Research on Large Scene Adaptive Feature Extraction Based on Deep Learning

Yahe Yang *, Iris Li, Ningjing Sang, Lipeng Liu, Xirui Tang *, Qiyuan Tian *

Posted Date: 11 September 2024

doi: 10.20944/preprints202409.0841.v1

Keywords: computing methodologies; artificial intelligence; computer vision; computer vision tasks; scene understanding

Article

Research on Large Scene Adaptive Feature Extraction Based on Deep Learning

Yahe Yang [1,*], Iris Li [2], Ningjing Sang [3], Lipeng Liu [4], Xirui Tang [5] and Qiyuan Tian [6]

[1] School of Business, George Washington University
[2] Courant Institute of Mathematical Sciences, New York University
[3] The Fu Foundation School of Engineering and Applied Science, Columbia University, ns3319@columbia.edu
[4] College of Engineering, Peking University
[5] College of Computer Sciences, Northeastern University
[6] George Washington University
* Correspondence: yahe.yang@gwmail.gwu.edu

Abstract: The proliferation of intelligent monitoring devices has led to the widespread adoption of background extraction technology across a multitude of domains, including intelligent transportation, video surveillance, human- computer interaction, and medical diagnosis. In this work, the model employs a multi-layer convolutional neural network structure, which enables the extraction and fusion of scene features from different scales in a layer-by- layer manner. This approach facilitates the comprehensive capture of complex scene information. The convolutional network comprises multiple layers, with each layer responsible for extracting features at a specific scale. The shallower layers capture more detailed feature information, whereas the deeper layers focus on more global scene features. The strategy of multi-scale feature fusion allows the model to fully capture multi-level and multi-dimensional information in the context of large-scale scenes. Furthermore, the incorporation of an attention mechanism enables the model to adaptively allocate attention to salient regions, thereby enhancing its capacity to discern intricate scenes by assigning greater significance to pivotal features. The experimental results demonstrate that the proposed method is effective in practice, as evidenced by its performance on public datasets.

Keywords: computing methodologies; artificial intelligence; computer vision; computer vision tasks; scene understanding

---

Adaptive feature extraction, Deep learning, Convolutional network, Multi-dimensional information.

1. INTRODUCTION

In the context of contemporary society, the acceleration of urbanisation and the continuous growth of the population have given rise to a number of significant challenges in relation to public safety and social management. The limitations of conventional video surveillance systems are readily apparent when confronted with the intricacies of contemporary challenges. Such systems are not only inefficient due to their reliance on human resources, but they are also susceptible to errors and miscalculations [1]. In particular, in the context of large-scale, multi-scenario surveillance environments, traditional methods are ill-equipped to meet the demands of real-time analysis and anomaly detection of voluminous video data. Consequently, the core issue in the development of video surveillance technology is the efficient and accurate extraction of key information from large volumes of video data.

The advent of machine learning has had a significant and far-reaching impact on numerous aspects of modern life. Its applications are pervasive, encompassing intelligent search, recommendation systems, audio and video processing, and a multitude of other domains where machine learning has brought about notable advancements. The rapid development of information technology and computing power has led to a diversification of the ways and scenarios in which data

 doi:10.20944/preprints202409.0841.v1

can be obtained and processed. However, this diversity also presents novel challenges for the algorithms in question [2]. The conventional machine learning algorithms are constrained in their capacity to extract valid features from raw data. Over the past few decades, the construction of a successful machine learning system has frequently necessitated a considerable investment of human resources, particularly when domain experts have been required to devise intricate feature extraction methodologies for the purpose of extracting useful features from raw data, such as pixel values in image samples, for utilisation in subsequent tasks. This process is often time-consuming and labour-intensive, and is dependent on domain-specific knowledge [3].

