Advances in Engineering Innovation (2024) Volume 14 EWA Publishing
Published online: 20 December 2024 DOI: 10.54254/2977-3903/2024.19130

Application and optimization of deep convolutional neural networks in multimodal emotion recognition

Qunli Xie

Software Engineering Institute of Guangzhou

qunlixie98@163.com

Abstract. With the development of artificial intelligence, emotion recognition has become a hot topic in the field of human-computer interaction. This paper focuses on the application and optimization of deep convolutional neural networks (CNNs) in multimodal emotion recognition. Multimodal emotion recognition involves analyzing data from different sources—such as voice, facial expressions, and text—to more accurately identify and interpret human emotional states. This paper first reviews the basic theories and methods of multimodal data processing, then details the structure and function of deep convolutional neural networks, particularly their advantages in handling various types of data. By innovating and optimizing network structures, loss functions, and training strategies, we have improved the model's accuracy in emotion recognition. Ultimately, experimental results show that the optimized CNN model demonstrates superior performance in multimodal emotion recognition tasks.

Keywords: Multimodal emotion recognition, Deep convolutional neural networks, Human-computer interaction, Feature extraction

1. Introduction

As a core component of human-computer interaction, emotion recognition not only enhances the interactivity of machines but also plays a significant role in fields such as healthcare, education, and customer service. Traditional single-modal emotion recognition methods, such as those relying solely on voice or facial expressions, often suffer from low accuracy due to the one-sidedness of the information. The emergence of multimodal emotion recognition technology, by integrating information from multiple sensory channels, provides richer data support for the recognition process, significantly enhancing the accuracy and robustness of emotion recognition. Deep convolutional neural networks, with their powerful feature extraction capabilities, demonstrate great potential in multimodal emotion recognition. However, effectively integrating information from different modalities and optimizing network structures to accommodate different types of data remain key challenges in current research. This paper delves into the characteristics of multimodal data and the latest advancements in deep learning technology, exploring the application and optimization methods of deep convolutional neural networks in multimodal emotion recognition, hoping to provide theoretical support and technical guidance for the emotional intelligence of future human-computer interaction systems.

2. Basic theories and methods of multimodal data processing

2.1. Types and characteristics of multimodal data

Multimodal data processing involves collecting and analyzing data from different sensory channels, such as voice, vision (facial expressions, body language), and text. Each type of data has unique expressive forms and intrinsic characteristics. For example, voice data includes information on tone, intensity, and rhythm; visual data includes changes in expressions and body movements; text data directly reflects the semantics and emotional tendencies of language. In multimodal emotion recognition, effectively integrating these different types of data is key to improving recognition accuracy [1].

2.2. Data preprocessing and feature extraction methods

Data preprocessing is the first step in multimodal emotion recognition, mainly involving data cleaning, synchronization, and normalization. For example, facial expression videos require frame extraction and facial feature point calibration for preprocessing, while voice data needs noise reduction and segmentation for optimization. Feature extraction involves transforming raw data into a form more suitable for machine learning models. Common feature extraction methods in multimodal emotion recognition include Mel-frequency cepstral coefficients (MFCC) for voice and convolutional neural networks (CNN) for visual features, as well as word embedding techniques like Word2Vec for text data.

2.3. Application of deep learning models in data fusion

In multimodal data processing, deep learning models, especially convolutional neural networks (CNN) and recurrent neural networks (RNN), are widely used for feature fusion. CNNs excel in processing spatial data, such as images and videos, while RNNs are suitable for sequence data processing, such as voice and text. Data fusion typically occurs at the feature level, where features from each modality are extracted separately and then combined and fed into one or more deep learning models for comprehensive analysis. Additionally, attention mechanisms are used to enhance the model's ability to recognize key information, further optimizing the processing of multimodal data [2].

3. Structure and function of deep convolutional neural networks

3.1. Basic structure and working principle

Deep convolutional neural networks (CNN) are a type of deep learning network structure specifically designed to handle data with a known grid structure, such as images. CNNs mainly consist of convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract local features from input data; by applying different filters, the network can capture multi-level features ranging from simple to complex. Pooling layers reduce the dimensionality of features and enhance the robustness of the model. Fully connected layers, typically at the end of the network, are responsible for mapping learned high-level features to the final output categories or results [3].

3.2. Advantages of deep convolutional neural networks in visual information processing

CNNs are widely used in the field of visual information processing, especially exhibiting outstanding performance in facial expression recognition and human action recognition. Through deep network structures, CNNs can automatically learn complex feature hierarchies, which are effective for understanding and interpreting visual content. Additionally, with increasing network depth, more detailed and advanced abstract features can be captured, thereby improving the accuracy and granularity of emotion recognition.

3.3. Application of deep convolutional neural networks in voice and text information processing

Although initially primarily used for image data processing, CNNs have also shown potential in handling voice and text data. In voice emotion recognition, by transforming voice signals into spectrograms, CNNs can effectively extract emotional features from voice signals. For text, by combining convolutional layers with recurrent layers, CNNs can capture local sequential features in the text and effectively process order and context information, which is crucial for understanding the emotional tone of the text [4].

