ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20231013001024

Original scientific paper

Improved ECA-ResTCN for Online Classroom Student Attention Recognition

Qun TU, Xiaoru ZHAO, Daqing GONG, Qianqian ZHANG*

Abstract: With the rapid rise of online classrooms, monitoring student engagement is critical but challenging for educators. This work explores how artificial intelligence (AI) and big data techniques can automatically evaluate student concentration levels in online courses. We developed an end-to-end ResTCN model combining ResNet and temporal convolutional networks (TCN) to extract spatial and temporal video features. Further, we introduced a CutMix data augmentation method and an efficient channel attention (ECA) module to enhance model training. Evaluated on a public dataset of student videos, our approach achieved 63.28% accuracy in classifying student engagement, outperforming state-of-the-art methods. The contributions are a novel spatiotemporal neural architecture, data augmentation strategy, and attention mechanism tailored for the student engagement recognition task. This demonstrates the potential of AI in creating smart education systems.

Keywords: attention mechanism; convolutional neural network; convolutional temporal network; student concentration

1 INTRODUCTION

Classroom teaching is the most fundamental teaching organization form in the process of talent cultivation in higher education [1]. The quality of teaching is generally an important data and indicator for evaluating the level of school teaching work. Meanwhile, students' classroom participation has always been an important research focus in classroom teaching, and it is positively correlated with academic performance [2].

Currently, teaching comes in various forms, with the main approaches being physical classroom teaching and online classroom teaching [3]. With the rapid development of online classrooms, the extensive learning resources have significantly expanded learners' channels for acquiring knowledge [4]. This has been effective in improving the knowledge levels of students. However, studies have shown that online courses have an extremely high "dropout rate," which previously exceeded 87% [5]. Research indicates that online teaching systems severely limit the interactivity of teaching [6], [7].

In the teaching process, learner feedback is one of the important criteria for measuring teaching quality and is essential for teachers. Learner feedback is not only reflected in knowledge assessment tests but is also more intuitively demonstrated in classroom teaching [8]. In offline physical teaching, while teachers can subjectively gain some teaching feedback during the course, the large number of students makes it impossible to pay timely and comprehensive attention to the concentration of each learner [9]. On the other hand, online classroom teaching makes it difficult for teachers to effectively monitor students' behaviour and assess the classroom's teaching effectiveness. In addition to learner classroom feedback, other evaluation indicators of classroom teaching quality are often presented through subjective methods such as open class evaluations and learner questionnaires. Compared to traditional teaching quality assessment methods, these approaches are somewhat less objective, lack sufficient data support, and do not produce results that are applicable for analyzing actual teaching quality [10].

Hence, researchers focus on quantitatively measuring students' concentration during classroom teaching using artificial intelligence (AI) and big data technology. An objective and supportive measurement of students' concentration assists teachers in adjusting their teaching methods and improving teaching quality, forming a positive feedback teaching mechanism [11]. This is also an important part of the teaching quality evaluation system in intelligent teaching.

In terms of end-to-end models for student concentration recognition, Zhang et al. innovatively improved and optimized the excellent model, Inflated 3D ConvNets (I3D), which is commonly used in the action classification field, and applied it to the field of automatic student engagement recognition [12]. Geng proposed an automatic concentration recognition method based on the 3D Convolutional Neural Network (C3D), which simulates appearance and motion information in videos and automatically identifies student concentration levels [13]. Abedi et al. introduced a new end-to-end hybrid neural network architecture consisting of the ResNet residual network and temporal convolutional network (TCN) [14]. Solanki et al. used the OpenFace tool to extract all possible features such as head posture, eye gaze, action units, and facial landmarks [15]. Mehta et al. proposed a 3D DenseNet Self-Attention Neural Network (DenseAttNet) [16]. Selim et al. introduced three new end-to-end hybrid deep learning models based on the ESEE dataset of Egyptian student electronic learning concentration videos and the public dataset DAiSEE on emotional states in electronic environments [17]. However, existing models suffer from issues such as excessive model parameters, poor long-distance feature capturing capability, inability for parallel computation, and low operational efficiency.

