Journal of
Data science and Modern Techniques

Research Article Open Access

Indoor Scene Recognition Method Combining Attention and Feature Suppression

Pengyu Hao[1*], Shengjun Xue[2], Xianyi Cheng[3] and Zexuan Ding[4]

[1]Silicon Lake College, Jiangsu 215332, China
[2]Nanjing University of Information Science and Technology, Jiangsu 210000, China
[3]Nantong Institute of Technology, Jiangsu 226001, China
[4]Singapore Institute of Management, Jin Wentai Road 599491, Singapore

[*]**Corresponding Author:** Pengyu Hao, Silicon Lake College, Jiangsu 215332, China, E-mail: pyhao1223@163.com

Received Date: August 01, 2024 **Accepted Date:** September 01, 2024 **Published Date:** September 04, 2024

Citation: Pengyu Hao, Shengjun Xue, Xianyi Cheng, Zexuan Ding (2024) Indoor Scene Recognition Method Combining Attention and Feature Suppression. J Data Sci Mod Tech 3: 1-12

Abstract

Due to the inconsistency in the dimensions and properties of the target features and scene features extracted by the ObjectNet, as well as the presence of redundant information that affects scene judgment, resulting in low recognition accuracy, a new indoor scene recognition method is proposed. Firstly, the attention mechanism is introduced into ObjectNet. Then, the target features output by ObjectNet are transformed to obtain features with the same dimensions as the original scene features. Finally, Context gating (CG) is used to suppress redundant information in the features. Experiments were conducted on two datasets, MIT Indoor67 and SUN 397. Compared with MR-CNNs networks, the recognition accuracy of ObjectNet increased by 3% and 2.11% respectively. The results show that by using attention mechanism and CG to suppress redundant feature information, the accuracy of indoor scene recognition is improved.

Keywords: ObjectNet; Deep Learning; Indoor Scene; Class Conversion Matrix; Feature Fusion

Introduction

In recent years, with the popularization of the Internet and the continuous improvement of people's living standards, the development of scene recognition technology has brought more and more services and convenience to people. Indoor scene recognition is a key part of scene recognition, and the development of indoor scene recognition technology has broad application prospects in smart homes, service robots, security monitoring and other fields [1-2].

Although many methods have achieved remarkable results, there exist issues such as inconsistencies in the dimensions and properties of extracted target features and scene features, as well as redundant information that affects scene recognition, leading to limited accuracy in indoor scene recognition. This paper proposes a new indoor scene recognition method by introducing an attention mechanism into the object detection network. The target features outputted by ObjectNet are transformed to match the same dimension as the original scene features. Subsequently, a Context Gating (CG) mechanism is employed to suppress redundant information in the features, thereby enhancing the role of target features in scene recognition.

Related Research

Early indoor scene recognition typically relied on features such as color, texture, and shape for identification. With the widespread application of operators like SIFT, SURF, and HOG, a popular classification method involved using various operators to extract environmental features and then training a benchmark model for scene discrimination. The widely used models were Bag of Words (BoW) and its improved versions. For instance, Lazebnik et al. [3] proposed a spatial pyramid architecture based on the BoW model for scene recognition. Espinace et al. [4] suggested using image segmentation to infer scene categories based on typical objects, but object segmentation in complex scene environments is itself a challenging task in machine vision. Additionally, some scholars have combined different models and features to achieve scene discrimination. Zhao et al. [5] combined color and local texture features, utilizing monocular vision and natural landmarks to effectively accomplish robot positioning tasks with good recognition results. However, these algorithms still heavily rely on manual operators for feature extraction, lacking generalization ability, which is significantly improved by deep learning methods in this aspect.