The advent of deep learning, however, has resulted in a fundamental shift in this regard. Deep learning facilitates end-to-end automatic feature extraction through the stacking of multiple layers of non-linear transformations, enabling the output of each layer to be transformed into higher-level and more abstract representations. The stacking of non-linear layers enables deep learning models to fit arbitrarily complex functions, thereby facilitating the automatic extraction of features without the necessity for manual design by human experts. Theoretically, this universal approximation capability of deep networks is widely accepted; however, in practical applications, the target loss function is usually constructed through the supervised learning paradigm [4]. However, the annotation information of the samples or the difference in sample quality will have a significant impact on the learning effect of the model, resulting in a challenging feature learning process. In particular, the current diversity of application scenarios makes it more challenging to obtain labelled data, while a large amount of unlabelled data is more common. This forces us to identify new solutions.

The advent of deep learning has marked a pivotal moment in the field of artificial intelligence, offering a transformative approach to handling intricate data sets and complex problems. It has facilitated significant advancements in numerous domains, including image recognition, speech processing, and natural language understanding. This has been achieved by developing multi-layer neural networks that empower computers to autonomously discern features and patterns from vast quantities of data [5]. In contrast to conventional machine learning techniques, a principal advantage of deep learning is its capacity to process and comprehend intricate unstructured data, thereby facilitating precise predictions and decision-making. One of the fundamental architectures of deep learning models is the convolutional neural network, which has proven to be an especially effective tool for image and video processing tasks. Convolutional neural networks extract local features through convolutional layers and perform feature dimensionality reduction through pooling layers in order to capture the spatial hierarchy present in the image. The combination of multi-layer convolution and pooling allows CNNs to extract more abstract and higher-level features layer by layer, thereby providing a deep understanding of complex image content. This feature extraction capability enables CNNs to excel in tasks such as image classification, object detection and face recognition [6].

Another significant deep learning architecture is that of recurrent neural networks, which demonstrate particular aptitude for the processing of sequential data. RNNs are capable of effectively capturing dependencies in the time dimension by sharing parameters in the sequence data, a technique that has been widely employed in tasks such as natural language processing, speech recognition, and time series prediction. Despite the gradient vanishing issues that are inherent to traditional RNNs, the capacity of deep learning models to address long-range dependencies has been markedly enhanced by the advent of variants such as Long Short-Term Memory Networks (LSTMs) and Gated Recurrent Units (GRUs).

The value of deep learning lies not only in its intricate network structure and robust feature extraction capabilities, but also in its end-to-end learning mode. This mode enables the entire system to automatically learn the optimal representation and decision rules from input to output, obviating the need for manual intervention. In the presence of a substantial corpus of training data, the deep learning model is capable of autonomously modifying its parameters in order to achieve an output that is as closely aligned as possible with the actual situation. This automated learning method significantly diminishes the necessity for input from domain experts, while enabling the model to accommodate a broader spectrum of potential applications [7].

 doi:10.20944/preprints202409.0841.v1

Nevertheless, the success of deep learning is not without its challenges. Firstly, the training of deep learning models frequently necessitates the availability of a substantial quantity of annotated data, which can prove costly or challenging to obtain in certain application domains. Secondly, the training process of deep learning models is frequently lengthy and resource-intensive, particularly when processing high-dimensional data, which necessitates substantial computational resources. Furthermore, the opaque nature of deep learning models hinders their interpretability, which can be a drawback in applications where high interpretability is a necessity.

In order to address these challenges, researchers have been pursuing a number of avenues of enquiry. For instance, transfer learning techniques permit models to utilise the knowledge acquired in one task to address a related task, thereby reducing their dependence on labelled data at scale. Furthermore, the development of hybrid models, which combine deep learning with other machine learning methods to leverage their respective advantages in order to improve overall performance, represents an important direction of research. Furthermore, the advancement of hardware technology, notably the emergence of graphics processing units (GPUs) and application- specific integrated circuits (ASICs), has markedly enhanced the training and inference speed.