In this paper, we have based our model on the ResTCN (the combination of the ResNet and TCN) model and integrated the ECA (Efficient Channel Attention) attention mechanism module to enhance the spatial feature extraction capability of the model. We have also introduced the CutMix data augmentation method to enrich the data distribution of the sample set. Experimental results demonstrate that our proposed model has improved accuracy in student concentration recognition.

2 MODEL DEVELOPMENT

2.1 Model Architecture

This study combines the ResNet convolutional neural network and the TCN to detect student engagement levels, which allows the model to capture both spatial and temporal features of the video. The TCN is used to extract features from the spatial feature maps extracted by the ResNet that to enhance the model's ability to extract temporal features between frames in video data.

The model transforms the input videos into feature vectors of dimensions (3, 224, 224) through the data preprocessing module. The vector matrix is then fed into the ResNet convolutional neural network for feature extraction, obtaining spatial feature vectors for video frames. ResNet, with the presence of skip connections, can easily extend to very deep models. On the one hand, the model has many network layers, and the stacked convolutional modules have strong feature extraction capabilities, which can effectively handle the complex video frames in the student engagement recognition dataset.

Therefore, ResNet and TCN are combined to model and jointly train spatial and temporal data within the video frame sequence. TCN was chosen because it has advantages in modelling longer sequences and retaining historical memory compared to general recurrent architectures like LSTM and GRUs. While ResNet extracts spatial features from individual frames, TCN models the temporal changes within the frame sequence and outputs the detected student engagement levels. The ResTCN model structure is illustrated in Fig. 1.

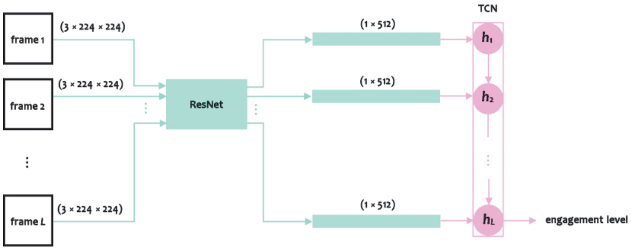

Figure 1 ResTCN model structure

The detailed training process of the ResTCN model is as follows: Given a video sequence, the input to ResTCN is an $L \times C \times H \times W$ tensor, where L, C, H, and W correspond to the number of frames, channels, frame height, and frame width, respectively. After removing the final fully connected layer of the standard ResNet18, it is used as a (trainable) feature extractor for single frames of the input video. It extracts feature vectors from consecutive frames as multi-dimensional inputs to the expanded TCN to simulate the temporal information within the video. The output at the final time step of TCN is passed through a fully connected layer and a Softmax function to detect the student's engagement level within the input video. In summary, the model takes a sequence of raw video frames as input and outputs categories corresponding to the student's engagement levels in the video. This is an end-to-end machine-learned classification process.

2.2 Data Augmentation with CutMix

Considering the issue of imbalanced sample distribution in the dataset, this study applies the CutMix technique for data augmentation. CutMix, proposed by Yun et al., [18] has shown promising results in various domains such as image classification, object detection, and speech recognition, and it has become a widely used data augmentation technique.

The fundamental idea of CutMix is to merge training samples of different student engagement levels, enhancing the diversity and richness of the dataset. Unlike traditional data augmentation methods, CutMix generates more diverse training data without altering the sample labels. By training on these newly generated samples, the model can better learn the relationships between different video frames from various engagement categories, thus improving the model's generalization ability.

The workflow of CutMix involves two main steps: random cropping and pixel-level mixing.

(1) Random Cropping and Pasting: Initially, a portion of the original training image is randomly cropped, and this cropped part is randomly placed into another training image. These two images are then combined to create a new training image.

(2) Pixel-Level Mixing: Subsequently, pixel-level mixing is performed on the combined image, with pixel intensity values weighted based on the area of the cropped region. This process generates the final mixed image.

