
A novel deep learning model to improve the recognition of students' facial expressions in online learning environments

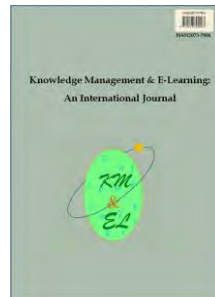
Heng Zhang

The University of Hong Kong, Hong Kong

Minhong Wang

The University of Hong Kong, Hong Kong

Zhejiang University, Hangzhou, China



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:

Zhang, H., & Wang, M. (2024). A novel deep learning model to improve the recognition of students' facial expressions in online learning environments. *Knowledge Management & E-Learning*, 16(1), 134–150. <https://doi.org/10.34105/j.kmel.2024.16.006>

A novel deep learning model to improve the recognition of students' facial expressions in online learning environments

Heng Zhang 

Faculty of Education
The University of Hong Kong, Hong Kong
E-mail: zhangheg@connect.hku.hk

Minhong Wang* 

Faculty of Education
The University of Hong Kong, Hong Kong
College of Education
Zhejiang University, Hangzhou, China
E-mail: magwang@hku.hk

*Corresponding author

Abstract: With the fast development of artificial intelligence and emerging technologies, automatic recognition of students' facial expressions has received increased attention. Facial expressions are a kind of external manifestation of emotional states. It is important for teachers to assess students' emotional states and adjust teaching activities accordingly. However, existing methods for automatic facial expression recognition have the limitations of low accuracy of recognition and poor feature extraction. To address the problem, this study proposed a novel deep learning model called DenseNetX-CBAM to improve facial expression recognition by utilizing a variant of densely connected convolutional networks (DenseNet) to reduce unnecessary parameters and strengthen the reuse of expression features between networks; moreover, convolutional block attention module (CBAM) was integrated to allow the networks to focus on important special regions and important channels when representing features. The proposed model was tested using 217 video clips of 33 students in an online course. The results demonstrated promising effects of the method in improving the accuracy of facial expression recognition, which can help teachers to accurately recognize students' emotions and provide real-time adjustment in online learning environments.

Keywords: Automatic facial expression recognition; Deep learning model; Recognition accuracy; Online learning environment

Biographical notes: Mr Heng Zhang is a Ph.D. candidate in the Faculty of Education, The University of Hong Kong. His research interests include technology-enhanced learning, aiming to understand and improve the processes and outcomes of computer technology-based teaching and learning.

Dr. Minhong (Maggie) Wang is Professor and Director of the Laboratory for Knowledge Management & E-Learning, Faculty of Education, The University of Hong Kong. She is a member of the Advisory Group on Academic Reviews of HKU. She is also K.uang-piu Chair Professor at Zhejiang University, and Eastern Scholar Chair Professor at East China Normal University. She is the Editor-in-Chief of Knowledge Management & E-Learning (indexed in Scopus

& ESCI). More details can be found at <http://web.edu.hku.hk/staff/academic/magwang>

1. Introduction

Online learning has been widely integrated in educational practice. Compared with traditional face-to-face classroom learning, online learning is more convenient, allowing students to learn without the constraints of time and place. However, there are concerns that online learning may limit direct emotional communication between teachers and students (Alawamleh et al., 2020). In online learning environments, it is difficult for teachers to quickly and accurately assess the emotional states of students and give timely feedback, which may affect online learning experience and outcomes.

Researchers have been investigating human emotions and the relationship between human emotions and facial expressions. Ekman and Friesen (1971) revealed that there are six primary human emotions including joy, sadness, fear, anger, surprise, and disgust. Moreover, he created a facial action coding system (FACS) to classify human expressions based on the correspondence between facial muscle movement units. The classification of the six emotions lays the foundation for the discrete emotion representation models. Mehrabian (2008) found that facial expressions account for 55% of emotional expressions, suggesting that facial expressions play a crucial role in the way learners express their emotions. These studies indicate that facial expressions are a kind of external manifestation of emotional states; emotional states can be detected through facial expressions.