4. Optimization strategies for deep convolutional neural networks in multimodal emotion recognition

4.1. Innovation and optimization of network structures

To accommodate the complex requirements of multimodal emotion recognition, the structure of deep convolutional neural networks needs innovation and optimization. A common approach is to introduce a multi-branch network architecture, where each branch processes a type of modal data, and these branches are then merged at a deeper layer of the network for feature fusion. Additionally, introducing skip connections (such as those in residual networks) can help improve the training efficiency of the network and address the problem of gradient vanishing in deep networks, allowing the network to learn more complex and deep feature representations.

4.2. Selection and optimization of loss functions and activation functions

Choosing appropriate loss functions and activation functions is crucial for optimizing the performance of deep convolutional neural networks [5]. In multimodal emotion recognition, cross-entropy loss functions are commonly used for classification problems, but for more complex emotional states, customized loss functions may be required to more precisely model the diversity and nuances of emotions. Regarding activation functions, the ReLU function is widely used in convolutional layers due to its non-saturating nature, but in the output layer, depending on the specific task requirements, Sigmoid or Softmax functions may be chosen to suit different output needs.

4.3. Optimization of training strategies, including data augmentation and transfer learning

Optimizing training strategies is essential to improve the network model's generalization ability for multimodal data. Data augmentation is an effective method, where introducing modified data samples during training (such as rotating, scaling, cropping images, or changing the speed of voice) can significantly enhance the model's robustness. Additionally, transfer learning, by utilizing models pre-trained on large datasets to initialize or fine-tune the network for specific tasks, can effectively reduce overfitting and increase training speed.

5. Selected multimodal emotion recognition

5.1. Experimental platform and dataset

In this study, we selected a well-recognized multimodal emotion recognition dataset, which includes detailed facial expression videos, voice recordings, and corresponding text descriptions. These data are widely used to assess the performance of emotion recognition systems. We also set up an experimental platform that supports the training and testing of deep convolutional neural networks, capable of processing and analyzing a large volume of multimodal data.

5.2. Experimental design and implementation process

The experiment aims to verify the effectiveness of deep convolutional neural networks in multimodal emotion recognition and the practical effects of optimization strategies. The experimental design includes several steps: firstly, preprocessing and feature extraction for all modal data; secondly, setting different network configurations and optimization strategies to assess their impact on model performance; finally, evaluating the final model's performance through cross-validation methods [6].

5.3. Experimental results and analysis

The experimental results show that the optimized deep convolutional neural network model exhibits excellent performance in the task of multimodal emotion recognition. Specifically, by introducing a multi-branch network structure and optimized loss functions, the model has shown significant improvements in accuracy and generalization capability. Furthermore, the application of data augmentation and transfer learning strategies has further enhanced the model's robustness on unseen data.

5.4. Comparative analysis with other models

To comprehensively evaluate the efficacy of our model, we compared it with several mainstream multimodal emotion recognition models. The results indicate that our model surpasses or at least matches these existing models in various evaluation metrics, particularly showing stable performance when dealing with highly imbalanced datasets [7].

6. Conclusion

Through in-depth research and experimental analysis, this paper has confirmed the effectiveness of deep convolutional neural networks in multimodal emotion recognition. The optimized network structure can more accurately parse and recognize complex human emotions, especially when integrating information from different modalities, displaying significant performance improvements. Additionally, the innovative training strategies used in this study, such as data augmentation and transfer learning, effectively enhance the model's generalization ability, allowing it to maintain high accuracy on new or unknown data. These achievements not only advance the technology of emotion recognition but also provide strong technical support for the emotional intelligence of future human-computer interaction systems. We hope future research will further explore more optimization techniques and application scenarios, aiming to achieve broader and deeper technological applications and social impacts.

References

[1] Pak, S., Park, G. S., Park, J., et al. (2024). Application of deep learning for semantic segmentation in robotic prostatectomy: Comparison of convolutional neural networks and visual transformers. *Investigative and clinical urology*, 65(6), 551-558.

[2] He, T., & Huang, W. (2024). Automatic identification of depressive symptoms in college students: an application of deep learning-based CNN (Convolutional Neural Network). *Applied Mathematics and Nonlinear Sciences*, 9(1).

[3] Kumar, R. P., Dipankar, M., Shibaprasad, S., et al. (2023). Optimization of microscopy image compression using convolutional neural networks and removal of artifacts by deep generative adversarial networks. *Multimedia Tools and Applications*, 83(20), 58961-58980.

[4] Liu, W., & Xia, M. (2023). The application of deep learning in computer vision: Facial emotion recognition based on convolutional neural network. *Journal of Physics: Conference Series*, 2646(1).

[5] Wang, S. (2023). Application of Deep Convolutional Neural Networks in Image Recognition and Classification in Library Management. *Wireless Personal Communications*, (prepublish), 1-18.

[6] Madhav, M., Ambekar, S. S., & Hudnurkar, M. (2023). Weld defect detection with convolutional neural network: an application of deep learning. *Annals of Operations Research*, (prepublish), 1-24.

[7] Cao, X., Yao, B., Chen, B., et al. (2023). Intelligent Tool Condition Monitoring Based on Multi-Scale Convolutional Recurrent Neural Network:Special Section on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications. *IEICE Transactions on Information and Systems*, E106.D(5), 644-652.