

Research on the Application of Deep Learning in Food Detection

Zheng Sen*, Orgil Demberel

School of Applied Sciences, Mongolian University of Life Sciences, Ulaanbaatar, Mongolia

Abstract: With the continuous development of deep learning-related technologies, image recognition has gradually shifted from traditional manual feature methods to data-driven representation learning models. In the context of dish detection, due to the large number of categories, small visual differences between categories, and complex background environments, the image recognition process poses certain challenges. This paper combines typical models such as convolutional neural networks, deep belief networks, and recurrent neural networks to analyze their feature learning mechanisms in fine-grained image recognition. The research shows that deep learning can enhance recognition capabilities in complex visual scenes, but model performance is still affected by factors such as data quality and training strategies, and there is still room for further optimization of related methods.

Keywords: Deep Learning; Image Recognition; Dish Detection; Convolutional Neural Network; Representation Learning

With the widespread adoption of information collection devices, the scale of image data continues to grow. Extracting effective information from unstructured visual data has become one of the important research directions in the current artificial intelligence field. Image recognition, as a core task in this field, has wide application value in scenarios such as industrial inspection, intelligent security, and smart healthcare[1].

In the context of catering informatization, dish detection has gradually become a research direction that has attracted attention and is being applied in systems such as automatic settlement and intelligent recommendation. However, due to the wide variety of dish categories, the small visual differences between categories, and the complexity of background environments and lighting conditions, the recognition task still poses certain challenges.

Traditional image recognition methods primarily rely on manually constructed feature representations, such as SIFT and HOG, whose performance heavily relies on domain experience and feature engineering. In complex visual tasks, the generalization ability and scalability of these methods are limited[2]. Against this backdrop, end-to-end modeling methods based on deep learning have gradually become an important research direction in the field of image recognition.

1. Theoretical Basis Related to Deep Learning

1.1 Expressing the Basic Idea of Learning

Deep learning is essentially a method for learning data representation through multiple layers of nonlinear transformations. Its core does not lie in the manual design of features, but rather in relying on the network structure to automatically learn high-dimensional representations with discriminative ability during the training process. In image recognition tasks, this process is typically manifested as a progressive evolution from pixel-level information to semantic-level expression[3]. Compared to traditional machine learning methods, such models significantly reduce their dependence on manual feature design, thus exhibiting stronger adaptability and expressive power in complex data distribution environments.

1.2 Deep Belief Network(DBN)

Deep belief networks are composed of multiple stacked restricted Boltzmann machines, and their training process typically employs a layer-wise unsupervised pre-training strategy to improve the initialization quality of deep networks. This method was of great significance in the early stages of deep learning development, but it has gradually been replaced by convolutional structures in current large-scale visual tasks[3].

1.3 Convolutional Neural Network(CNN)

Convolutional Neural Networks (CNNs) are arguably one of the most widely used and

representative models in current image recognition tasks. Their main characteristic lies in the introduction of local receptive fields and weight sharing mechanisms. Such a structure can better utilize the spatial correlation of images, reduce the number of parameters to a certain extent, and improve overall computational efficiency[2][4]. In the task of dish detection, CNNs continuously extract image features layer by layer, enabling information to gradually transition from relatively basic texture features to higher-level semantic representations. This allows the model to more accurately capture key visual information even under complex background conditions.

1.4 Recurrent Neural Network(RNN)

Recurrent neural networks (RNNs) introduce a temporal recurrent structure, enabling the model to possess sequence modeling capabilities. In visual tasks, they are commonly used to handle structured output problems, such as image sequence description or character-level recognition tasks. When combined with convolutional networks, a more expressive hybrid structural model can be formed[3].

2. Key Modeling Mechanisms in Image Recognition

2.1 Hierarchical Semantic Representation Mechanism

One of the significant advantages of deep models lies in their hierarchical feature expression capability. Typically, shallow networks primarily extract basic visual information such as edges and colors, while as the number of network layers increases, the middle and high layers further learn texture structures and semantic features. This expression method, which gradually transitions from local details to overall semantics, endows the model with good stability and robustness in complex visual environments[1]. In the task of dish detection, this hierarchical mechanism can assist the model in more effectively distinguishing dishes with similar appearances but different categories, thereby enhancing the overall recognition performance.

2.2 Parameter Optimization and Training Strategy

The training process of a deep model can be viewed as a non-convex optimization problem, with the objective of continuously updating

network parameters through backpropagation to gradually reduce the loss function. In practical training, to enhance convergence stability, strategies such as momentum methods, adaptive learning rate adjustments, and regularization are often employed to minimize fluctuations during the optimization process and improve the overall training performance[1][3]. Simultaneously, the Dropout mechanism is frequently utilized in network training. By randomly masking certain neurons during the training phase, it diminishes the model's reliance on local features, thereby mitigating overfitting to some extent and endowing the model with enhanced stability and generalization capabilities on unseen data.