With the fast development of artificial intelligence (AI) and emerging technologies, automatic recognition of students' facial expressions to support teaching and learning has received increased attention (Kazemitabar et al., 2019; Tonguç & Ozaydın Ozkara, 2020). In particular, deep learning approaches, a class of machine learning method effective for image recognition based on neural network architectures with multiple layers of processing units have been utilized for facial expression recognition, mainly through three steps: face detection, feature extraction, and classification (Revina & Emmanuel, 2021; Zou et al., 2018). However, the existing deep learning approaches have some limitations, such as low accuracy of recognition and poor feature extraction (Li & Deng, 2022), which affect the application of these approaches to support online teaching and learning.

This study aimed to address the limitation of existing approaches by proposing a new model for facial expression recognition, called DenseNetX-CBAM. It is a new deep learning model designed by improving the structure of densely connected convolutional networks (DenseNet) (Huang et al., 2018) and integrating the convolutional block attention module (CBAM) (Woo et al., 2018). Experiments were conducted to evaluate the performance of the proposed model in recognizing students' facial expressions in an online learning environment.

2. Literature review

2.1. Online learning

Online learning refers to education that occurs through the Internet, either synchronously or asynchronously, instead of in a physical classroom environment. Dhawan (2020) defines online learning as “a learning experience in a synchronous or asynchronous environment using different electronic devices with access to the Internet”. Online learning can improve the efficiency of learning to a greater extent by allowing for flexibility in the time and place of learning and teaching (Bender & Vredevoogd, 2006; Wong et al., 2019; Wong, 2023). Moreover, online learning with computer-based support can improve learning performance by fostering self-regulated learning (Chiu, 2022; Wei & Chou, 2020) and higher-order thinking (Cárdenas-Robledo & Peña-Ayala, 2019), which can be further improved by harnessing the potential of generative artificial intelligence (Zhu et al., 2023). However, the lack of face-to-face communication with teachers and peers and the lack of sufficient and timely feedback from teachers are the drawbacks of online learning (Mukhtar et al., 2020; Tang et al., 2023).

2.2. Facial expressions recognition technologies

To support interaction and communication in online learning environments, it is important to access students’ emotional states and give timely feedback. In this context, automatic recognition of students’ facial expressions to detect their emotions has received increased attention (Lasri et al., 2023; Ashwin & Guddeti, 2020). Traditional methods for facial expression recognition involve local binary patterns (Ojala et al., 2002), histogram of oriented gradient (Dalal & Triggs, 2005) and scale-invariant feature transform (Lindeberg, 2012). In general, these methods are constrained by human rules, and it is difficult to extract the deeper features of facial expressions.

With remarkable advances in information technologies, researchers have been exploring new methods, in particular deep learning models, for recognizing facial expressions. For instance, Krizhevsky et al. (2012) designed AlexNet, and Nayak and Sarvaiya (2022) improved AlexNet by introducing multi-scale convolution and batch normalization for conducting facial expression recognition recently. Meanwhile, lightweight convolutional neural networks combining attention modules have been employed to solve the problem of noise interference in non-face regions and avoid model overfitting (Shao et al., 2018). Furthermore, Park et al. (2017) applied the pruning algorithm and global maximum pooling on the GoogleNet model to retain the face location information, which greatly optimized the operation speed and accuracy. The development of residual networks solved the problem of vanishing gradients in neural networks and improved the feature extraction capability by adding identity mapping (He et al., 2016). Moreover, the generative adversarial network (Goodfellow et al., 2020) played a pivotal role in partially or entirely generating facial images that maintain context consistency and discriminating images.

In addition to optimizing foundation models and networks, loss functions have also been reformed. Sun et al. (2014) used a contrastive loss function for expression recognition. Schroff et al. (2015) utilized the triplet loss function to train expression recognition networks. It has the benefit of training smaller samples with lower variance, resulting in better performance in expression recognition. Wen et al. (2016) designed a center loss function with the aim of focusing on intra-class distribution uniformity to minimize intra-class variance. What is more, Cai et al. (2018) introduced the island loss

function that narrowed the intra-class variation while enlarging the inter-class variation to strengthen the discriminating power for deep features. Wang et al. (2018) employed a variation of the Softmax loss function by adding angular spacing and a cosine residual term. By maximizing the decision edges of the learned features in the feature space, this approach led to a significant decrease in intra-class distance and larger class spacing. More recently, the adaptive correlation-based loss function was developed to generate embedded feature vectors with high correlation for the within-class samples (Fard & Mahoor, 2022). However, existing deep learning methods have the weakness of poor performance because they are easily disturbed by the loss of information during feature propagation and by irrelevant features during feature extraction.

