Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/107/20241116

Exploiting Convolutional Recurrent Neural Networks for Enhanced EEG-based Emotion Recognition

Gengyu Li

College of Information and Communication Engineering, North University of China, Taiyuan, Shanxi, 030051, China

2105054220@st.nuc.edu.cn

Abstract. Emotion recognition is a branch of artificial intelligence that analyzes human emotional states through facial expressions, voice, or physiological signals. It enhances human-computer interaction, facilitating more personalized and empathetic technology experiences, crucial for fields like mental health, customer service, and human-robot interaction. In recent years, research on emotion recognition using these tools has grown rapidly, involving multiple interdisciplinary fields. With the aid of electroencephalogram (EEG)-based brain-computer interfaces (BCIs), the emotional states of users can be sensed and analyzed. It offers a direct, non-intrusive insight into user emotions, enhancing user experience and system responsiveness. This approach is crucial for developing adaptive artificial intelligence (AI) in fields like healthcare for personalized treatments and in entertainment for immersive experiences, advancing human-technology symbiosis. This paper compares five current machine learning (ML)-based emotion recognition methods leveraging EEG signals, aiming to evaluate their effectiveness and applicability in emotion recognition. The paper concludes that while both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have their strengths, the combination of them provides the best performance in EEG-based emotion recognition.

Keywords: Brain-computer interfaces, emotion recognition, convolutional neural network,Long short-term memory network, machine learning.

1. Introduction

Emotional recognition is a significant domain within artificial intelligence (AI), focusing on the interaction between computational systems and human emotions. Automatic emotion recognition technology, which utilizes electroencephalogram (EEG) signals and brain-computer interfaces (BCIs), identifies and interprets emotional states [1]. Advancements in these technologies, including reduced equipment costs, have enabled deeper research into the relationship between emotional states and EEG fluctuations, thus accelerating the rapid development of emotion recognition technology. Emotional states significantly impact decision-making processes, playing a role in either facilitating or hindering problem-solving. Positive emotional states not only enhance an individual's emotional intelligence but also contribute to greater success in personal and professional realms. Furthermore, a deeper understanding of one's emotional states aids in better mental health management and optimized work performance. Automatic emotion recognition systems provide critical insights into emotional dynamics, promoting effective communication between individuals and between humans and computational systems. Integrating automatic EEG-based emotion recognition technology into AI systems is expected to transform the way humans interact with their environment, enriching personal relationships and being crucial for developing efficient human-computer interaction AI systems. As emotion recognition technology continues to evolve, it will play an increasingly vital role in how AI systems adapt to and support human emotional needs [2,3].

Typically, voice cues, physiological data, or facial expressions are used as the basis for automatic emotion recognition. Though physiological signals are more robust and noise-resistant, they more accurately reflect emotional state fluctuations than external expressions like voice and facial expressions, which are subject to both individual subjective intentions and external environmental influences. Consequently, in the field of emotional computing, emotion recognition based on physiological signals—particularly multi-channel EEG signals—has drawn more attention in recent years.

Significant benefits have been demonstrated by deep learning in managing complicated data, particularly unstructured data. Its capacity to autonomously extract features—as opposed to classical machine learning, which depends on human created features—is its biggest asset. Because of this, deep learning now performs remarkably well in domains including biological signal processing, computer vision, and natural language processing. For EEG signal processing, deep learning has become a hot research direction.

Convolutional Neural Networks (CNNs) are primarily used for processing data with spatial structure and were first widely applied in the field of image processing [4]. For EEG signals, CNNs can capture the spatial correlations between different electrode positions, automatically extracting spatial features and avoiding the need for manually designing complex features. The advantage of CNNs lies in their ability to effectively extract local spatial information, especially for signals with local correlations. However, CNNs, which mainly target spatial features, have relatively limited capability in capturing time series features, thus performing poorly in tasks that rely on temporal information.

Long Short-Term Memory (LSTM) is a recurrent neural network that excels in modeling time-dependent data, particularly due to its ability to capture long-term dependencies via a unique memory cell structure [5]. Unlike traditional RNNs, LSTM mitigates the vanishing gradient problem, making it ideal for tasks involving long-sequence dependencies. However, while effective in temporal feature extraction, LSTM often neglects spatial information, such as inter-electrode relationships in EEG data.