Among deep learning-based methods, there are currently many approaches that combine target features for indoor scene recognition. Inspired by ImageNet [11], Zhou et al. [6] proposed a new dataset called Places and selected 205 scene categories from it to train a dedicated scene recognition network called Places-CNN. Its recognition accuracy far surpasses traditional manually designed feature methods, providing researchers with new methodological guidance. Due to the complexity of indoor scenes containing multiple targets, global features can be challenging to represent these target features. Based on this, Antonio et al. [7] proposed combining global and local features to identify indoor scenes, utilizing local features to represent target characteristics and improving recognition accuracy. Herranz et al. [8] introduced a multi-scale feature-based method that feeds images of different sizes into their respective target and scene networks for feature extraction, addressing the issue of image size matching with the recognition network. However, increasing image sizes lead to increased algorithmic complexity. Wang et al.[9] presented a knowledge-guided disambiguation strategy that uses target features extracted from a knowledge network to generate soft labels for scene images, guiding the scene network to minimize the loss function. This effectively addresses the issues of small inter-class differences and large intra-class variations, but the utilization rate of target features remains low. To improve target utilization, Seong et al. [10] proposed an end--to-end trainable network called FOSNet. The network employs a Scene Consistency Loss (SCL) algorithm to calculate losses in image patches, effectively improving the utilization of target features in scene images.

Due to the complexity of indoor scenes, uneven illumination, and high repetition of colors and textures, the aforementioned semantic segmentation methods based on RGB color images suffer from issues such as mis-segmentation of object edges and misclassification of categories. This makes it impossible to achieve precise understanding of environmental semantic information by intelligent agents. Recent research has found that compared to methods based on

ordinary RGB color images, RGB-D-based approaches can utilize additional depth information from the scene. This depth information is less affected by illumination and can reflect the positional relationship between objects, complementing the RGB color information.

Couprie et al. [14] found that auxiliary depth information can reduce the segmentation error rate for objects with similar depth, appearance, and positional information. With the emergence and development of depth cameras like Kinect [15], it has become easy to obtain depth information for images. However, finding a way to fuse RGB color information with depth information and exploiting their complementarity has been a challenging problem. Some simple methods stack depth information onto the RGB color channels and train the network assuming RGB-D data with four input channels. But directly fusing depth information as a fourth channel does not fully utilize the encoded scene structure information.

Gupta et al. proposed the HHA (Horizontal disparity, Height above ground, Angle with the inferred gravity direction) depth information representation method, which converts depth images into three different channels (horizontal disparity, height above ground, and the angle of the surface normal). However, HHA only emphasizes the complementary information between each channel's data while ignoring the independence of each channel, and it requires a significant amount of computation.

Hazirbas et al. proposed a new fusion architecture called FuseNet (Fusion Network) that integrates complementary depth information into the semantic segmentation framework, improving segmentation accuracy. However, it does not achieve multi-scale fusion.

Hu et al. proposed ACNet (Attention Complementary Features Network), designing an attention auxiliary module to balance feature distribution and enabling the network to focus more on effective regions of the image. While maintaining the original RGB-D feature branch, it fully utilizes the fused features of RGB information and depth information. However, because scene images contain multiple target information, this method may not positively impact scene recognition for all targets, and some may even have a negative effect on recognition performance. Additionally, when fusing scene features with target features, due to the difference in their dimensions, simple stacking or concatenation can lead to issues such as feature loss.

Multi-Head Attention Machanism

The Attention Mechanism in deep learning is a method that mimics the human visual and cognitive system. It allows neural networks to focus their attention on relevant parts when processing input data. By introducing the Attention Mechanism, neural networks can automatically learn and selectively focus on important information in the input, improving the model's performance and generalization ability.

It's worth noting that the attention mechanism does not inherently know how to determine the importance of information; instead, it requires learning through a significant amount of training data. During the training process, the model encounters numerous inputs and their corresponding outputs. Through continuous learning and optimization, the model gradually learns to identify which information is crucial and which can be disregarded for a given task.

In specific implementations, attention mechanisms are usually combined with encoder decoder architectures. The encoder maps an input sequence represented by symbols (usually vectors) $(x_1,...,x_n)$ to a continuous representation sequence $z = (z_1,...,z_n)$. After receiving z, the decoder generates an output sequence represented by symbols $(y_1,...,y_n)$, where y_i is generated at each time step, with i representing any number from 1 to n. At each step, the model automatically consumes the symbol generated in the previous step. For example, when generating y_2, y_1 is used as additional input.

The left and right halves of Figure 1 show the complete connections of the encoder and decoder, respectively.