2.3. Research questions

This study aimed to address the limitation of existing approaches to automatic facial expression recognition by creating a new deep learning model to improve the recognition of facial expressions of students in online learning environments. To foster effective interaction and communication in online learning environments, it is important to provide AI-based facilities that help teachers to detect students' emotional states and give timely feedback. This study aimed to answer the following research questions:

RQ1: How to design a new deep learning model to improve the recognition of students' facial expressions in an online environment?

RQ2: Does the designed model outperform other state-of-art models in recognizing students' facial expressions in an online environment?

3. Material and method

3.1. Learning environment

This study designed a new model to improve the recognition of facial expressions. The performance of the model was evaluated by using the dataset containing video clips of facial expressions of students in an online session as the input to the model. During the online session, students were asked to work individually to complete five programming tasks using the Java language. The user interface used by students was presented in the form of a simple program editor, allowing students to compile and execute the programs they wrote. Students could also run unit tests to verify the correctness of their solutions. In order to induce students to produce different facial expressions, the online learning system periodically generated interference behaviors, such as automatically adding or deleting some characters during the tasks.

3.2. Dataset

The source of the dataset in this research was provided by the Gdansk University of Technology DevEmo dataset (Manikowska et al., 2023). The dataset contains video clips of facial expressions of students in the online session. The participants were 212 students majoring in computer science. The video clips of 33 students completing all five tasks were selected. The selected dataset contained 217 video clips, each of which was annotated and labeled as one of five primary emotional expressions including happiness,

surprise, anger, confusion, and neutral; no videos were labeled as fear or disgust. A summary of the dataset is presented in Table 1.

Table 1
Dataset content description

Attribute	Value
Number of Video Clips	271
Number of Participants	33
Gender	28 males, 5 females
Age	19-22
Average video duration	3 Seconds
Label	Anger, Surprise, Happiness, Confusion, Neutral

3.3. Data pre-processing

The average duration of each video clip in the dataset is 3 seconds, and each second of one video clip contains 30 frames. This study utilized OpenCV to segment each video to obtain images, with a segmentation interval of 15 frames, forming an image dataset. In order to reduce the interference of other parts of the human body on the recognition results, human face detection and cropping were performed on images in the dataset. In addition, all images were annotated with different facial expression labels. Besides, we operated the data augmentations operations about changing the brightness of images and rotating images by 45° to enlarge the dataset. Data normalization was also applied to the image dataset. Moreover, all images were shuffled, and the image dataset was split into the training set, validation set and test set according to the ratio of 6:2:2. Additionally, we resized all images to a uniform size of 128×128 .

3.4. Proposed model for facial expression recognition

This study proposes a deep learning approach by improving DenseNet and integrating a hybrid attention mechanism to automatically recognize the facial expressions of students. DenseNet has the peculiarity of allowing features to be reused to obtain better performance. What is more, the attention mechanism can reduce the interference of non-key information, allowing the network to focus on valuable information. Therefore, this study modified the structure of the DenseNet structural framework and added the convolutional block attention module (CBAM) before the global average pooling layer. CBAM is a lightweight attention mechanism module with the capability to improve the feature extraction of our model.

3.4.1. DenseNet

The DenseNet is a type of convolutional neural network (Huang et al., 2018). It creates dense connections between layers through dense blocks to enhance feature propagation. Thus, neural networks can alleviate the vanishing gradient problem caused by layers of networks being too deep. The initial DenseNet mainly consists of four dense blocks and three transition layers.

The DenseNet differs from other general convolutional neural networks. The unique dense block of DenseNet allows the model to no longer rely on the feature vector

output from the last layer as the sole basis for classification. Furthermore, each layer of the dense block utilizes dense connections between layers to receive additional input from all preceding layers. As shown in Fig. 1, X_0 is the input feature vector of convolution block 1 (conv block 1). X_0 is transformed to the output feature vector X_1 through H , a nonlinear transformation function. H includes batch normalization (BN), rectified linear unit (ReLU) and convolution. Therefore, the end output feature vector of conv block 1 is $[X_0, X_1]$ due to the structure of the dense block. The output feature vector of the n th convolution block in the dense block can be expressed as:

$$F_n = [X_0, X_1, \dots, X_{n-1}, H_n([X_0, X_1, \dots, X_{n-1})] \tag{1}$$

The structure with dense connections is beneficial for feature propagation, which is conducive to improving feature extraction. It also reduces the number of parameters in the model because the input of each layer contains feature information of the preceding layers. Therefore, a few feature maps need to be extracted when extracting the features of each layer.