2. RELATED WORK

The supervised contrastive learning method proposed by Khosla et al. markedly enhances the clustering effect of the representation of similar samples in the feature space by introducing supervised information into the contrastive learning framework. In traditional contrastive learning, samples of different classes are only compared in an unsupervised manner. While this may enhance feature discrimination, the aggregation of similar samples in the feature space may not be optimal. Khosla's method [8] incorporates label information, thereby facilitating the aggregation of samples belonging to the same class in the feature space, resulting in more discernible and compact clusters. This convergence not only facilitates enhanced accuracy in classification tasks, but also enhances the consistency and robustness of the model when processing similar samples. Furthermore, the proposed method employs a meticulously crafted loss function that prompts similar samples to converge in the feature space while simultaneously displacing samples of disparate classes, thereby refining the structural configuration of the feature space and enhancing the model's performance in subsequent tasks. The introduction of supervised information has demonstrated the superiority of the contrastive learning method over traditional approaches in a range of tasks, while also paving the way for a new synthesis of supervised and contrastive learning.

Zhong et al. [9] put forth an innovative approach to enhance the robustness and stability of feature representation. This approach involves introducing cluster-generated pseudo-labels (CGPL) into the contrastive learning framework. In particular, the data is initially grouped through the application of an unsupervised clustering algorithm, and subsequently, these clustering results are introduced as pseudo-labels into the contrastive learning process. The fundamental concept of this methodology is to utilise the clustering outcomes as a reference point, thereby facilitating a more concentrated clustering of samples within the same cluster in the contrastive learning process within the feature space.

This approach enables the model to not only capture the intrinsic structure of the original data but also to enhance these structures through contrastive learning, thereby improving the clarity and robustness of the feature representation in the clustered space. In comparison to the conventional unsupervised contrastive learning approach, the incorporation of pseudo-labelling effectively harnesses the insights derived from clustering, while circumventing the potential issues associated with unstable feature representation, such as the influence of noise or inaccurate clustering outcomes. Furthermore, this method is capable of generating representations with high discrimination and aggregation in the feature space, even in the absence of a substantial amount of labelled data. This enables an improvement in the performance of the model in downstream tasks.

The method proposed by Zhong et al. not only improves the quality of the clustering results, but also enhances the interpretability of the feature representations, thereby enabling the model to better understand and utilise the implicit structure present in the data. In practical applications, this

 doi:10.20944/preprints202409.0841.v1

strategy combining clustering and contrastive learning demonstrates robustness in a variety of complex scenarios, particularly in the context of high-dimensional and heterogeneous datasets. This method offers a novel approach to the advancement of unsupervised learning and contrastive learning, and also presents a more efficacious avenue for feature representation learning.

In a recent study, Li et al. [10] proposed an innovative semi-supervised learning (Semi) method with the aim of achieving a smoother and more consistent representation in the feature space. This was achieved by generating pseudo-labels on unlabeled samples and using these pseudo-labels to construct a graph structure. In their approach, pseudo-labelling is employed not only in the context of traditional supervised learning, but also serves to guide the model in the acquisition of more coherent feature representations through the utilisation of graph structures. This enables adjacent data points to exhibit greater similarity in the feature space. The crux of this method lies in leveraging the comparative information embedded within the graph structure, thereby enabling the pseudo-labels to more accurately reflect the intrinsic structure of the data. This approach effectively mitigates the potential adverse effects associated with inaccurate pseudo-labels.

By combining contrastive learning with graph structure, Li et al. were able to enhance the smoothness and robustness of feature representations, enabling the model to generate consistent feature representations even in the presence of unlabeled data. This resulted in not only an improvement in the model's performance in semi- supervised learning but also a notable enhancement in its generalization ability when processing new samples.

3. METHODOLOGIES

In this section, we propose a large-scene adaptive feature extraction model based on deep learning, which uses a multi-layer convolutional neural network structure to realize the layer-by-layer extraction and fusion of scene features at different scales, ensuring the comprehensive capture of complex scene information.