(1) Scaled Dot-Product Attention

Scaled Dot-Product Attention takes as input queries and keys of dimension dk, and values of dimension dv. It first computes the dot product of the query with all keys, divides the result by the square root of dk, and then applies a softmax function. The output is the weighted sum of

the values, where the weights are given by the softmax output. The formula is as follows:

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V$$

(2) Multi-Head Attention

Compared to performing a single attention function using keys, values, and queries, it is more advantageous to perform linear projections h times with different keys, values, and queries, learning linear projections in dimensions dk, dk and dv, respectively. For each projected version of queries, keys, and values, we execute the attention function in parallel and obtain a dv-dimensional output value for each. These output values are then concatenated and projected again to obtain the final result, as shown in Figure 1.

Multi-head attention allows the model to attend to information from different representational subspaces at different positions simultaneously [15]. Using only one attention head would inhibit this capability.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^o$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$Wherein, W_i^Q \in \Re^{d\bmod el \times d_Q}, W_i^K \in \Re^{d\bmod el \times d_K}, W_i^V \in \Re^{d\bmod el \times d_V}, W^o \in \Re^{hd_V \times d\bmod el}$$

All occurrences of W here represent projection matrices. In this context, we employ h=8 parallel attention layers, also known as heads. For each layer, we use $d_k=d_v=d_{model}/h = 64$, where dmodel refers to the dimensionality of the word embedding vector, which can be deduced to be 64*8 = 512. Due to the reduced dimensionality per head, the overall computational cost is comparable to using a single attention head with the full dimensionality.

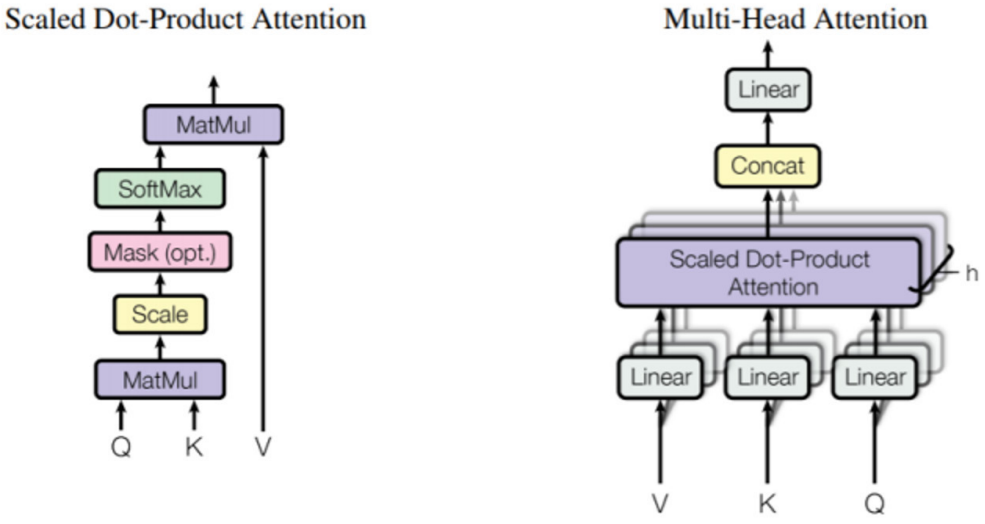

Figure 1: Attention mechanism

The multi-head attention mechanism plays the following key roles in indoor scenes:

(1) Improving model performance: The attention mechanism enables neural networks to focus more on key

information when processing sequential data while ignoring unimportant parts. This helps improve the accuracy and performance of the model, especially when dealing with long sequential data.

(2) Exploring relationships between sequences from different perspectives: The multi-head attention mechanism maps Query and Key to different subspaces in a high-dimensional space, calculates their similarity, and combines attention information from various subspaces. The aim is to explore relationships between sequences from different viewpoints and enhance the performance of the attention layer by synthesizing these relationships.

(3) Reduce computational complexity: The multi--head attention mechanism reduces the computational requirements for individual vectors by focusing computations on different subspaces, thereby decreasing the amount of computation and helping to prevent overfitting.

(4) Introduce nonlinear activation: After each layer of the encoder, a feedforward network (Linear) is typically connected. This network consists of two layers of linear transformations, which are used to introduce nonlinear activation, change the space of attention output, and enhance the expressive power of the model.