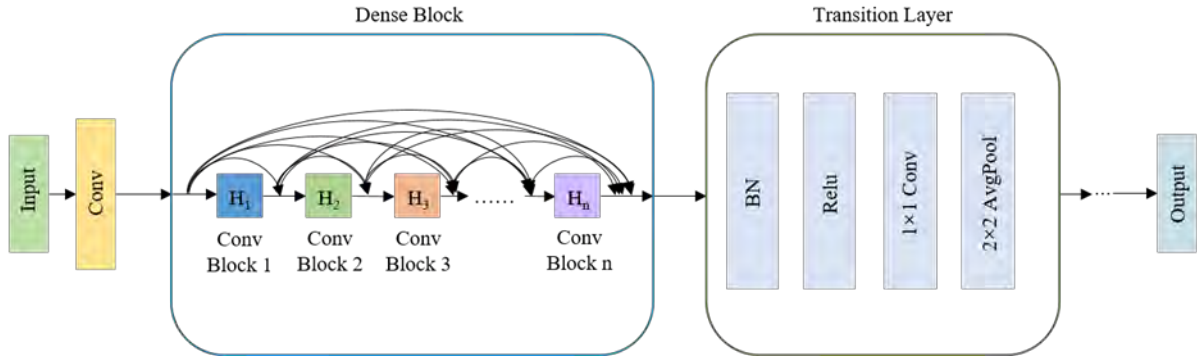


Fig. 1. Structure of DenseNet

The transition layer is mainly made up of one 1×1 convolution layer and one 2×2 average pooling layer. It connects the adjacent dense blocks. Moreover, it reduces the number of channels to control the complexity of the model.

3.4.2. Attention mechanism

The attentional mechanism is a technique in neural networks that mimics cognitive attention. Attention modules emphasize crucial target regions in the visual range and disregard irrelevant information within images. The attention mechanism improves the structure of neural networks and is widely applied in deep learning tasks. When processing images with neural networks, it is vital to allow networks to pay attention to the essential target regions adaptively, and the attention module is an approach to achieving adaptive attention to the target areas.

Attention modules can be divided into spatial attention modules (SAM) (Almahairi et al., 2016), channel attention modules (CAM), and hybrid attention modules. Firstly, SAM aims at finding the spatial information in the original image, transforming it into another space and retaining the key information, weighting the output for each location, focusing on the specific target region, and improving the feature representation of the target region. Secondly, CAM can assign higher weight values to important

channels and suppress useless channels. Thirdly, the hybrid attention module is a combination of SAM and CAM (see Fig. 2), and CBAM adopted in this study is a representative module of hybrid attention modules.

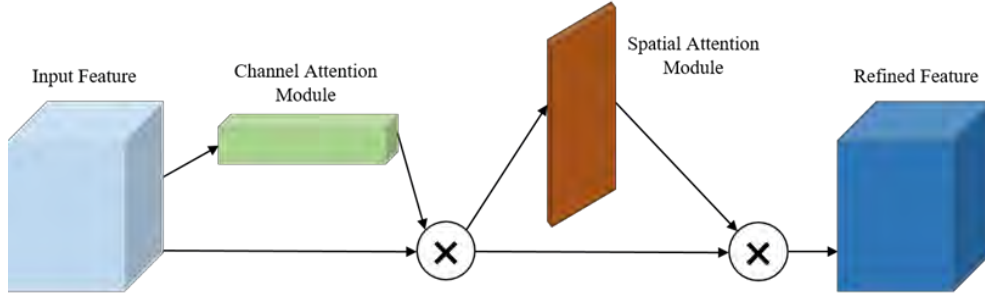


Fig. 2. Structure of CBAM

CAM typically employs weights to give more emphasis to the relevant area, with higher weights indicating a greater degree of attention. When recognizing images, it is difficult to avoid the effects of image rotation, distortion, and scale changes in image recognition tasks, but SAM can well preserve important information in images and reduce the impact of images from operations such as transformations. By mixing cross-channel and spatial information for feature information extraction, CBAM has the advantage of more accurate feature extraction.

