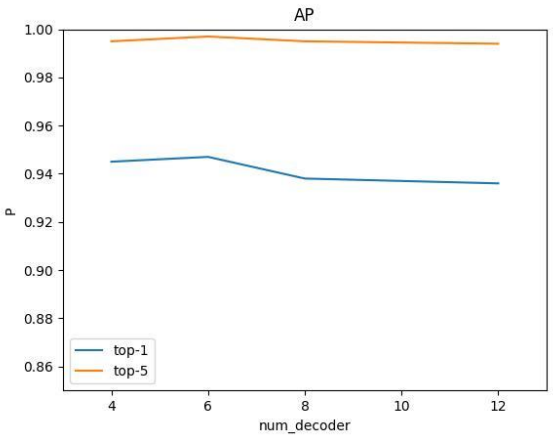


Figure 3. Classification average precision results on the impact of sparse decoding units.

The experimental results show that for the mixed dataset used in this experiment, the average classification precision was stable in the early stage when the number of decoders was less than 6, but decreases slightly when the number of decoders was greater than 6, where 6 is equal to the number of encoders.

4. Conclusion

The MoE classification model based on the pre-expert model proposed in this paper showed superior performance on the mixed dataset with multiple scenarios, and its performance in the field of industrial display manufacturing was better than that of traditional transfer learning based on real-world pre-trained model [14]. Meanwhile, there are hyper parameters such as the number of decoding branches and optimizer type which need to be artificially set for different scenarios, and which will not only affect the size of the model but also affect the performance of the model.

5. Impact of Research

This paper proposed an exploratory defect classification algorithm based on CNN and MoE model. The gating unit for feature fusion and separation was proposed, which fused and separated the extracted features from multiple pre-expert models. The feature encoding unit based on the MoE layer was proposed to complete the semantic transformation of features, and also the sparse decoding unit based on gating mechanism was proposed to complete the image classification task. Although this paper only focused on the image classification task, this paper explored a feature extraction method that multiplexes the pre-task model, a feature fusion and separation method with basis representation, and a sparse feature encoding and decoding method by MoE layers. This paper provided more possibilities for the future tasks such as object detection and segmentation of the MoE model based on CNN.

6. References

1. Cai, Weilin, et al. "A Survey on Mixture of Experts." (2024).
2. Shazeer, Noam, et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." (2017).
3. Xue, Fuzhao, et al. "Go Wider Instead of Deeper." (2021)
4. Hazimeh, Hussein, et al. "DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning." (2021).
5. Zhuang, F. Z., et al. "Survey on transfer learning research." Journal of Software 26(2015):26-39.
6. Redko, Ievgen, et al. "A survey on domain adaptation theory: learning bounds and theoretical guarantees." (2020).
7. Jane, Y. Nancy, et al. "2-HDCNN: A two-tier hybrid dual convolution neural network feature fusion approach for diagnosing malignant melanoma." Computers in biology and medicine 152(2023):106333.
8. Carion, Nicolas, et al. "End-to-End Object Detection with Transformers." (2020).
9. He, Ju, et al. "TransFG: A Transformer Architecture for Fine-grained Recognition." (2021).
10. Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations 2021.
11. Zhang, Ming Guang, W. H. Li, and M. Q. Liu. "Adaptive PID Control Strategy Based on RBF Neural Network Identification." International Conference on Neural Networks & Brain IEEE, 2005.
12. Zhai, Ziming. "Variational Auto-Encoder Reconstruction Networks for Classification of Hyperspectral and LiDAR Data." Journal of Physics: Conference Series 2562.1(2023).
13. Visin, Francesco, et al. "ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks." Computer Science 25.7(2015):2983-2996.
14. Mattins, R. Faerie, et al. "Object detection and classification of butterflies using efficient CNN and pre-trained deep convolutional neural networks." Multimedia Tools and Applications 83.16(2024):48457-48482.