Integrating Multimodal Data for Deep Learning-Based Facial Emotion Recognition

Jialu Li

Software School, Fudan University, Shanghai, 200433, China

22302010078@m.fudan.edu.cn

Abstract. With the rapid development of neural networks, emotion recognition has become a research area of great concern. It has important applications not only in marketing and human-computer interaction but also holds significant importance for improving emotional computing and user experience. This paper studies various methods for emotion recognition in images and videos, utilizing convolutional neural networks (CNN), multi-layer perceptron (MLP), and fusion models. The Facial Expression Recognition 2013 (FER2013) image dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) audio and video dataset serve as the basis for this study. The experimental results indicate that ResNet18 outperforms others in image emotion recognition, attributed to its residual block design and the incorporation of regularization techniques. In the realm of video emotion recognition, the audio model based on MLP demonstrates a superior ability to identify emotional information. Although the fusion of image and audio models theoretically could enhance accuracy, the randomness of video frames prevents the fusion model from achieving the desired effect. Future research might further explore the application of time series models in video emotion recognition to capture continuous emotional changes within videos.

Keywords: Emotion recognition; convolutional neural networks; multilayer perceptron; model fusion.

1. Introduction

Emotion recognition is an important research topic in the field of neural networks, showing extensive application potential across various domains. Emotion recognition systems can enhance the intelligence of human-computer interaction, enabling computers and devices to better understand and respond to users' emotions. For instance, in marketing, emotion recognition technology can analyze consumers' reactions to advertisements or products, assisting companies in optimizing the user experience. In the health sector, it can monitor the mental health status of patients. In human-computer interaction, it allows robots to adjust their behavior based on emotional feedback, providing more personalized services [1,2]. With the expansion of these application scenarios, developing high-precision and efficient emotion recognition systems has become particularly important.

To meet the growing demand, researchers have achieved significant breakthroughs in neural network technology, particularly in emotion recognition tasks, where the performance of neural network models has been substantially improved. From early multi-layer perceptrons (MLPs) and radial basis function networks (RBF) to today's convolutional neural networks (CNNs) and deep learning methods, emotion recognition techniques have advanced rapidly [3,4]. These models not only excel in recognizing facial expressions but also process multimodal data, such as audio and video, to help the system understand emotional information more comprehensively [5]. Reviews indicate that the multi-layer structure of deep neural networks can effectively capture complex emotional features, thereby enhancing recognition accuracy and generalization ability.

In this context, this paper aims to explore how deep learning models can better identify emotions in images and videos. Firstly, the author trained and optimized the ResNet18 convolutional neural network model on the Facial Expression Recognition 2013 (FER2013) image dataset for efficient image emotion recognition [6]. Secondly, a multi-layer perceptron (MLP) model is trained on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset for audio emotion recognition [7]. Lastly, this work attempted to combine image and audio models to test their performance in video emotion recognition tasks. By comparing the effects of a single image model,

a single audio model, and a fusion model, this work analyzed their performance and discussed potential directions for future improvement.

2. Method

2.1. Dataset

This study utilized two datasets for emotion recognition tasks: one for image and audio training, and another for video testing, ensuring a comprehensive evaluation of the model's performance across different modalities.

The FER2013 dataset, a widely used facial expression recognition dataset, contains 35,887 grayscale images, each with a resolution of 48×48 pixels. These images are categorized into seven emotional categories: anger, disgust, fear, joy, sadness, surprise, and neutral. In this study, the dataset was employed to train and test image emotion recognition and to evaluate the performance of various convolutional neural network models on emotion classification tasks [6].

The RAVDESS Affective Speech dataset, specifically designed for affective speech and performance research, contains 1,440 voice audio files recorded by 24 professional actors (12 men and 12 women). Each audio sample features emotional voice speeches, with emotion categories including: neutral, angry, happy, sad, afraid, disgusted, and surprised. This study utilized the audio data from RAVDESS to train a MLP model for recognizing emotions expressed through speech [7].

To test the multimodal fusion method, the RAVDESS video dataset was also employed in this study. The dataset comprises 735 video samples, each synchronized to contain both visual (facial expressions) and audio (voice expression of emotion) information. This work extracted audio and image information from the video datasets for audio and visual emotion recognition, respectively, and assessed the fusion effect of audio and image models.

2.2. Models

2.2.1 Image classification

In the field of image recognition, numerous neural network models are available for emotion recognition tasks. This project selects the ResNet18 neural network model. Compared to traditional convolutional neural networks, such as VGGNet and AlexNet, ResNet effectively addresses the common issues of gradient disappearance and gradient explosion in deep networks by introducing residual blocks [8]. The skip connections in the residual blocks allow the input to bypass several layers directly, bypassing the activation function and passing initial information to deeper layers of the network. This not only ensures efficient information transmission within the deep network but also reduces the loss of information that may occur as the network's depth increases [9]. This innovative structural design enables ResNet to construct deeper networks without compromising performance due to excessive depth.