3.1. Convolution Operations with Multi-Layer Structures

The core of the model consists of a multi-layer convolutional neural network, each of which is responsible for extracting feature information at a specific scale. The shallow convolutional layer focuses on capturing detailed features, while the deep convolutional layer focuses more on global scene features. This multi-layer structure enables the model to extract and fuse multi-scale and multi-dimensional information layer by layer, so as to achieve a deep understanding of large-scale scenarios.

The input scene data is $X \in R^{H\times W\times C}$, where $H, W, and\ C$ represent the height, width, and number of channels of the input image, respectively. For the l-layer convolutional layer, its convolutional kernel $K^l \in R^{K_H\times K_W\times C^l_{in}\times C^l_{out}}$,
where K_H and K_W are the height and width of the convolution kernel, and C^l_{in} and C^l_{out} represent the number of
input and output channels, respectively. In a convolution operation, the output feature plot F^l of layer l can be obtained by the following Equation 1.

$$F^l_{i,j,c} = \sum_{m=1}^{k_H}\sum_{n=1}^{k_W}\sum_{d=1}^{C_{in}} K^l_{m,n,d,c}\cdot X_{i+m-1,j+n-1,d} + b^l_c \quad (1)$$

Where the variables i and j represent the position index of the output feature map, c denotes the channel index, and b^l is the corresponding offset.

The hierarchical convolution operation enables the shallow convolutional layer (such as the first and second layers) to extract low-level features in the scene including the edges, textures, and etc,

 doi:10.20944/preprints202409.0841.v1

whereas the deeper convolutional layer is capable of capturing high-level semantic information and global features through a larger receptive field.

3.2. Multi-Scale Feature Fusion

The design of the multi-layer convolutional network enables each layer of the feature map to contain information of varying scales. To effectively integrate this information, a multi-scale feature fusion strategy was employed. This entailed upsampling or downsampling the output feature maps of each layer to ensure they had the same dimensions in the spatial domain. Subsequently, these feature maps were added or stitched together element by element to form the final fusion feature representation F_{fusion} is expressed as Equation 2.

$F_{fusion} = Concat(\{U(F^1), U(F^2), \cdots, U(F^L)\})$ (2) Where the function $U(\cdot)$ represents the upsampling operation. F^l represents the feature diagram of layer l. L represents the total number of layers. In this manner, the integrated feature representation encompasses both comprehensive detail and global semantic information, thereby enabling the model to conduct a comprehensive analysis of the entire scene.

3.3. Attention Mechanisms

In order to enhance the model's capacity to focus on salient regions, we incorporate an attention mechanism based on feature fusion. The fundamental concept of the attention mechanism is to adaptively modulate the model's attention to varying regions by quantifying the significance of each position within the feature map. In accordance with the fusion feature representation F_{fusion}, the attention weight $A_{i,j}$ is initially calculated for each position, which is expressed as Equation 3.

$$A_{i,j} = \frac{\exp(Q_{i,j} \cdot K_{i,j})}{\sum_{i,j} exp(Q_{i,j} \cdot K_{i,j})} \quad (3)$$

Where the parameters $Q_{i,j}$ and $K_{i,j}$ represent query vectors and key vectors, respectively, obtained from F_{fusion} through linear transformations. Subsequently, a weighted sum of the fusion features is performed. According to the attention weight $A_{i,j}$, in order to obtain an enhanced feature representation and express as Equation 4.

$$F_{attn} = \sum_{i,j} A_{i,j} \cdot V_{i,j} \quad (4)$$

Where $V_{i,j}$ represents a vector of values obtained from F_{fusion} through the application of a linear transformation. The model employs an attention mechanism that enables the adaptive allocation of computing resources to critical areas of the scene, thereby enhancing the understanding and analysis of complex scenes.