This study conducts a comparative analysis of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. Using a 75-25 split on the preprocessed DEAP dataset, both models produced favorable performance outcomes, demonstrating their effectiveness in EEG-based emotion recognition [6] LSTM performed particularly well in processing sequential data, effectively capturing dependencies between time steps, making it very suitable for sequence prediction tasks. Although the data is not continuous, its inherent sequential characteristics allow LSTM to achieve significant predictive accuracy. In contrast, CNN is known for its computational efficiency, enabling faster predictions. This efficiency is a key consideration in this study for introducing CNN, as it can speed up processing without significantly reducing accuracy, making it highly valuable in applications where speed is crucial. The study found that LSTM, designed for processing time-related data, can effectively track and capture long-term dependencies, giving it an advantage in tasks that require memory across multiple time steps. CNN, on the other hand, has become the standard model for spatial pattern recognition tasks through the efficient data processing capabilities of its convolutional layers.

2. Related work

Deep learning has proven superior to traditional machine learning approaches in fields such as computer vision, natural language processing, and biomedical signal analysis. This trend is particularly evident in EEG-based emotion recognition, where deep learning models either act as classifiers post-feature extraction or function as end-to-end systems, bypassing manual feature engineering. For example, Yang et al. integrated differential entropy (DE) from EEG signals with continuous CNNs for classification [7]. Similarly, Song et al. designed DE features based on the spatial relationships between EEG electrodes

DOI: 10.54254/2755-2721/107/20241116

and used graph convolutional neural networks (GCNs) for classification [8]. Some deep learning architectures are inherently data-driven, eliminating the need for predefined feature extraction. For example, Alhagry et al. proposed a model capable of end-to-end processing of raw EEG signals, which learns features through LSTM recurrent neural networks (RNNs) and classifies using fully connected layers [9]. Additionally, Yang et al.A parallel convolutional recurrent neural network (CRNN) has demonstrated notable effectiveness in enhancing the accuracy of EEG-based emotion recognition [10]. However, the challenge of extracting highly discriminative features from EEG signals persists. Developing more sophisticated deep learning architectures that can directly extract and classify features from raw EEG data is crucial for advancing research in this area.

3. Method

3.1. Dataset

The DEAP dataset collection was conducted in two phases to provide comprehensive emotional and physiological data [6]. In the first phase, 14-16 participants rated 120 one-minute YouTube music videos across five emotional dimensions, with results recorded in CSV or XLS files. The second phase involved 32 volunteers providing physiological data, including raw EEG signals from 40 out of 120 videos, and facial expression videos for 22 subjects. The dataset also includes YouTube video links and a questionnaire.

EEG signals, originally sampled at 512 Hz, were preprocessed to 128 Hz, involving downsampling, filtering, artifact removal, and were provided. Each Python file contains a Data array with dimensions of 40 × 40 × 8064, representing data from 40 EEG channels for each video, with 8064 data points per channel, totaling 322,560 data points, and a labels array of shape 40 × 4, representing EEG data and emotional dimensions, respectively.

In the dataset's feature extraction process, Fast Fourier Transform (FFT) was utilized to reduce the data dimensions from (40,40,8064) to (58560,70), optimizing computational efficiency and improving model performance. The extracted features represent five critical EEG frequency bands: Theta (4–8 Hz), Delta (1–4 Hz),Beta (14–31 Hz)Alpha , (8–14 Hz), and Gamma (31–50 Hz). Using the PyEEG Python library, 70 features were derived. By transforming signals from the time to the frequency domain, FFT effectively computes the Discrete Fourier Transform (DFT) of the time series, aiding in the identification of relevant frequency-specific patterns. By employing an iterative method to calculate DFT coefficients, FFT significantly reduces computation time and complexity, while also minimizing rounding errors that may occur during computation. The model uses 14 electrode channels and five frequency bands, with a window size of 256 samples, and calculates the average power of the frequency bands within a 2-second window. The step size is 16 samples, thus updating every 0.125 seconds.

3.2. CNN

CNNs were originally prevalent in image processing due to their ability to efficiently handle data with spatial structure [4,11]. This strength has since been leveraged in other fields, including EEG-based emotion recognition.Image data has distinct spatial features, such as local correlations between pixels, which allows CNNs to automatically extract local features through convolution operations, significantly enhancing the performance of tasks like image classification and object recognition. This capability is also applicable to EEG signal processing, particularly in effectively capturing the spatial structural features present in EEG signals. EEG signals, which originate from the electrical activity at various electrode locations in the brain, have spatial distributions and inter-electrode correlations that are crucial for analyzing electroencephalogram data. CNNs can automatically extract these spatial features through their convolutional layers, eliminating the need for manually designing complex feature extraction algorithms, thus simplifying the feature engineering process.