In this project, ResNet18 consists of 18 layers, including multiple residual blocks, each containing two 3x3 convolutional layers [9]. This structure not only enhances the model's performance in image emotion recognition but also reduces computational complexity through a shallower network architecture, ensuring the model performs well in environments with limited computing resources. Additionally, the residual structure in ResNet18 enables the model to capture more details and features when dealing with more complex emotion classification tasks.

To further improve the model's generalization ability and prevent overfitting during training, this paper introduces two techniques: Dropout and L2 regularization. Dropout avoids overreliance on specific neurons by randomly dropping a certain percentage of neurons during training, thus enhancing the network's robustness. L2 regularization limits the occurrence of excessively large weights by imposing a penalty term on the weights, preventing the model from overfitting to the training data. The combination of these two regularization techniques helps the model better adapt to previously unseen image data, thereby improving its performance in real-world scenarios

2.2.2 Audio classification

For the audio emotion recognition task, this paper adopts a MLP classifier. The MLP is a feedforward neural network with a relatively simple structure, consisting of an input layer, multiple hidden layers, and an output layer. The model is trained using a backpropagation algorithm, which gradually adjusts the weights in the network to minimize the cross-entropy loss function, thereby optimizing the model's sentiment classification performance [10].

In terms of audio feature extraction, this study utilizes the librosa library to extract a series of key audio features that can effectively represent emotional information in audio. The main features extracted include:

Mel-frequency cepstral coefficients (MFCC): The audio signal is converted into a spectrum through Fourier transform, and then the features obtained by Mel scale and cepstral transform can effectively capture the spectral envelope in speech, which is suitable for emotion recognition tasks.

Chroma features: These represent the tonal information in the audio. This feature maps frequency information in the audio to 12 different scales and is suitable for analyzing tonality and emotion in speech or music.

Mel-spectrogram: By mapping the audio spectrum to the Mel scale, it reflects the frequency energy distribution, simulating the perception of human ear hearing, and is especially suitable for emotion recognition.

These extracted audio features are standardized and then input into the MLP model for training. The hidden layer of the MLP model uses the ReLU activation function to introduce non-linearity, thus improving the model's ability to learn complex features. The output layer uses the softmax activation function to convert the model's output into a probability distribution for each emotion class. Through multiple rounds of training, the MLP gradually learns the mapping relationship between audio features and emotion categories, thus achieving effective recognition of audio emotions.

2.2.3 Fusion model

To enable the image model to handle video emotion recognition tasks, this work employed the method of intercepting random frames to extract static image information from the videos. Specifically, a number of frames are randomly selected from each video, and these frames are emotionally classified one by one using a pre-trained image model such as ResNet18. For each frame, the image model outputs an emotion probability distribution that represents the probability values the model assigns to the frame corresponding to different emotions.

To obtain an emotional prediction for the entire video, the author averages the emotional probability distributions of all the randomly selected frames. For each video, the probability values for each emotion across all frame classifications are summed and then averaged to derive the overall emotion probability distribution for the video. This approach allows the image model to output a probability distribution that reflects the emotional characteristics of the entire video, rather than relying solely on the prediction of a single frame. This frame-level probability fusion method helps reduce the influence of individual frame emotion fluctuations or misjudgments on the overall emotion recognition result.

Next, this work fused the emotion probability distribution from the image model with that generated by the audio model. The audio model produces an independent emotion probability distribution by processing the synchronized audio information in the video. The weighted average or simple average of the probability distributions from the image and audio models can be combined to generate a final emotion probability distribution. By fusing the results of the image and audio models, this work selected the emotion with the highest average probability as the final emotion output for the video.

2.2.4 Evaluation index

This study employs several common metrics to evaluate model performance: (1) Accuracy: This metric measures the proportion of correct predictions to the total number of predictions, providing an overview of the model's overall performance. (2) Precision: It reflects the accuracy of positive

identification by measuring the proportion of true positive predictions to all positive predictions. (3) Recall: This metric indicates the model's ability to capture positive classes by measuring the proportion of true positive class predictions to all actual positive classes. (4) F1 Score: It is the harmonic average of precision and recall, providing a balanced assessment when these two metrics are out of balance.

3. Experiments and Results

3.1. Training Details

In this experiment, this work configured several key hyperparameters to train the multimodal emotion recognition model effectively: (1) Learning Rate: the learning rate is set to 0.0001 to control the model's weight update speed. (2) Batch Size: The batch size was chosen as 32, balancing between computational efficiency and model accuracy during training. (3) Number of Epochs: The model was trained for 6 epochs to ensure sufficient learning, without risking over fitting. (4) Optimizer: Adam optimizer is used, which was selected due to its effectiveness in minimizing the loss function while efficiently handling the gradient descent process. (5) Dropout Rate: A dropout rate of 0.2 was applied to prevent over fitting by randomly dropping neurons during training. (6) L2 Regularization: The L2 regularization coefficient was set to 0.001, which adds a penalty to the model's weights, preventing large weight values and helping to improve generalization.