3.4.3. Integrating improved DenseNet with CBAM

In this study, a novel facial expression recognition model DenseNetX-CBAM was developed by improving the DenseNet and introducing the CBAM to neural network structure. The whole structure of the proposed model is shown in Fig. 3. In the input section of the model, the input is an RGB three-channel facial expression image with a pixel size of 128×128 . The feature extraction network uses 7×7 and 3×3 convolution kernels for initial feature extraction and dimensionality reduction of the input image to obtain a 32×32 feature map of sixty-four channels, and four dense blocks and three transition layers are alternately connected to further optimize the extraction of information and improve the reusability of features.

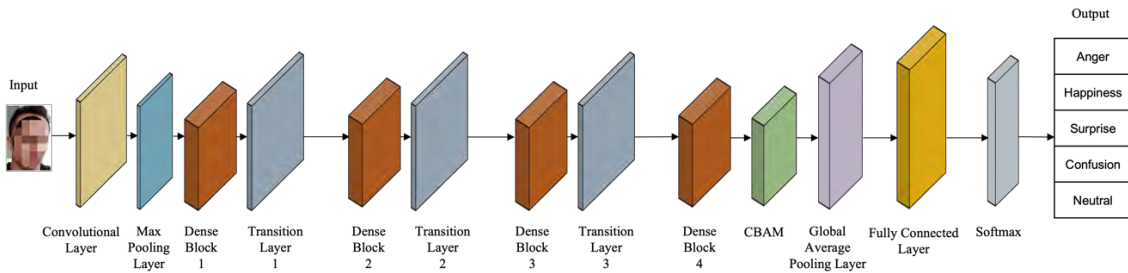


Fig. 3. Overall structure of DenseNetX-CBAM

The number of convolutional blocks in the four dense blocks is set to 3, 6, 12 and 8, respectively. Convolutional blocks extract features by using 1×1 and 3×3 convolutional kernels with a growth rate of 16. In the transition layer, the compression

rate of parameters related to channel downscaling is set to 0.5. The size of the feature map changes through three transition layers from 32×32 to 4×4 .

In this study, the CBAM module is added following dense block 4 which strengthens the feature learning of important channels and spaces using CAM and SAM to improve the accuracy of the model. First, in the CAM section of CBAM, a $4 \times 4 \times 262$ facial expression feature map output by dense block 4 undergoes maximum pooling and average pooling operations on their height and width to obtain two $1 \times 1 \times 262$ facial expression feature maps. The two feature maps are transported into the two-layer neural network to obtain two $1 \times 1 \times 262$ facial expression feature vectors. Besides, the input feature map of SAM is generated by adding the above two feature vectors and multiplying the face expression feature map with the channel attention weight. Second, in the SAM section of CBAM, the maximum pooling and average pooling operations are performed on the input feature maps to get two $4 \times 4 \times 1$ facial expression feature maps. The channels of the two feature graphs are concatenated, and the dimensions are reduced by applying a 3×3 convolution kernel. After activation with the sigmoid function, the spatial attention weight is acquired. Finally, the optimized features of facial expressions are obtained by multiplying the facial expression feature map with the spatial attention weight. Therefore, by using the CBAM, attention maps can be generated for the extracted feature information in the channel dimension and spatial dimension in turn. This is conducive to enhancing the channel and spatial dimension information of locally important features, thus accelerating the convergence speed of the model and increasing the expression recognition rate.

3.5. Experimental setting

In this study, a computer with Intel i7-9700 CPU, 32GB RAM and GeForce 2080 GPU was utilized for the model training, validating and testing. The model was trained for 50 epochs, using the adaptive moment estimation (Adam) optimizer with an initial learning rate of 0.0001 and a batch size of 8. Moreover, the loss function employed the cross-entropy loss.

3.6. Performance metrics for model evaluation

The performance of the proposed model was tested in terms of accuracy, macro F1-score, parameters, and confusion matrix.

