Convolutional Neural Network Interpretability Analysis for Image Classification

Haopeng Fang[1, *]

[1]School of Mathematics and Physics, Lanzhou Jiaotong University, Gansu 730000, China

[*] Corresponding Author

Abstract

In order to understand the basis for decision-making by convolutional neural networks in image classification tasks, and then optimize the model and reduce the cost of parameter adjustment, it is necessary to conduct interpretability analysis of convolutional neural networks. To this end, the article takes the fruit image classification task as the starting point, uses a variety of category activation maps, and analyzes the reasons for the results given by the model from multiple angles. The article uses the ResNet model for fine-tuning first. After achieving good classification performance, it conducts basic analysis of semantic features, occlusion analysis, as well as CAM-based interpretability analysis and LIME interpretability analysis to provide convolutional neural networks. Certain interpretability. Experimental results show that the basis for decision-making by convolutional neural networks is consistent with the semantics understood by humans.

Keywords

image classification, convolutional neural networks, interpretability analysis, activation map

1. Introduction

Interpretability analysis is a method of describing the internal structure of a model in a way that humans can understand. It has a lot to do with human cognition and biases. Its advantage is that it can help humans understand the working principle of the model, break the "black box" of the neural network, provide human-

understandable explanations for the decisions made by the algorithm, ensure that humans can trust the decisions made by the neural network, and at the same time endow the model with Strong interpretability is conducive to ensuring its robustness, analyzing the reasons for model judgment errors, and then improving the model. At the same time, the last feature layer of the convolutional neural network contains rich semantic and spatial location information. This information is difficult for humans to understand and is difficult to display in a visual way. This article extracts the information of this layer and uses Heatmaps study which areas in an image enable a convolutional neural network to make decisions, providing insights for deep learning in classification tasks.

2. Research methods and ideas

The ResNet is a convolutional neural network model proposed by He Kaiming and others. Its main idea is to add a residual structure module to the network, which solves the problem of difficult training of deep neural networks, making it possible for the depth of the network to reach 1,000 layers, and with time Accuracy can also be improved by increasing the number of coating layers. The simple and efficient ResNet model has become the most popular convolutional neural network structure in the industry. This experiment uses this model as the experimental model. Currently, there are three main ways to achieve interpretability of deep neural networks:

1. Based on data interpretability, through data analysis and visualization technology realizes the visualization of the model and intuitively displays the key basis of the model results.

2. Based on the interpretability of the model, an interpretable model is constructed. The model not only outputs the results, but also outputs the reasons for the results.

3. Based on the interpretability of the results, the trained model is considered as a "black box". Based on a given batch of inputs and corresponding outputs, combined with the observed behavior of the model, the reasons for the corresponding results are inferred. This paper The experiment adopts this method.

The research ideas of this article:

Step 1: Construct a picture classification data set, and use crawler technology to obtain a total of 6,310 fruit pictures in 30 categories (about 200 pictures in each category). The pictures are screened as much as possible to include various scenes of target objects (such as different backgrounds, Pictures of different shapes and different parts).

Step 2: Use transfer learning to train the model. The ResNet152 model was used in the experiment. Only the last layer of the training model was fine-tuned so that the output of the fully connected layer corresponds to the number of categories in the current data set. After hyperparameter adjustment and iteration of the data set 30 times, , the model converged, and the accuracy obtained on the test set was 87.58%.

Step 3: Calculate the semantic features of the test set images, and extract the output of the intermediate layer of the trained image classification model as the semantic features of the input image. Calculate the semantic features of all images in the test set, use UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction method to reduce it to two dimensions, and visualize it. Analyze the semantic distance of different categories.

Step 4: Perform CAM-based analysis on the trained models respectively

(Class Activation Mapping) interpretability analysis, high-resolution fine-grained interpretability analysis based on Guided Grad-CAM and LIME (Local Interpretable Model Explanations) interpretability analysis.

