Enhancing Psychological Counseling through Accurate Speech Emotion Classification Using Self-Attention-Based Deep Convolutional Neural Networks

Abstract

Speech Emotion Recognition (SER) has emerged as a vital technology in fields such as psychological counseling, human-computer interaction, and personality analysis. However, conventional CNN- or RNN-based methods often struggle to simultaneously capture fine-grained spectral details and long-range temporal dependencies, leading to limited accuracy and weak generalization across diverse emotional contexts. To address these challenges, the present work introduces a Self-Attention-Based Deep Convolutional Neural Network optimized with the Addax Optimization Algorithm (SA-DCNN-AOA). The framework integrates multi-scale feature extraction through a Multi-Scale Feature Transformer (MSFT) and a self-attention mechanism to enhance local–global feature learning, while the Addax Optimization Algorithm (AOA) performs automatic hyperparameter tuning for improved convergence and reduced overfitting. Using the CASIA Chinese Emotional Corpus containing 7,200 samples across six emotion classes, along with natural speech data for real-world validation, the proposed system follows a robust preprocessing pipeline involving pre-emphasis filtering, MFCC-based feature extraction, and temporal normalization. Experimental results demonstrate that the SA-DCNN-AOA effectively models both spectral and temporal features, achieving superior classification accuracy and generalization compared to conventional approaches, thereby offering a scalable and adaptive solution for real-time emotion recognition applications.

Related work

Zhou, L. and An, W., [1] suggested a DeepPsy-Based Multi-Source Mental Health Identification (DMHI) method that overcomes the drawbacks of questionnaire-based methods by integrating numerous data sources. LSTM is used to record temporal dependencies, 2D-CNN is used to extract daily online patterns, and a deep learning network is used to merge underlying features with online trajectory patterns, labeled with psychological data. With results of 0.71 accuracy, 0.75 recall, and 0.72 F1-score, 75% of impacted students were successfully identified. Improved detection accuracy and temporal modeling are benefits, but reliance on multi-source data and significant computational complexity are drawbacks.

Liu, Z.T., [2] created a few-shot SER technique using Meta-Transfer Learning with Domain Adaptation (MTLDA). To reduce overfitting, forgetting, and target domain adaptability problems, the procedure uses CASIA for pre-training and few-shot learning on Emo-DB and SAVEE with meta-transfer learning with domain adaptation. The Emo-DB results show 65.12% WAR and 64.50% UAR, whereas the SAVEE results show 58.84% WAR and 53.26% UAR. Robust performance with small sample sizes is a benefit; middling accuracy and reliance on pre-training databases are drawbacks.

Han, T., [3] suggested a Deep Residual Shrinkage Network with Bi-GRU (DRSN-BiGRU) technique that uses a self-attention mechanism to reduce noise and extract useful features. It combines convolutional layers, residual shrinkage, bi-directional gated recurrent units, and fully connected layers. Verification and optimization across the CASIA, IEMOCAP, and MELD datasets are part of the procedure. Accuracy results were 86.03%, 86.07%, and 70.57%, respectively, exceeding CNN-LSTM, CNN-BiLSTM, and DRN-BiGRU. Higher computing costs and more sophisticated models are drawbacks, but better accuracy and noise reduction are advantages.

Lin, L. and Tan, L., [4] suggested the Multi-Distributed SER utilizing MFCC and Parameter Transfer (MDSPT-SER) approach, which combines MFCC features with a pre-trained Inception-v3 network and a single-layer LSTM for emotion recognition. For the final prediction, the fully connected and classification layers are fine-tuned after MFCC features are extracted, fed into LSTM, and feature extracted using Inception-v3. Results show that, when compared to conventional SER frameworks, classification performance is greater across multi-distribution speech datasets. The use of pre-trained models and a high computing demand are drawbacks, but enhanced generalization and transferability are benefits.

Lu, C., [5] developed an Attentive Time–Frequency Neural Network (ATFNN) approach combines a time–frequency attention mechanism with a Time–Frequency Neural Network (TFNN) to address issues in SER. The method uses a Transformer-based frequency encoder (F-Encoder) and a Bi-LSTM-based time encoder (T-Encoder) to jointly learn time-frequency patterns. Additionally, by emphasizing long-range dependencies, F-Attention and T-Attention improve emotional discrimination. Tests on IEMOCAP, ABC, and CASIA verify better results than the most advanced techniques. Higher model complexity and training costs are drawbacks, whereas good feature discrimination and generalization are advantages.