2.3 Feature Extraction Module

The feature extraction module in this study comprises two main components: the ResNet spatial feature extraction module and the TCN temporal feature extraction module. Regarding the spatial feature extraction module, ResNet takes as input images of dimensions (16, 3, 224, 224). Through three stages of convolution operations and the utilization of a fully connected layer, it maps the feature vectors to four class outputs. Therefore, for video frames with input data dimensions of (2, 16, 3, 224, 224), the ResNet spatial feature extraction module processes these inputs by stacking residual building blocks, resulting in a final output of dimensions (6, 16, 32). In this context, "2" represents the batch size for processing data, "16" denotes the combination of 16 frames of video images, "3" signifies the number of channels in each frame, and "224" stands for the width and height of the images. After processing by ResNet, the spatial features of the images are encoded into a 1 × 32 feature vector.

To capture the temporal dependencies between video frames, the TCN model is applied to extract features from the spatial features of dimensions (2, 16, 32). It captures dependencies and causality features among video frames over time. The TCN model takes an input sequence with a length of 16, and each time step features a dimension of 32. The model primarily consists of multiple temporal convolution blocks, each comprising a temporal convolution layer and multiple residual convolution layers. The temporal convolution layer utilizes a 7 × 1 convolution kernel with a kernel size of 7, 128 channels, a stride of 1, padding of 6, and no batch normalization. Each residual convolution layer contains two convolution layers and a dropout layer. These two convolution layers share the same kernel size, channel numbers, and padding size, but they employ different dilated convolutions between the convolution layers. The dilation size follows an exponential increase, starting from 2^{10} and going up to 2^6 for seven steps. Each residual convolution layer uses batch normalization and ReLU activation functions, with a dropout layer following the second convolution layer. Additionally, both the input and output channel numbers for each temporal convolution block are set at 128. Finally, a fully connected layer maps the TCN model's output features to four-dimensional outputs, which are input to a Softmax function to produce classification scores for student engagement, as shown in Tab. 1.

Table 1 The vector dimensions of the input and output

Layers	Input	Output
ResNet18	2, 16, 3, 224, 224	2, 16, 32
Transpose	2, 16, 32	2, 32, 16
TCN: TemporalBlock0	2, 32, 16	2, 128, 16
TCN: TemporalBlock1-7	2, 128, 16	2, 128, 16
Softmax	2, 128, 16	1

2.4 Classification Module

After combining the feature vectors extracted from each frame, the feature sequence is fed into the classification module. The classification module's primary function is to perform a final dimension reduction on the extracted spatiotemporal feature maps, compressing the spatiotemporal fusion features to a fixed dimension. In this study, a four-class video spatiotemporal feature extraction network was built, resulting in this fixed dimension being 4. Consequently, the classification results now contain both spatial and temporal information about the video.

Finally, this study employs the classical Softmax classification network, as depicted in Tab. 1, to calculate classification scores for the compressed output, which is a 1×4 feature vector. This computation results in the final classification weights and ultimately yields the output categories (Very Engaged, Engaged, Not Engaged, Highly Not Engaged).

2.5 Attention Module

In the field of computer vision, attention mechanisms can also play a significant role. This is mainly manifested by the model's ability to actively learn pixel spatial information, temporal feature information, and background modelling during the training process. Furthermore, it can provide some degree of interpretability for "black-box" structures like deep neural networks. Fig. 2 illustrates the structure of the ECA-ResNet, where the ECA module is inserted into the residual branch.