The model weights use PyTorch's default initialization method, ensuring that the model starts learning from a good starting point and that the initial weights have appropriate variance, avoiding the problem of disappearing or exploding gradients. To ensure repeatability of the results, the random seed for ResNet18 is set to 42, while in the audio MLP model, the random seed is set to 68.

3.2. Result Comparison

The performance comparison is demonstrated in Table 1. The detailed performance of MLP on different classes are shown in Table 2 and Table 3 shows the effectiveness of fusion model.

Table 1. Performance comparison of different CNN models.

	Test accuracy	F1 score	Precision	Recall	dropout	L2regularization
VGG19	0.6367	0.6392	0.6438	0.6367	/	/
VGG19	0.6411	0.6416	0.6446	0.6411	0.2	1e-3
Resnet18	0.6461	0.6521	0.6634	0.6461	/	/
Resnet18	0.6551	0.6549	0.6606	0.6551	0.2	1e-3
MobileNetV2	0.6386	0.6408	0.6453	0.6386	/	/

Table 2. Performance of MLP model.

Classes	Precision	Recall	F1-score	Support
0	0.72	0.70	0.71	37
1	0.62	0.62	0.62	34
2	0.62	0.73	0.67	41
3	0.59	0.55	0.56	44
4	0.67	0.62	0.65	16
5	0.67	0.62	0.64	47
6	0.70	0.74	0.72	31
accuracy			0.65	250
macro avg	0.66	0.65	0.65	250
weighted avg	0.65	0.65	0.65	250

The main feature of VGG is that it uses a smaller convolutional kernel and deep network structure, which significantly improves the performance of convolutional neural networks. But the problem is that there are too many parameters. It will not only occupy large memory, but also produce over fitting.

To solve this problem, Mobile Net introduces deep separable convolution. This greatly reduces the number of parameters and computations while maintaining high accuracy. It can run efficiently on mobile or embedded devices.

While Mobile Net excelled in lightweight, Resnet's Residual Block design solves the problem of disappearing gradients in deep networks, allowing gradients to propagate effectively to support deeper network training and maintain good performance.

Table 3. Effectiveness of fusion model.

	resnet18 on video	audio model	fusion model
percision	0.1178	0.67	0.66

4. Discussion

Among different CNN models, ResNet18, which incorporates L2 regularization and dropout layers, performs best on the FER2013 dataset for the following reasons: (1) Compared to a neural network with a larger structure, such as VGG19, the overall architecture size of ResNet18 is more compatible with the size of the FER2013 dataset. (2) Compared to the lightweight design of MobileNetV2, ResNet18 has more parameters and can therefore learn more features. (3) Adding L2 regularization inclines the model to choose smaller weights by imposing penalties on larger weights, which helps to avoid overfitting. (4) Adding dropout breaks the dependency between neurons and enhances the model's robustness.

The accuracy of the fusion model is not higher than that of the single model as expected because, in video processing, image analysis inevitably exhibits bias. The number of frames selected at random is highly likely to miss important frames that can represent emotions, so image processing alone cannot adequately represent the emotions of the entire video. Even the ResNet18 neural network, which performs well on the FER2013 dataset, does not excel in the recognition of video frames selected at random. This paper does not explore time-series-dependent models, which may perform better when analyzing videos with continuous image changes.

5. Conclusion

In this paper, the ResNet18 network is utilized to identify the emotions of people in images, and an audio model trained by an MLP (Multi-Layer Perceptron) is employed to identify emotions in videos. Initially, the image model with higher accuracy is trained through the FER2013 dataset, and the audio model with higher accuracy is trained using the audio portion of the RAVDESS training set. According to pre-experimental speculation, it was anticipated that the accuracy would increase after fusing the two models. However, for videos, there has never been a reliable method to correlate the mood of a single frame with the mood of the entire video. The outcome is that using only the audio model yields better results, whereas incorporating the image model in video analysis does not yield satisfactory outcomes. In the future, the author could explore the use of time series models, such as Temporal Convolutional Networks, Hidden Markov Models, or Recurrent Neural Networks, to address the continuous changes in video content

References

[1] Zeng Zhihong, Maja Pantic, Thomas S. Huang. Emotion recognition based on multimodal information. Affective information processing. London: Springer London, 2009: 241-265.

[2] Zhang Jianhua, Yin Zhong, Chen Peng, et al. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 2020, 59: 103-126.

[3] Mellouk Wafa, Wahida Handouzi. Facial emotion recognition using deep learning: review and insights. Procedia Computer Science, 2020, 175: 689-694.

[4] Lieskovska, Eva and Jakubec, Marov and Jarina, Roman, et al. A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 2021, 10(10): 1163.

[5] Saxena Anvita, Ashish Khanna, Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems, 2020, 2(1): 53-79.

[6] Challenges in Representation Learning: Facial Expression Recognition Challenge. URL: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data. Last Accessed: 2024/10/26

[7] Livingstone Steven R., Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 2018, 13(5): e0196391.

[8] Li Zewen, Liu Fan, Yang Wenjie, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] Rumelhart David E., Geoffrey E. Hinton, Ronald J. Williams. Learning representations by back-propagating errors. nature, 1986, 323(6088): 533-536.