Network Structure

Overview of Network Architecture

This paper uses Inception as the basic network architecture, and the network framework consists of two parts: the scene recognition network (PlacesCNN) and the object detection network (ObjectNet) [11], as shown in Figure 2. PlacesCNN extracts scene features, while ObjectNet extracts object features, and the two types of features are ultimately fused. However, in this process, due to the different categories and nature of object and scene features, and the fact that some objects have a smaller role in scene recognition (such as a computer in a bedroom), or may even have a counterproductive effect (such as a fully stocked bar in a restaurant), directly overlaying or connecting features may not fully utilize the role of object features. To improve the utilization efficiency of object features in indoor scene recognition, this paper applies an attention mechanism to transform the object features in ObjectNet, so that the dimensions of the object features are the same as those of the scene features, reducing the loss of object feature information. CG is used to suppress redundant information in the features (such as computers appearing in bedrooms, sofas appearing in restaurants, etc.), reducing the weight of irrelevant features and allowing the network to focus more on relevant target areas of the image, thereby improving the utilization efficiency of the object features.

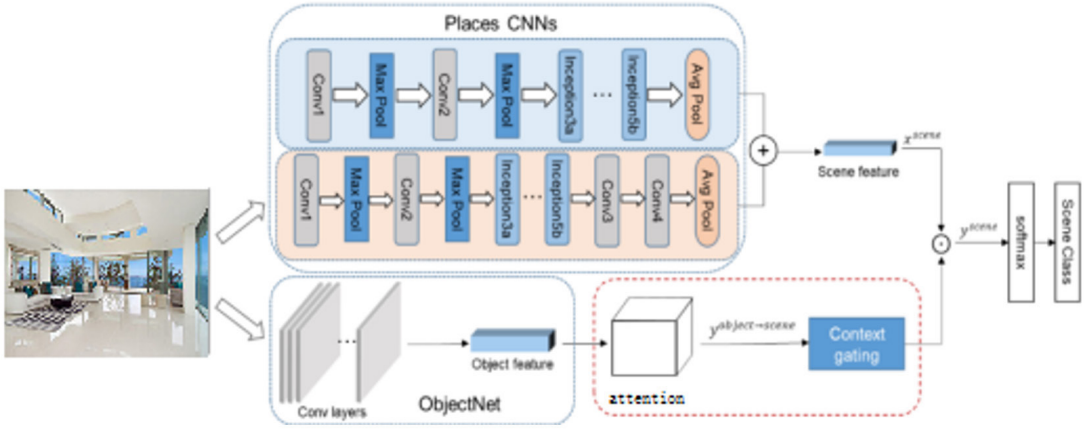

Figure 2: Network Structure

Firstly, object attributes and positional relationships in the image are expressed through position descriptors, and GloVe model is used to extract semantic information from the image to generate word vectors. Then, a data preprocessing algorithm is used to ensure data formatting, and the extracted semantic features are input into an LSTM model based on an attention mechanism. The attention mechanism improves the accuracy of feature recognition,

and its output is used as input for the subsequent CNN. Finally, scene classification is performed using the Softmax function.

Feature Transformation

To transform the object features extracted by ObjectNet into scene features, avoid the direct overlay or connection of two distinct features, namely object features and scene features, and improve the fusion effect at the feature level, an attention mechanism is introduced into the combined method of object detection and scene recognition, as shown in Figure 3. The attention mechanism is placed after ObjectNet to process the object features. The input object features are extracted from ObjectNet, and the output features represent the attention-adjusted representation. The attention calculation is as follows:

To convert the object features extracted by ObjectNet into scene features, prevent the direct superposition or concatenation of two distinct feature types—object features and scene features—and enhance the fusion effectiveness at the feature level, an attention mechanism is introduced in the integrated approach of object detection and scene recognition. As depicted in Figure 2, the attention mechanism is positioned subsequent to ObjectNet to manipulate the object features. The input object features, denoted as ${}_{i}nyx^{object}$, are extracted from ObjectNet, while the output features are designated as ${}_{i}nyy^{object\to scene}$. The computation of attention is outlined below:

$$y^{object\to scene} = Wx^{object} + b$$

Where in, ${}_{i}nyx^{object} \in R^{n}, W \in R^{n\times m}, b \in R^{m}$ Variable n is the dimensionality of the input vector, and m is the dimensionality of the output vector. If the ImageNet dataset is used to train the object module, then n=1000. If the Places 2 dataset is used to train the scene module, then m=365. The relationship between objects and scenes can be analyzed from the attention weights. If an object frequently appears in a particular scene, the corresponding weight will be higher; conversely, if an object rarely appears, the weight will be lower.