2.3 Data-Driven Modeling Feature

Deep learning models essentially rely on data-driven mechanisms, and their performance is closely related to the scale, quality, and distribution structure of training data. In the task of dish detection, due to the imbalance of categories and significant sample differences, data augmentation and resampling strategies become important means to improve model performance[3].

3 Analysis of Dish Inspection Tasks

3.1 Task Characteristics and Difficulties

Dish detection is often regarded as a fine-grained visual recognition task, where the difficulty lies not only in the large number of categories, but also crucially in the uneven distribution of discriminative information among different categories. In practical scenarios, different dishes often exhibit high similarity in color composition and overall shape, making effective distinguishing information more concentrated in local detail areas. Simultaneously, dishes within the same category can be influenced by variations in cooking methods, plating habits, shooting angles, and lighting conditions, leading to significant differences in appearance within the category, thus resulting in a certain degree of discreteness in feature distribution.

Furthermore, in real-world application scenarios, background factors such as tableware, desktop texture, and lighting variations are common. These non-target information may be mistakenly used as discriminative evidence by the model during feature extraction, thereby affecting the overall generalization ability. Under the combined influence of various factors, the

phenomenon of blurred or even overlapping class boundaries in the feature space, known as the "feature aliasing" problem, is prone to occur. Therefore, how to stably extract fine-grained features with discriminative ability under complex background conditions has become one of the key issues in this task.

3.2 Analysis of Feature Learning Process

In the framework of convolutional neural networks, image features are typically formed gradually through multiple layers of nonlinear transformations. Shallow networks primarily extract basic visual information such as edges and colors, while intermediate layers further learn texture structures and local shape features. Higher-level networks gradually integrate these into abstract representations with semantic meaning[2][4].

In the task of dish detection, this hierarchical feature learning process is crucial. Since the differences between different dishes often concentrate in local detail regions, the model needs to retain key detail information as much as possible when abstracting semantic features; otherwise, excessive feature compression may affect recognition performance. From the perspective of implementation mechanism, this issue is closely related to the expansion of convolutional receptive fields and the information compression during pooling: a larger receptive field helps to obtain global information but may weaken local fine-grained features, while a smaller receptive field is not conducive to forming stable semantic representations. Therefore, in the design of network structure, reasonably coordinating the expression and fusion of features at different levels is of great significance for improving dish detection performance.

3.3 Network Structure Adaptability Analysis

Given the characteristics of dish detection tasks, model structure design often requires a trade-off between expressive power and generalization ability. Increasing network depth appropriately can enhance the model's ability to express complex features, but it may also introduce issues such as unstable gradient propagation and overfitting. To address this, the residual connection mechanism is widely used, which enables shallow information to participate in deep representation through cross-layer feature propagation, thereby alleviating optimization

difficulties during deep network training[5].

Furthermore, multi-scale feature fusion holds significant importance in dish detection tasks. Due to inconsistent target scale distribution, single-scale features often struggle to balance local details and overall structural information simultaneously. By fusing features from different levels, the model's adaptability to complex scenarios can be enhanced to some extent[6]. Simultaneously, to minimize background interference, attention mechanisms or feature re-weighting methods are typically introduced to strengthen the model's focus on key areas. Overall, the core of model structure design lies in optimizing feature representation based on task characteristics, thereby enhancing discriminative ability and generalization performance.

4 Conclusion

This article analyzes the application of deep learning in the field of image recognition, focusing on the representative fine-grained visual task of dish detection as the research object. It organizes and discusses the currently mainstream deep learning models and their feature learning mechanisms. Through comparative analysis of different model structures and their modeling methods, it is evident that deep learning methods can gradually learn feature representations with strong discriminative ability in complex visual scenes, thereby improving the overall recognition performance to a certain extent. However, in practical applications, its performance is often not completely stable and is still largely influenced by the characteristics of data distribution and the setting of training strategies. Especially in cases where sample diversity is insufficient or class distribution is imbalanced, model performance may fluctuate or even decline[1][2].

References

- [1] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *nature*, 2015, 521(7553): 436-444.
- [2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. *Advances in neural information processing systems*, 2012, 25.
- [3] Goodfellow I, Bengio Y, Courville A, et al. *Deep learning*[M]. Cambridge: MIT press, 2016.
- [4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image

- recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [6] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.