3.6.1. Accuracy

Accuracy is the main performance metric used to evaluate classification models. Higher accuracy indicates better performance of the model. Accuracy is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where *TP* is true positive prediction; *TN* is true negative prediction; *FP* is false positive prediction; and *FN* is false negative prediction

3.6.2. Macro F1-score

The F1 score is an evaluation index in deep learning models. It assesses the precision and recall of the model at the same time and can be regarded as a harmonic average of the

precision and recall of the model. When addressing multiple classification tasks, the macro F1-score is often used to measure the performance of the model. The number of categories is set to N . Therefore, the macro F1-score is the mean of F1 scores for all categories, and it is calculated as follows:

$$F1 \text{ score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3)$$

$$\text{Macro F1-score} = \frac{\sum_1^N (F1 \text{ score})}{N} \quad (4)$$

3.6.3. Parameters

In deep learning models, parameters represent the total number of parameters contained in the model. It is an important index to measure deep learning models, corresponding to the consumption of computer memory resources. The calculation formulas for the parameters of the convolutional layer and for the fully connected layer of the model are as follows:

$$\text{Parameter}_1 = (K_h \times K_w \times C_i + 1) \times C_o \quad (5)$$

$$\text{Parameter}_2 = (N_i + 1) \times N_o \quad (6)$$

where C_i is the input channel, K_h is the height of the convolution kernel, K_w is the width of the convolution kernel, "+1" means the bias, C_o is the output channel, N_i is the number of input nodes, and N_o is the number of output nodes

3.6.4. Confusion matrix

To better evaluate the performance of the model proposed in this study and observe the change in performance, the confusion matrix is used to compare and analyze the recognition rate of each type of expression. The horizontal coordinates of the confusion matrix represent the predictive labels, the vertical coordinates represent the true labels, and the values of the coordinate points where the predictive labels intersect with the true labels represent the accuracy rates of the corresponding labels. Therefore, the values on the main diagonal in the confusion matrix represent the accuracy rate of the corresponding expressions, and the rest of the area represents the similarity with other expressions. Furthermore, the shades of blue are used to represent the accuracy rate and the darker color means the higher accuracy.

4. Results

4.1. Comparison of different models

The proposed DenseNetX-CBAM model was compared with the following five state-of-the-art deep learning models.

- GoogLeNet (Szegedy et al., 2014): GoogLeNet model has the inception module to aggregate visual information at different sizes for facilitating feature extraction. It shows promising performance in image classification.
- MobileNetv2 (Sandler et al., 2018): MobileNetv2 is a mobile deep learning model. It is based on the inverted residual structure, and the intermediate

expansion layer of the network filters the feature by using nonlinear lightweight depth convolutions.

- VGG16 (Simonyan & Zisserman, 2015): VGG16 is a deep learning model widely used in large-scale image classification and recognition tasks.
- ResNet50 (He et al., 2016): ResNet50 uses the residual network structure to effectively alleviate the problem of model degradation, thereby achieving a deeper network structure design.
- EfficientNet (Tan & Le, 2019): EfficientNet uses a series of fixed scaling factors to uniformly scale the network dimension, demonstrating excellent accuracy and efficiency.

The performance of all six models were tested using the aforementioned dataset as the input to the models. Table 2 presents the performance all six model in terms of accuracy, macro F1-score, and parameters.

Table 2

Accuracy, macro F1-score, and parameters of the six models

Model	Accuracy	Macro F1-Score	Parameters
GoogLeNet	0.740	0.716	5998172
MobileNetv2	0.752	0.735	3447620
VGG16	0.758	0.742	25501932
ResNet50	0.784	0.769	138357642
EfficientNet	0.816	0.792	1695328
DenseNetX-CBAM	0.856	0.837	705195

Fig. 4 presents the performance of these models in terms of confusion matrix. The results show that DenseNetX-CBAM outperforms the other five state-of-the-art models in all criteria.