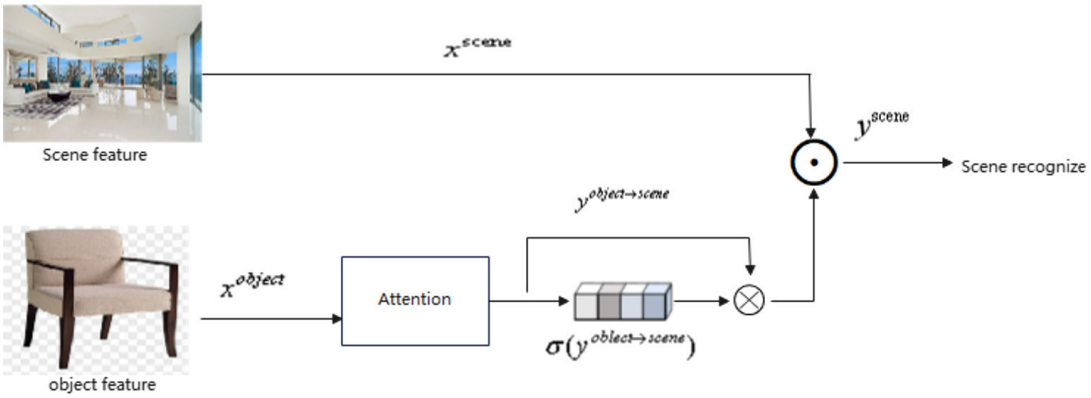

Figure 3: Feature Conversion

Redundant Information Suppression

Each scene image may contain various objects, and when a particular object appears in a scene image, it is highly likely that this scene belongs to a specific category. For example, the presence of a bed is significant for identifying a bedroom, or a bathtub for a bathroom. However, other irrelevant features can influence the network's judgment. To reduce the weight of irrelevant features, this paper introduces Context Gating, whose calculation is shown in Figure 2:

$$y'^{scene} = (y^{oblect \to scene} \otimes \sigma(y^{oblect \to scene}))$$

Based on the characteristics of the scene, the output is:

$$y^{scene} = y'^{scene} \Theta x^{scene}$$

Wherein, x represents the scene features extracted by the scene network, o represents the element-wise product at corresponding positions in the matrix, and σ(x) is the sigmoid activation function. Due to the characteristics of the sigmoid activation function, its left end approaches 0 asymptotically, while its right end approaches 1, forming a "gate" that restricts information. When irrelevant information passes through this gate, the function value tends towards 0, suppressing the irrelevant information.

Method Comparison

The approach proposed in this paper, integrating the multi-head attention mechanism with Context Gating (CG) technology, exhibits significant advantages in design philosophy compared to other methods. The following is a detailed comparison with ResNet-50 with Transfer Learning, DenseNet-121 with Data Augmentation, and VGG-16 with Spatial Pyramid Pooling (SPP).

ResNet-50, a deep residual network, excels in indoor scene recognition tasks through transfer learning. In contrast, the method in this paper further optimizes the processes of feature extraction and suppression by introducing the multi-head attention mechanism and CG technology. The attention mechanism enables the model to focus on salient features, while CG technology effectively mitigates the interference from redundant features, thereby enhancing the model's recognition capabilities in complex scenes.

DenseNet-121 leverages dense connections to enhance feature reuse and achieves remarkable results in indoor scene recognition with data augmentation techniques. The method presented in this paper, however, elevates recognition accuracy further through the enhancement of feature extraction via the multi-head attention mechanism and the optimization of feature suppression using CG technology. The multi-head attention mechanism allows the model to attend to multiple crucial feature regions simultaneously, improving recognition outcomes, while CG technology plays a pivotal role in suppressing irrelevant features.

VGG-16, combined with Spatial Pyramid Pooling (SPP) technology, excels in multi-scale feature extraction by fusing features across different scales. In comparison, the method in this paper achieves higher multi-scale feature extraction and fusion capabilities through the superiority of the attention mechanism in feature selection and the contribution of CG technology in suppressing redundant features. The attention mechanism enables the model to dynamically adjust its focus on different features, thereby more effectively processing complex scenes.