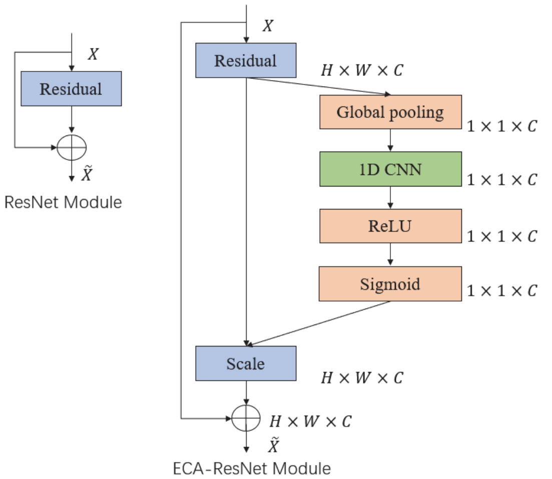

Figure 2 ECA-ResNet single module structure

The process begins with reducing the feature dimension to 1/*r* of the input dimension. This reduced feature then undergoes ReLU activation, followed by an upscaling to the original dimension through a Fully Connected layer. This approach offers several advantages over using a single Fully Connected layer: 1) it introduces greater non-linearity, enabling a better fit to the complex interrelations between channels; 2) it significantly reduces the number of parameters and computational workload.

3 EXPERIMENTS AND DISCUSSION

3.1 Data Description

This paper utilizes the DAiSEE dataset created by Gupta et al. [19] to evaluate the performance of the ResTCN model and compare it with innovative approaches. The dataset comprises 9,068 videos from 112 online course students, primarily focusing on students' emotional states, including boredom, confusion, engagement, and frustration. The dataset categorizes students' engagement levels into four levels: 0 (very high), 1 (high), 2 (low), and 3 (very low). The videos in the dataset have a length of 10 seconds, a frame rate of 30 frames per second, and a resolution of 640 × 480 pixels. This research primarily addresses the issue of student engagement recognition, focusing on a four-class classification with labels [0, 1, 2, 3], excluding emotional factors. The engagement labels in the DAiSEE dataset are ordered from high to low, corresponding to four distinct states: very engaged, engaged, disengaged, and highly disengaged. The ratio of training set, validation set and test set in this article is 6:2:2. The training set, validation set, and test set are distributed according to random sampling.

3.2 Model Configuration

In this section, we set the hyperparameters of the student engagement recognition model that fuses the ResTCN spatiotemporal features. As shown in Tab. 2, we configure various hyperparameters, including the learning rate, batch size, total global iterations, image dimensions, dropout rate, optimizer, activation function, loss function, and whether to perform random shuffling. For the optimizer's hyperparameters, as displayed in Tab. 2, we specify a momentum value, weight decay, and learning rate of 0.9, 0.01, and 0.001, respectively.

Table 2 Model training super parameter

Parameter	Value
Momentum	0.9
Weight decay	0.01
Learning rate	0.001
Batch size	2
Input	16, 3, 224, 224
Epoch	100

3.3 Prediction Results and Analysis

3.3.1 Comparison with Baseline Models

All experiments in this paper have been completed following the comparison of different embedding positions for the attention mechanism channel, resulting in an accuracy of 63.28%, as shown in Tab. 3. It can be observed from Tab. 3 that the accuracy is on par with the results of the latest models. Tab. 3 provides a comparison of state-of-the-art end-to-end models in the field of student attention recognition on the same dataset. It is evident from the table that student attention recognition is a challenging deep learning task, and the accuracy of mainstream four-class results currently remains in the range of 60-63%.

Table 3 Comparison of accuracy results in students' concentration

Model	Accuracy
C3D (2016)	48.1%
I3D (2019)	52.4%
LRCN (2019)	57.9%
DERN (2019)	60.0%
HDC-ASER (2020)	60.03%
Neural Turing Machine (2021)	61.3%
DFSTN (2021)	58.8%
C3D+TCN (2021)	59.9%
DenseAttNet (2022)	62.59%
ECA-ResTCN	63.28%

3.3.2 Comparison with or without CutMix

In this subsection, we analyze the impact of the CutMix data augmentation module on the model's classification performance by comparing the original model with the addition of the CutMix module. As shown in Tab. 4, both models exhibit varying degrees of accuracy improvement with the inclusion of the data augmentation module. Specifically, after incorporating the Mixup data augmentation technique, the accuracy of the ResNet + TCN model is 61.57%, representing a 0.54% increase over the original ResTCN model. In the field of machine learning, enhancing model accuracy is no easy feat, and each improvement demands a significant amount of time and effort. Sometimes, improvements in accuracy do not directly correlate with the workload. While a 0.54% accuracy boost might not seem substantial, it demonstrates the practical value of Mixup data augmentation for the ResTCN model.