3. Experimental results and analysis

3.1. Semantic feature analysis

The visualization result after UMAP dimensionality reduction is shown in Figure 1.

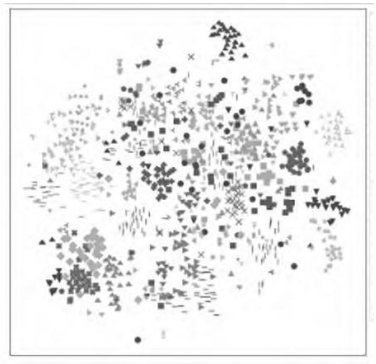

Figure 1: Visualization results diagram

The experiment found that objects of the same category are mostly clustered together, indicating that the model has fully extracted the semantic features of the image. It was also found that the classification boundaries of some categories are very blurry (such as cucumber and bitter melon) , indicating that the semantics between them are very similar and cannot be easily distinguished by the model, which is also the reason why the model makes mistakes.

3.2. Model training settings

Use the small slider to slide to block different areas on the input image, and observe which areas will significantly affect the model's classification decision after being blocked. Then change the slider size, sliding step size, and compare the effect.

The experiment first used a medium occlusion slider. When the slider blocked part of the object in the picture, the probability of the model predicting the correct category dropped significantly. Then use a small occlusion slider to slide the object in the picture, and the probability of the model predicting the correct category is not significantly affected. Use two sliders of the above size at the same time, fix one of them, slide the other in the picture, and then fix it The slider moves once, and the other slider still slides in the picture. It is found that the probability of the model predicting the correct category decreases slightly. The experiment shows that the convolutional neural network does based on the features that humans are more concerned about when making decisions. It relies on By understanding certain local features, correct decisions can be made even without seeing global features.

3.3. Category Activation Heatmap Interpretability Analysis

Input an original image into the model, and the predicted category of the model is banana, as shown in Figure 2. The Grad-CAM (Gradient-weighted Class Activation Mapping) algorithm is used to generate an interpretability analysis heat map and map it to the original image. On the figure, Figure 3 is obtained. The highlighted area in the picture is the area of interest when the model predicts a banana. Figure 4 shows the area of interest when the model predicts a kiwi. The highlighted area in the picture corresponds to the label area. It can be seen that for different categories, the model can find their own distinctive or informative areas, which shows that the model has category discriminability, that is, for any picture or any category, a heat map is drawn, and the highlighted part It shows the area that the neural network focuses on a specific category on the original image, that is, the attention of the neural network. This algorithm helps to further evaluate the model, discover the shortcomings of the model, and then improve the model structure.