Experiments and Analysis

Dataset and Experimental Platform

The indoor scene dataset used in this paper comes from MIT indoor67 [12], which contains 67 indoor categories with a total of 15,620 images, and each category has at least 100 images. The SUN397 [13] dataset consists of 108,754 images, including 397 image categories. The ImageNet dataset, which ObjectNet uses to extract targets from scene images, contains 1,000 object categories. Some scenes from the datasets are shown in Figure 4.

The specific implementation of the algorithm in this paper uses the deep learning framework Tensorflow . The experimental environment is the Ubuntu 15 operating system, accelerated by two NVIDIA 1080Ti Graphics Processing Units (GPUs).

Training Time per Epoch: On two 1080Ti GPUs, the training time for each epoch is approximately 2 hours for the MIT Indoor67 dataset, and around 5 hours for the larger SUN397 dataset.

Total Training Time: Generally, deep learning models require tens to hundreds of epochs to achieve convergence. Assuming 100 epochs of training on the MIT Indoor67 dataset, the total training time is estimated to be 200 hours (approximately 8 days). For 50 epochs on the SUN397 dataset, the total training time is projected to be 250 hours (approximately 10 days). It's important to note that these estimates are based on specific hardware configu-

rations, and actual times may vary depending on different hardware or training parameter settings.

Memory Requirements: Due to the introduction of attention mechanisms and the Context Gating module, the memory requirements for a single forward and backward pass are relatively high. Each of the two 1080Ti GPUs, with 11GB of memory, is sufficient to meet the memory demands of this study.

Storage Space: The substantial intermediate results, model parameters, and log files generated during training require ample storage space. The total storage requirement is estimated to be in the range of tens to hundreds of GB, depending on the number and frequency of checkpoints saved during training.

Figure 4: Examples of some scenarios in the MIT Indoor67 dataset

Analysis of Experimental Results

The experiment was trained using gradient descent method, with a decay coefficient of 0.0001, a mini--batch size of 256, and an initial learning rate of 0.001. The evaluation index used was accuracy calculated through the confusion matrix, which is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Wherein, the meanings of TP (True positive), TN (True negative), FP (False Positive), and FN (False Negative) are shown in Table 1.

Table 1: confusion matrix

Real results	Predict results	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

TP: Positive samples predicted as positive by the model.

FP: Negative samples predicted as positive by the model.

FN: Positive samples predicted as negative by the model.

TN: Negative samples predicted as negative by the model.

The results indicate that by transforming the target features and suppressing redundant information, the expressive ability of the features and the effectiveness of feature fusion are enhanced. As shown in Table 2, Sum and

Concat represent two fusion methods of feature summation and concatenation, respectively. On the MIT Indoor67 dataset, the recognition accuracy of the proposed algorithm in this paper reaches 87.40%, which is up to 3% higher than several other algorithms; on the SUN397 dataset, the recognition accuracy of the proposed algorithm reaches 79.75%, which is 2.11% higher than other algorithms.

Table 2: The accuracy of this algorithm and other algorithms on the MIT Indoor67 and SUN397 datasets

Algorithm	MIT Indoor67	SUN397
MR-CNNs	85.40	77.64
MR-CNNs+ObjNet+Sum	86.21	78.10
MR-CNNs+ObjNet+Concat	86.73	78.26
The algorithm in this article	88.40	79.75

Compared to the detection method of the original network, the proposed method in this paper has a significant inhibitory effect on the redundant information contained in the features extracted by the object detection network.

To test the performance of the method in this paper on other network architectures, this paper takes ResNet-18 as an example and analyzes the recognition accuracy of the model on the MIT Indoor67 dataset through the combination of CCM and Context Gating, as shown in Table 3.

Model Analysis

Table 3: Comparison of the combined effect of attention and Context Gating

Attention	CG	Accuracy	Recall	F1
-	-	85.40	78.60	81.86
√	-	86.45	79.30	82.72
-	√	85.97	80.22	82.99
√	√	88.25	82.02	86.03

According to the data in Table 2, when the network is equipped with only CCM or CG, the accuracy rates are 86.35% and 85.91% respectively, representing an increase of 1.05% and 0.57% compared to the original network's 85.40%. However, when CCM and CG are used together, the accuracy rate reaches 88.25%. The experiments prove that the method in this paper has improved the accuracy of indoor scene recognition to a certain extent.