CutMix technology increases the diversity and richness of the overall sample by fusing different samples. It helps the model learn more complex features, such as textures, backgrounds, colours, and other aspects from different samples, even from different classes. These features enable the model to better distinguish between images of different categories. Therefore, this experiment conducted a preliminary comparative training of these two modules, and the results indicate that CutMix achieved better results in training the ResTCN model, with an accuracy of 61.75%, which is a 0.72% improvement over the original model. The comparative accuracy results with the introduction of different data augmentation modules are presented in Tab. 4.

Table 4 Comparison of the accuracy results of the initial model

Model	Accuracy
ResNet + TCN	61.03%
Mixup + ResTCN	61.57%
Cutmix + ResTCN	61.75%

3.3.3 Comparison with or without Attention Mechanism

In this section, this article verifies the impact of the attention mechanism on the classification accuracy of the model. After training for 200 epochs, we tested the trained models on the test set. As shown in the table below, with the addition of the ECA attention mechanism, the model achieved an accuracy of 63.10%, surpassing the ResTCN composite model and outperforming the model with the SE (Squeeze-and-Excitation Networks) attention mechanism. This indicates that the ECA attention mechanism is better at capturing student attention information within the spatiotemporal fusion features, thus helping the model improve its accuracy in attention discrimination. Similarly, we also observed a certain improvement in accuracy in the ResTCN model with the addition of the SE attention mechanism, demonstrating that attention mechanisms enhance the efficiency and accuracy of feature extraction in ResTCN, as shown in Tab. 5. Although ResNet and TCN can simultaneously extract spatial and temporal features from the sample, they do not make a choice based on the importance of the feature channel. In other words, their mechanism of treating all channels equally during the feature extraction process easily results in a large amount of redundant student concentration information in the extracted feature maps. At the same time, it is also possible to ignore the concentration in the more important feature map channels. information. The introduction of the attention mechanism further enhances the model's ability to perform targeted feature extraction from spatiotemporal fusion feature maps. These mechanisms assign higher weights to channels containing important information about student attention and lower weights to those with a lot of irrelevant information, thereby enhancing the overall performance of the model.

Table 5 Comparison of experimental results of two different backbones

Model	Accuracy
Cutmix + ResTCN	61.75%
Cutmix + SE-ResTCN	62.87%
Cutmix + ECA-ResTCN	63.10%

4 CONCLUSION

In conclusion, this work makes three key contributions. First, we designed an end-to-end ResTCN neural network architecture combining ResNet and TCN to effectively extract spatiotemporal features from video data for student engagement recognition. Second, we implemented a CutMix data augmentation technique to expand the diversity of training samples. Third, we incorporated a lightweight ECA attention module to enhance feature learning by weighting inter-channel relationships. Experiments demonstrated that our model achieves state-of-the-art accuracy of 63.28% on a public dataset of student engagement videos. The results highlight the potential of AI techniques to enable smart education systems that can automatically monitor student concentration and feedback. Future work can explore personalization of models using additional student metadata and testing on larger video datasets. Overall, this research takes an important step toward intelligent tools for evaluating and improving online classroom teaching quality.

Acknowledgments

This research was supported by the Yunzhong University Project 2022 (2022MU025), the Shanghai Data

Exchange project, the Fundamental Research Funds for the Central Universities (JD2330), and Funds for First-class Discipline Construction in Beijing University of Chemical Technology (XK1802-5).

5 REFERENCES

[1] Zhang, X., Wang, J., Zhang, H., & Hu, J. (2017). A heterogeneous linguistic MAGDM framework to classroom teaching quality evaluation. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(8), 4929-4956. https://doi.org/10.12973/eurasia.2017.00966a

[2] Lee, A. V. Y. (2021). Determining quality and distribution of ideas in online classroom talk using learning analytics and machine learning. *Educational Technology & Society*, 24(1), 236-249.