4. EXPERIMENTS

4.1. Experimental Setups

In order to ascertain the efficacy of our proposed large-scene adaptive feature extraction model, we conducted experiments on the standard CIFAR-10 and CIFAR-100 datasets. These two datasets are widely used in image classification tasks and contain a substantial number of multi-class image samples, which can effectively assess the feature extraction and classification performance of the model in diverse scenarios. A model architecture comprising five layers of convolutional neural networks was constructed. Each convolutional layer was followed by the ReLU activation function and the maximum pooling layer, which enabled the extraction of features at varying scales in a sequential manner. In order to enhance the model's ability to understand complex scenes, a self-attention mechanism has been introduced on the output feature map of the convolutional network. This generates a weighted feature representation, calculated by determining the importance weight

 doi:10.20944/preprints202409.0841.v1

of each position. The model training employs the cross-entropy loss function, and the L2 regularisation and dropout mechanisms are utilised to prevent overfitting.

4.2. Experimental Analysis

The intra-class distance and inter-class distance of features serve as crucial indicators for evaluating the impact of feature extraction. The intra-class distance is defined as the mean distance between samples of the same class in the feature space. A smaller intra-class distance indicates that the samples of the same class are more tightly clustered in the feature space, which suggests that the extracted features are more consistent in distinguishing between the same class. The inter-class distance is defined as the mean distance between samples of disparate classes. A larger inter-class distance signifies that the classes are more disparate in the feature space, thereby enhancing the classifier's capacity to distinguish between different classes. The optimal outcome of feature extraction is the minimisation of intra-class distance and the maximisation of inter-class distance. This approach enhances the classification performance and robustness of the model. Figure 1 is an experimental comparison of the within-class and between-class distances of the features, comparing CGPL, Semi, and our methods. The intra-class and inter-class distances for each method are shown in the diagram, which shows the advantages of our method in feature extraction.

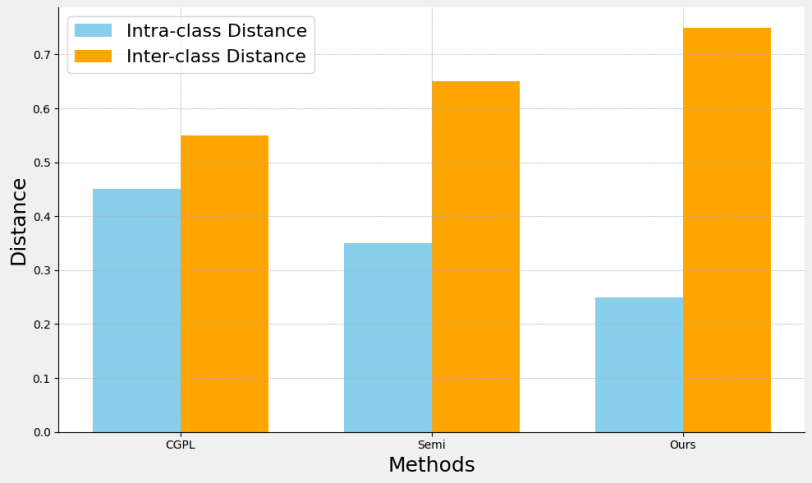

Figure 1. Comparison of Intra-class and Inter-class Distances.

Visualization of feature embedding is an intuitive method for evaluating and understanding the effectiveness of feature extraction from models. By downgrading high-dimensional features to a two- or three-dimensional space and visualizing the distribution of samples of different classes in that space, researchers can observe the aggregation and separation of samples of different classes in the feature space. Ideally, feature embedding should be characterized by a close spatial aggregation of samples of the same kind, without significant separation between samples of the same kind. Figure 2 is a comparison of the visual experiments of feature embedding, showing the distribution of features of CGPL, Semi, and our method in 2D space.