4.2. Ablation study results

An ablation study was conducted on the same dataset to investigate how the models integrating various attention mechanism modules impact the performance for recognizing facial expressions. The implementation steps are as follows:

- 1) Baseline: This model uses the initial DenseNet framework without any improvements or the addition of attention modules.
- 2) DenseNetX: This model improves upon the DenseNet framework but does not include any attention modules.
- 3) DenseNet-CBAM: This model adds both SAM and CAM to the baseline model.
- 4) DenseNetX-SAM: This model adds SAM to the DenseNetX framework without CAM.
- 5) DenseNetX-CAM: This model adds CAM to the improved DenseNetX framework.
- 6) DenseNetX-CBAM: This model adds both SAM and CAM to the DenseNetX framework.

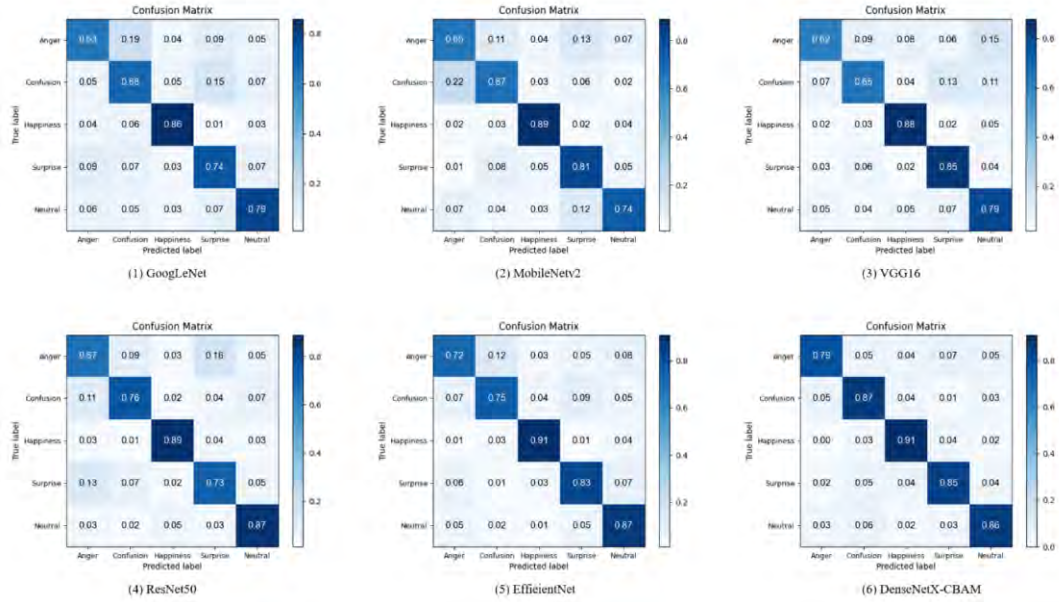


Fig. 4. Confusion matrix of the six models

Table 3 reports the Ablation study results in terms of accuracy, macro F1-score and parameters of integrating different attention mechanism modules based on DenseNet and DenseNetX. The results show the proposed DenseNetX-CBAM model based on the combination of DenseNetX and CBAM is more effective for increasing the model recognition rate.

Table 3
Ablation study

Model	Accuracy	Macro F1-Score	Parameters
DenseNet (Baseline)	0.782	0.758	7045704
DenseNetX	0.794	0.775	690162
DenseNet-CBAM	0.806	0.792	7083524
DenseNetX-SAM	0.825	0.814	704974
DenseNetX-CAM	0.831	0.819	705062
DenseNetX-CBAM	0.856	0.837	705195

The optimization of DenseNet led to the evolution of DenseNetX, where DenseNetX was progressively better than the previous one. Therefore, we compared the performance of the optimized model with the initial model. Fig. 5 presents the results of the DenseNet-CBAM model and DenseNetX-CBAM model for recognizing each class of facial expressions in the dataset. Confusion matrices suggest that the combination of DenseNetX and CBAM is more effective in improving the performance of facial expression recognition.