The initial convolutional layer uses the Rectified Linear Unit (ReLU) activation function and contains 128 convolution kernels, each of size 3. The number and size of the kernels have been extensively fine-tuned through hyperparameter optimization, including grid searches and manual

DOI: 10.54254/2755-2721/107/20241116

adjustments. The input shape of the first 1D convolutional layer is (70,1), and it uses 'same' padding and a stride of 1 to ensure that the spatial dimensions of the input remain unchanged during the convolution process.

The convolutional layer's output is normalized via Batch Normalization to achieve zero mean and unit variance. A 1D max pooling layer with a window size of 2 is then applied, downsampling the feature map by extracting the maximum values within each window. With default settings for padding ('valid') and stride ('none'), the dimensions of the resulting feature map are computed using the following formula:

$$n_{out} = \left\lceil \frac{n_{in}+2p-k}{s} \right\rceil + 1 \tag{1}$$

Although CNNs excel at extracting spatial features, their performance in handling time series data is relatively limited. EEG signals contain not only spatial structural information but also temporal continuity and dynamic changes. Traditional CNNs are primarily optimized for spatial features, and their convolutional operations are less capable of capturing time series features. Therefore, when tasks rely on temporal information, such as time series prediction or dynamic pattern recognition, CNNs may not perform as well as other models specifically designed to handle time series data, like RNNs or LSTMs.

To fully leverage temporal information, it is often necessary to combine CNNs with other models that process time series data. This combined strategy can compensate for the shortcomings of CNNs in temporal feature extraction, enabling a comprehensive analysis of the complex spatial and temporal features in EEG signals. The integrated use of CNNs and time series models, such as RNNs or LSTMs, can enhance the overall effectiveness of electroencephalogram data analysis and improve the performance of emotion recognition and other complex analytical tasks.

3.3. LSTM

LSTMs, a specialized variant of RNNs, were developed to address the limitations of traditional RNNs in managing long-term dependencies. Originally designed for processing time series data, LSTMs are particularly well-suited for EEG signals due to their ability to capture temporal continuity. LSTMs achieve this through a gating mechanism, consisting of input, forget, and output gates, which collaboratively regulate the flow and retention of information [5,12]. All gates employ the sigmoid activation function, outputting values between 0 and 1 to control the degree of information retention. The mathematical representations of these gates are as follows:

$$i_t = \sigma\left[\omega_i\left(h_{t-1}, x_t\right) + b_i\right] \tag{2}$$

The equation determines the portion of new information at time step t to be retained in the cell state. The input gate's activation value, computed via the sigmoid function, quantifies the importance of this incoming information, ensuring that relevant data is preserved for long-term dependencies..

The equation for the forget gate is:

$$f_t = \sigma\left[\omega_f\left(h_{t-1}, x_t\right) + b_f\right] \tag{3}$$

The role of the forget gate is to determine which information needs to be discarded. Its output is modulated by the sigmoid function to decide how much of the previous information to retain in the cell state, thereby facilitating the clearance of information that is no longer useful.

The equation for the output gate is:

$$o_t = \sigma\left[\omega_0\left(h_{t-1}, x\right) + b_0\right] \tag{4}$$

The role of the output gate is to generate the final output activation value, determining the content of the output at the current time step. The activation value, calculated by the sigmoid function, is used to adjust the impact of the cell state, ensuring the effectiveness of the output information.

In experiments, bidirectional LSTM layers are primarily applied. The design of these equations ensures that the LSTM network can effectively manage the storage and forgetting of information when processing long sequences of data, thereby improving the performance of traditional neural networks in tasks involving long-term dependencies.

3.4. Convolutional Recurrent Neural Networks (CRNNs)

CNNs and LSTMs are combined in CRNNs to process data with temporal and spatial dependencies efficiently. Local patterns in photos or other data with spatial structure are examples of the spatial characteristics that are first extracted by the CNN in a convolutional neural network (CRNN). To be processed further by the LSTM, these spatial features are converted into high-dimensional feature vectors. Capturing temporal dependencies in the input data is the LSTM's main job. Through the interaction of input, forget, and output gates, the LSTM—an improved recurrent neural network—dynamically modifies the flow and forgetting of information, effectively modeling important temporal dependencies in sequences while ignoring short-term dependencies or unimportant noise.Through this technique, the LSTM can more accurately capture richer temporal information and simulate long-term contextual connections in sequential data. LSTMs are usually coupled in series with CNNs in the integrated architecture of CRNNs to handle the sequence of feature vectors collected by the CNN, supporting time series information for further tasks such as regression or classification. CNNs and LSTMs work together to create CRNNs, which are especially useful for jobs requiring the capture of both spatial and temporal data, like speech recognition and video processing. LSTMs' potent capacity to capture long-term sequence dependencies is a major contributing reason to the CRNN architecture's success.