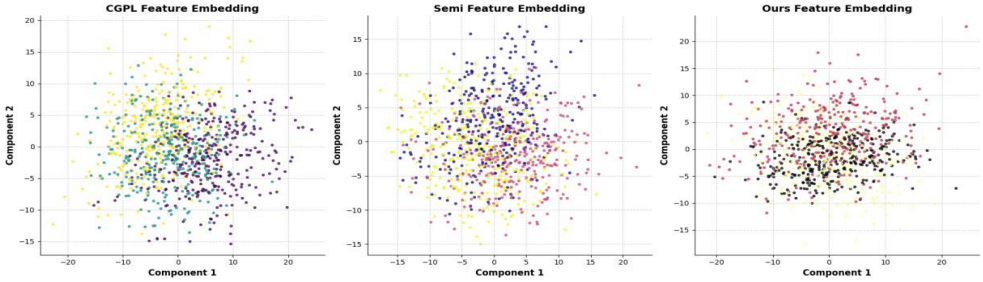

Figure 2. Feature Embedding Comparison Results.

 doi:10.20944/preprints202409.0841.v1

Pairwise distance correlation is an indicator that evaluates the effect of feature extraction by calculating the correlation between the distance of paired samples in the feature space and their distance in the original space. High correlation indicates that the feature extraction retains the geometry of the original data. Figure 3 shows the pairwise distance correlation comparison results.

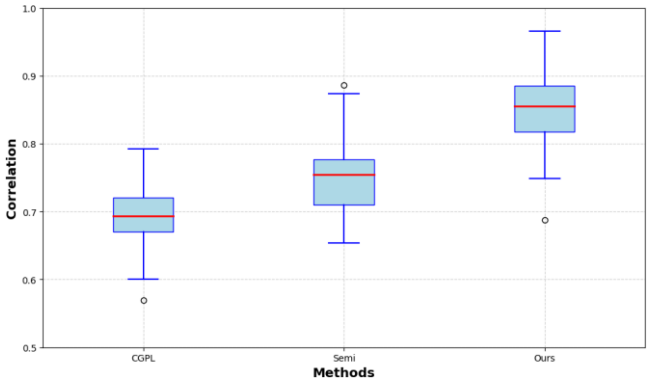

Figure 3. Pairwise Distance Correlation Comparison.

5. CONCLUSIONS

In conclusion, the model presented in this study employs a multi-layer convolutional neural network structure to extract and fuse multi-scale features, and incorporates an attention mechanism to enhance the model's focus on pivotal regions. This method is capable of capturing a comprehensive and accurate feature representation in large scenes, thereby providing substantial support for the comprehension and processing of complex scenes. The experimental results demonstrate that the proposed method exhibits a notable enhancement in performance for large-scale scene tasks, effectively addressing the challenges posed by scene complexity and diversity.

References

1. Xue, Guangdong, et al. "An adaptive neuro-fuzzy system with integrated feature selection and rule extraction for high-dimensional classification problems." IEEE Transactions on Fuzzy Systems 31.7 (2022): 2167-2181.
2. Zhao, Huimin, et al. "Feature extraction for data-driven remaining useful life prediction of rolling bearings." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-10.
3. Fard, Ali Pourramezan, and Mohammad H. Mahoor. "Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild." IEEE Access 10 (2022): 26756-26768.
4. Li, Wei, Qian Huang, and Gautam Srivastava. "Contour feature extraction of medical image based on multi-threshold optimization." Mobile Networks and Applications 26.1 (2021): 381-389.
5. Roy, Arunabha M. "Adaptive transfer learning-based multiscale feature fused deep convolutional neural network for EEG MI multiclassification in brain–computer interface." Engineering Applications of Artificial Intelligence 116 (2022): 105347.
6. Chen, Wuge, et al. "Fault feature extraction and diagnosis of rolling bearings based on wavelet thresholding denoising with CEEMDAN energy entropy and PSO-LSSVM." Measurement 172 (2021): 108901.
7. Lu, Siyu, et al. "Multiscale feature extraction and fusion of image and text in VQA." International Journal of Computational Intelligence Systems 16.1 (2023): 54.
8. Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33 (2020): 18661-18673.
9. Zhong, Huasong, et al. "Deep robust clustering by contrastive learning." arXiv preprint arXiv:2008.03030 (2020): 03030.
10. Li, Junnan, Caiming Xiong, and Steven CH Hoi. "Comatch: Semi-supervised learning with contrastive graph regularization." Proceedings of the IEEE/CVF international conference on computer vision. (2021): 9475-9484.