Figure 2: Original picture

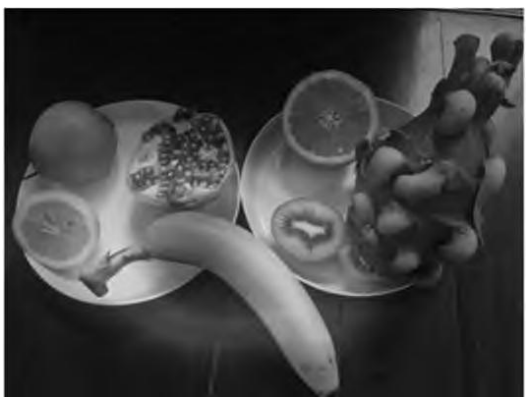

Figure 3: Picture processed by Grad-CAM

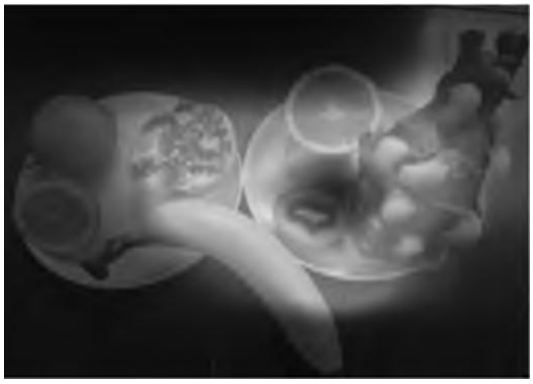

Figure 4: Picture processed by Grad-CAM

3.4. Gradient-based interpretation methods

Perform Guided Grad-CAM interpretability analysis on Figure 2, and draw a fine-grained heat map that is both category discriminative and high-resolution, as shown in Figure 5. The principle of the Guided Backpropagation algorithm is to find the impact that each pixel of the image will have on the network output of a specific category. The prediction result of a certain category is derived from the partial derivative of each pixel of the original image to obtain Figure 6. Multiply the banana's Grad-CAM heat map and the GuidedBackpropagation heat map element by element to obtain the corresponding pixel heat map. This heat map has both high resolution and category discriminability. The gradient calculation of the model can highlight the key features of bananas. It is more efficient and has better experimental results.

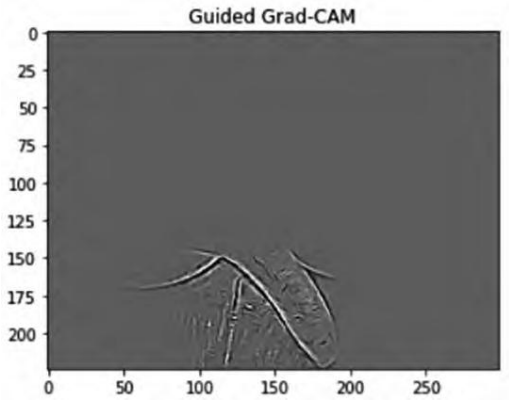

Figure 5: Picture processed by Guided Grad-CAM

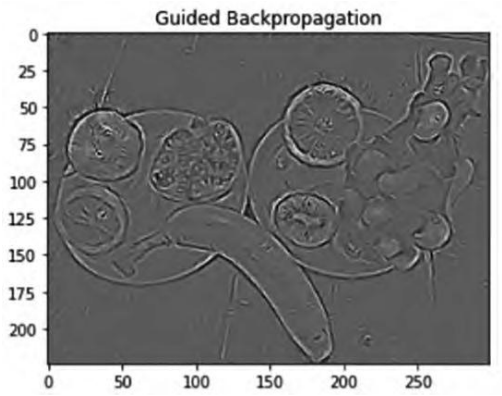

Figure 6: Image processed by Guided Backpropagation

3.5. Gradient-based interpretation methods

The principle of the LIME algorithm is to convert the original image into an interpretable feature representation, perturb the samples, obtain N perturbed samples, restore these N samples to the original feature space, and use the predicted values as the real values, using the Interpret the feature data to create a simple data representation and observe which superpixels have larger coefficients.Use the LIME algorithm to perform interpretability analysis on Figure 2, and the visualization results are shown in Figure 7. This algorithm helps analyze the reasons why the model makes decisions and the reasons why the model makes mistakes, thereby optimizing the model structure.

Figure 6: Image processed by LIME

4. Conclusion

Aiming at the problem of fruit classification, this paper analyzes the semantic features extracted by the model, conducts dimensionality reduction visualization and occlusion interpretability experiments, and finally introduces class activation maps for

analysis. Experiments show that the convolutional neural network makes decisions based on more consistent semantic concepts understood by humans, and it can make correct decisions based on part of an object. CAM visualization can help understand the principles of the model and analyze the reasons for model prediction errors. However, CAM visualization can only explain the area of concern of the model during classification, and cannot explain why the convolutional network can locate the relevant area during training, that is, the network is not known. How this knowledge is learned requires further research.

References

[1] Alex K, Ilya S, Geoff H. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, 2012: 1106–1114.

[2] Karen S, Andrew Z. Very Deep Convolutional Networks for Large–Scale Image Recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, United States, 2015: 1–14.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. CVPR, 2016: 770–778.

[4] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. IEEE Computer.