Shahin, I., [6] developed a novel transfer learning technique dubbed Task-based Unification and Adaptation (TUA) to bridge the gap between intensive upstream training and task-specific downstream customization. Using multidimensional characteristics, the method blends task-specific flexibility with task-challenging unification to achieve powerful SER. Recognition rates in the studies were 91.2%, 84.7%, and 88.5% for the ESD, RAVDESS, and SUSAS datasets, respectively. The advantages include high accuracy and versatility on a range of datasets, but the disadvantages include growing model complexity and reliance on intensive upstream training.

Methodology

The present work proposed a SA-DCNN-AOA for effective speech emotion recognition. The workflow begins with data collection and preprocessing using the CASIA Chinese Emotional Corpus containing 7,200 samples across six emotions, supplemented with natural speech data for practical evaluation. Speech signals undergo pre-emphasis filtering to enhance high-frequency features, followed by normalization through truncation or padding to ensure a uniform length of 400 frames. MFCCs along with delta and delta-delta coefficients are extracted to form 400×40×3 feature maps. A MSFT module then captures both local and global features by applying multi-branch convolutions and a transformer encoder with advanced position encoding to strengthen temporal modeling. The extracted representations are fed into the SA-DCNN, where convolutional layers learn spectral and temporal cues, pooling, normalization, and dropout improve generalization, and a self-attention mechanism models long-range dependencies. Finally, fully connected layers with a softmax classifier generate the emotion predictions. To further boost performance, AOA is employed for hyperparameter optimization, automatically tuning learning rate, dropout probability, convolutional filters, kernel sizes, and attention head dimensions through iterative exploration and exploitation. This integrated workflow ensures robust feature representation, efficient hyperparameter search, and high-accuracy speech emotion classification suitable for real-world scenarios.

Research objectives

1. To develop a self-attention-based deep convolutional framework (SA-DCNN) capable of capturing both local spectral and global temporal dependencies in speech signals.

2. To integrate a Multi-Scale Feature Transformer (MSFT) for enhanced multi-resolution temporal modeling and feature fusion.

3. To employ the Addax Optimization Algorithm (AOA) for automatic hyperparameter tuning, improving convergence and preventing overfitting.

4. To evaluate the proposed SA-DCNN-AOA framework on the CASIA Chinese Emotional Corpus and natural speech data for real-world validation.

5. To achieve high accuracy, stability, and generalization in speech emotion recognition for practical applications such as counseling and personality assessment.

References

1. Zhou, L. and An, W., 2022. Data classification of mental health and personality evaluation based on network deep learning. Mobile Information Systems, 2022(1), p.9251598.
2. Liu, Z.T., Wu, B.H., Han, M.T., Cao, W.H. and Wu, M., 2023. Speech emotion recognition based on meta-transfer learning with domain adaption. Applied Soft Computing, 147, p.110766.
3. Han, T., Zhang, Z., Ren, M., Dong, C., Jiang, X. and Zhuang, Q., 2023. Speech emotion recognition based on deep residual shrinkage network. Electronics, 12(11), p.2512.
4. Lin, L. and Tan, L., 2022. Multi-Distributed Speech Emotion Recognition Based on Mel Frequency Cepstogram and Parameter Transfer. Chinese Journal of Electronics, 31(1), pp.155-167.
5. Lu, C., Zheng, W., Lian, H., Zong, Y., Tang, C., Li, S. and Zhao, Y., 2022. Speech emotion recognition via an attentive time–frequency neural network. IEEE Transactions on Computational Social Systems, 10(6), pp.3159-3168.
6. Shahin, I., Nassif, A.B., Thomas, R. and Hamsa, S., 2023. Novel Task-Based Unification and Adaptation (TUA) Transfer Learning Approach for Bilingual Emotional Speech Data. Information, 14(4), p.236.
7. Zhang, X., Lin, M., Hong, Y., Xiao, H., Chen, C. and Chen, H., 2025. MSFT: A multi-scale feature-based transformer model for arrhythmia classification. Biomedical Signal Processing and Control, 100, p.106968.

8. Alshehri, M.S., Saidani, O., Alrayes, F.S., Abbasi, S.F. and Ahmad, J., 2024. A self-attention-based deep convolutional neural networks for IIoT networks intrusion detection. IEEE Access, 12, pp.45762-45772.
9. Hamadneh, T., Kaabneh, K., Alssayed, O., Eguchi, K., Gochhait, S., Leonova, I. and Dehghani, M., 2024. Addax Optimization Algorithm: A Novel Nature-Inspired Optimizer for Solving Engineering Applications. International Journal of Intelligent Engineering & Systems, 17(3).