4. Results

In the experimental results, the outcomes and conclusions based on the aforementioned methods were discussed. Various model architectures were constructed, and multiple training sets were attempted. As shown in Table 1, the LSTM model achieved an accuracy of 88.6% with a 75-25 training-test dataset split, while the CNN model obtained 87.4% accuracy, and the CRNN model achieved an accuracy of 90.6%.

Table 1. Performance comparison of various models.

	Accuracy	Loss
CNN	88.6%	0.741
LSTM	87.4%	0.401
CRNN(CNN+LSTM)	90.6%	0.398

In order to provide a fair comparison, the performance of the CNN, CRNN, and LSTM models was evaluated in this study utilizing a shared dataset and a 75-25 split. With 88.6% accuracy, the CNN model successfully captured geographical features, however it had issues with long-term temporal connections. Time series data was handled exceptionally well by the LSTM model, which achieved 87.6% accuracy because to its sophisticated gating mechanisms. With a 90.6% accuracy rate, the CRNN model—which combines the temporal modeling of LSTM with the spatial feature extraction of CNN—performed better than both, showcasing its prowess in handling complicated data with both spatial and temporal information.

5. Discussion

This study indicates that the combined CNN-LSTM model outperforms the use of either model alone on the dataset, yet it has some shortcomings. Firstly, the current experiments only considered the basic integration of CNN and LSTM. Future research could explore more complex model combinations and optimization strategies, such as incorporating additional deep learning algorithms or feature fusion methods, to further enhance performance. Secondly, the experimental data and application scenarios in this study are limited, so the adaptability to different datasets and real-world application environments still needs further validation. Future work should consider testing the generalization capabilities of these models across a broader range of application scenarios.

Combining multiple algorithms has distinct advantages. Different algorithms excel at processing specific types of data and tasks. By effectively integrating them, the strengths of each algorithm can be fully leveraged, thereby enhancing the overall system performance. For instance, whereas LSTMs are better at identifying long-term dependencies in time series data, CNNs excel at extracting spatial features from image data. Combining the two not only utilizes CNNs' powerful feature extraction capabilities but also benefits from LSTMs' modeling of temporal dynamics, leading to a more comprehensive understanding of complex data patterns. To further enhance model performance and adaptability, future research may look into various deep learning model and algorithm types, such as Transformers and Graph Neural Networks (GNNs). By continuously optimizing and expanding model combinations, more accurate and efficient solutions can be achieved across various domains and tasks.

6. Conclusion

Technology that uses EEG signals for emotion identification has had a big impact on the field. An examination of deep learning-based techniques has been provided in this research, with an emphasis on how well CNNs and LSTM networks distinguish emotional states from EEG data. The findings underscore the complementary strengths of CNNs and LSTMs and the integration of these two approaches within a CRNN framework has yielded the most promising results, achieving the highest accuracy rate of 90.6% in experiments. The success of the CRNN model can be attributed to its ability to harness both spatial and temporal features, offering a more holistic analysis of EEG data. This dual capability addresses the limitations inherent in models that rely solely on either spatial or temporal data, thus enhancing the predictive power and accuracy of emotion recognition systems. In the future, the continued development and refinement of these deep learning models will be instrumental in advancing systems that can more effectively interact with and respond to human emotions.

References

[1] Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., ... & Marttinen, P. (2022). EEG based emotion recognition: A tutorial and review. ACM Computing Surveys, 55(4), 1-57.
[2] Dadebayev, D., Goh, W. W., & Tan, E. X. (2022). EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques. Journal of King Saud University-Computer and Information Sciences, 34(7), 4385-4401.
[3] Liu, H., Zhang, Y., Li, Y., & Kong, X. (2021). Review on emotion recognition based on electroencephalography. Frontiers in Computational Neuroscience, 15, 758212.
[4] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.
[5] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.
[6] Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, 3(1), 18-31.
[7] Yang, Y., Wu, Q., Qiu, M., Wang, Y., & Chen, X. (2018). Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In 2018 international joint conference on neural networks, 1-7.
[8] Alhagry, S., Fahmy, A. A., & El-Khoribi, R. A. (2017). Emotion recognition based on EEG using LSTM recurrent neural network. International Journal of Advanced Computer Science and Applications, 8(10), 355-358.

DOI: 10.54254/2755-2721/107/20241116

[9] Yang, Y., Wu, Q., Fu, Y., & Chen, X. (2018). Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In International Conference of Neural Information Processing, 433-443.
[10] Song, T., Zheng, W., Song, P., & Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Transactions on Affective Computing, 11(3), 532-541.
[11] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., ... & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics, 10(20), 2470.
[12] Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586.