[3] Julija, M., Biruta, S., & Jevgenija D. (2021). Influence of the Pandemic Caused by Covid-19 to the Teaching Staff of the Higher Education Institutions. *Journal of Service, Innovation and Sustainable Development*, *2*(2), 1-11. https://doi.org/10.33168/SISD.2021.0201

[4] Lu, D. & Guo, F. (2022). Application of wearable motion sensor in business English teaching. *Computer Science and Information Systems*, *19*(3), 1481-1498. https://doi.org/10.2298/CSIS210320020L

[5] Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings*, 5825-5834.

[6] Yuan, T. (2021). Algorithm of classroom teaching quality evaluation based on Markov chain. *Complexity*, *2021*, 1-12. https://doi.org/10.1155/2021/9943865

[7] Chen, X. (2022). Multimedia teaching system based on art interaction technology. *Computer Science and Information Systems*, *19*(3), 1517-1532. https://doi.org/10.2298/CSIS220405026C

[8] Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, *26*, 5-28. https://doi.org/10.1007/s11092-013-9179-5

[9] Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504-509. https://doi.org/10.1002/pra2.487

[10] Mok, K. H., Xiong, W., & Bin Aedy Rahman, H. N. (2021). COVID-19 pandemic's disruption on university teaching and learning and competence cultivation: Student evaluation of online learning experiences in Hong Kong. *International Journal of Chinese Education*, *10*(1), 221258682110070. https://doi.org/10.1177/22125868211007011

[11] Stephen, A. & Ahmad, N. F. (2023). Evaluation of The Success of Business Travel Management System Using Delone& McLean Approach. *Journal of System and Management Sciences*, *13*(4), 199-213. https://doi.org/10.33168/JSMS.2023.0412

[12] Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., & Li, J. (2019). An novel end-to-end network for automatic student engagement recognition. *ICEIEC*, 342-345. https://doi.org/10.1109/ICEIEC.2019.8784507

[13] Geng, L., Xu, M., Wei, Z., & Zhou, X. (2019). Learning deep spatiotemporal feature for engagement recognition of online courses. *SSCI*, 442-447. https://doi.org/10.1109/SSCI44817.2019.9002713

[14] Abedi, A. & Khan, S. S. (2021, May). Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. *CRV*, 151-157. https://doi.org/10.1109/CRV52889.2021.00028

[15] Solanki, N. & Mandal, S. (2022, December). Engagement Analysis Using DAiSEE Dataset. *ICARCV*, 223-228. https://doi.org/10.1109/ICARCV57592.2022.10004250

[16] Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, *52*(12), 13803-13823. https://doi.org/10.1007/s10489-022-03200-4

[17] Selim, T., Elkabani, I., & Abdou, M. A. (2022). Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm. *IEEE Access*, 10, 99573-99583. https://doi.org/10.1109/ACCESS.2022.3206779

[18] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, 6023-6032.

[19] Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885. https://doi.org/10.48550/arXiv.1609.01885

Contact information:

Qun TU
School of Economics and Management,
Beijing University of Chemical Technology,
Beijing 100029, China
No.15, Beisanhuandong Road, Chaoyang District, Beijing, China
E-mail: tuqun@buct.edu.cn

Xiaoru ZHAO
School of Economics and Management,
Beijing Jiaotong University,
Beijing 100044, China
No. 3, Shangyuancun, Haidian District, Beijing, China
E-mail: 21125640@bjtu.edu.cn

Daqing GONG
School of Economics and Management,
Beijing Jiaotong University,
Beijing 100044, China
No. 3, Shangyuancun, Haidian District, Beijing, China
E-mail: dqgong@bjtu.edu.cn

Qianqian ZHANG
(Corresponding author)
School of Information, Beijing Wuzi University,
Beijing 101149, China
No. 1, Fuhe Street, Tongzhou District, Beijing, China
E-mail: zhangqianqian@bwu.edu.cn