Conclusion

This paper proposes an improved indoor scene recognition method that combines object detection and scene recognition, aiming to address issues such as the inconsistency in nature and dimensions between object features and scene features, as well as feature information redundancy in indoor scene recognition. Through an attention mechanism, object features in scene images are transformed into scene features, and Context Gating (CG) is utilized to suppress redundant information in the features, thereby enhancing the role of object features in indoor scene recognition. Preliminary experimental results have been achieved.

This method could be particularly beneficial in smart home systems, where precise indoor scene recognition is crucial for automating household tasks and enhancing user experiences. For instance, smart home devices could better interpret different room types and adjust lighting, temperature, and even suggest activities based on the

identified scene.

Service robots would greatly benefit from this enhanced recognition method. Robots in environments such as hospitals, hotels, or homes for the elderly could more accurately identify and navigate different rooms, ensuring efficient and safe task completion.

Security monitoring is another area where this method proves advantageous. Enhanced indoor scene recognition can improve the accuracy of surveillance systems in detecting and identifying unauthorized access or unusual activities within different indoor environments. This leads to more reliable security systems that can promptly alert security personnel to potential threats.

However, the method proposed in this paper faces the following challenges:

One primary issue is the need for large-scale, annotated indoor scene datasets. These datasets are crucial for training convolutional neural networks from scratch, but they are often scarce, which limits the development and refinement of robust scene recognition models.

(2) Another challenge is the computational complexity associated with the attention mechanism and feature transformation processes. While these techniques enhance recognition accuracy, they also demand significant computational resources, which may not be readily available in all application settings.

Indoor scene recognition technology has broad application prospects in today's digital era. With the continuous development of technologies such as the Internet of Things, big data, and artificial intelligence, indoor scene recognition technology will play an important role in various fields.

References

1. Wang XQ, Li G, Plaza A, et al. (2021) Ship Detection in SAR Images via Enhanced Nonnegative Sparse Locality-Representation of Fisher Vectors[J]. IEEE Transactions on Geoscience and Remote Sensing, 59: 9424-38.

2. Lin Haitao, Liu Zichang, Cheang Chilam, et al. (2022) SAR-Net:Shape alignment and recovery network for category-level 6D object pose and size estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6697-707.

3. Lazebnik S, Schmid C, Ponce J, et al. (2006) Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 2169-78.

4. Espinace, Kollar T, Soto A, et al. (2010) Indoor Scene Recognition Through Object Detection [C]// IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 1406-13.

5. Zhao ZengShun, Feng Xiang, Wei Fang, et al. (2013) Learning Representative Features for Robot Topological Localization[J]. International Journal of Advanced Robotic Systems, 10: 1-12.

6. Zhou Bolei, Agata Lapedriza, Xiao Jianxiong, et al. (2014) Learning Deep Features for Scene Recognition using Places Database[J] Advances in Neural Information Processing Systems, 1: 487-95.

7. Antonio T, Pawan S, et al. (2009) Recognizing Indoor Scenes[C] Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 413-20.

8. Herranz L, Jiang S, Li X, et al. (2016) Scene Recognition with CNNs: Objects, Scales and Dataset Bias[C]//Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA, 571- 9.

9. Wang L, Guo S, Huang W, et al. (2017) Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNs [J]. IEEE Transactions on Image Processing, 26: 2055- 68.

10. Seong H, Hyun J, Kim E, et al. (2020) FOSNet: An Endto-End Trainable Deep Neural Network for Scene Recognition[J]. IEEE Access, 1.

11. Russakovsky Olga, Deng Jia, Su Hao, et al. (2015) ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 115: 211-52.

12. Nascimento G, Laranjeira C, Braz V, et al. (2017) A Robust Indoor Scene Recognition Method based on Sparse Representation[J]. Congress on Pattern Recognition. Springer, 408-15.

13. Xiao J, Hays J, Ehinger KA, et al. (2010) SUN database: Large-scale scene recognition from abbey to zoo[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 3485-92.

14. He K, Zhang X, Ren S, et al. (2016) Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 770-8.

15. Chu Jun, Su Yawei, Wang Lu (2018) Scene classification with adaptive adjustment of learning rate and sample training method[J]. Pattern Recognition and Artificial Intelligence, 31: 625-33.