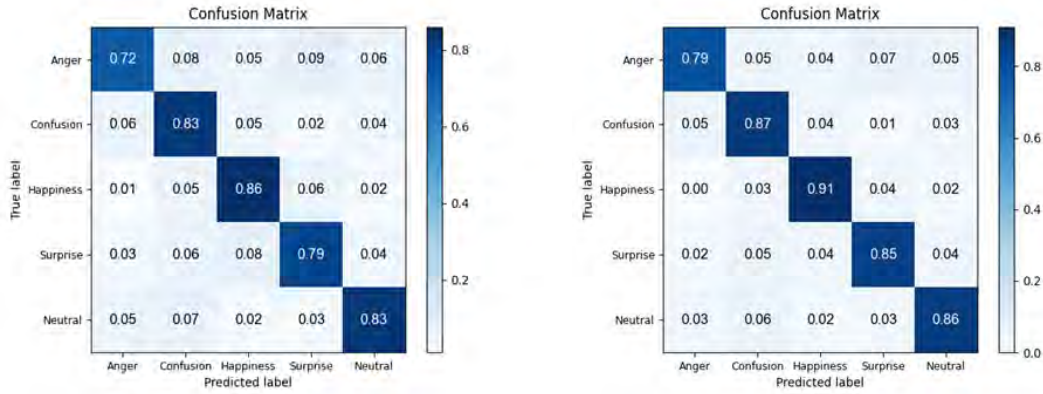


Fig. 5. Confusion matrix of DenseNet-CBAM (Left) and DenseNetX-CBAM (Right)

5. Discussion

The findings of the study are discussed as follows.

5.1. How to design a new deep learning model to improve the recognition of students' facial expressions in an online environment?

This study proposed a novel deep learning model called DenseNetX-CBAM to improve the recognition of students' facial expressions in an online learning environment. The model was designed by combining the superior capabilities of DenseNet and CBAM. Firstly, the structure of DenseNet was utilized and ameliorated to reduce the number of unnecessary parameters and strengthen the reuse of expression features between networks. It has an adequate improvement over traditional machine learning methods such as KNN and SVM (Patil & Patil, 2022). Secondly, CBAM was integrated into the model, which can help the networks to focus on effective information by assigning different weights to the facial expression features of different channels. CBAM is a hybrid attention module that combines SAM and CAM. SAM can highlight important spatial regions while suppressing less informative regions; CAM can help the model to focus on the most important channels to enhance feature representations. In addition, CBAM is a lightweight module with a simple structure, which makes it a convenient and efficient modification of deep learning models.

5.2. Does the designed model outperform other state-of-art models in recognizing students' facial expressions in an online environment?

By comparing the performance of DenseNetX-CBAM with other state-of-the-art models, this study reveals that DenseNetX-CBAM achieved the best performance in recognizing students' facial expression in an online learning environment. DenseNetX-CBAM outperformed other models in terms of recognition accuracy, precision and recall of the model, and the number of parameters. The number of parameters of DenseNetX-CBAM

was far less than that of other models, indicating that the proposed model can effectively save computing resources.

To further assess the model, we investigated the impact of model components on model performance through the ablation study. The results show that the DenseNetX integrated with CBAM (combination of CAM and SAM) has a greater capability than the DenseNetX integrated with merely CAM or SAM. Moreover, it can be seen by observing the confusion matrix that DenseNetX-CBAM has a significant improvement in the accuracy of recognizing five facial expressions on the dataset compared to DenseNet-CBAM. For emotions that are relatively difficult to recognize, such as anger, the recognition accuracy increased from 0.72 to 0.79. For easily recognizable expressions, such as happiness and confusion, the recognition accuracy increased from 0.86 to 0.91 for happiness and 0.83 to 0.87 for confusion. These findings suggest that the performance of DenseNetX-CBAM is far superior to other models.

5.3. Limitations and future work

This study has several limitations. First, a dataset of facial expressions of students learning Java programming in an online learning environment may not fully represent the facial expressions of students in other online courses. Second, the recognition rate of the facial expression recognition models may decrease in complex environments like those with varying levels of lighting. Such environments can limit the efficacy of the expression recognition system. Third, the dataset of video clips used in this study contained five primary emotional expressions of students learn in an online learning environment including happiness, surprise, anger, confusion, and neutral, not including other emotional expressions (e.g., disgust and fear). More empirical studies are needed to investigate the possibility of including other expressions in analysis.

6. Conclusion

To foster effective interaction and communication in online learning environments, it is important to assess students' emotional states and give timely feedback, which can be supported by AI technology. This study proposed a novel deep learning model to improve recognition of students' facial expressions in an online learning environment. The proposed model utilized a variant of densely connected convolutional networks (DenseNet) to reduce the number of unnecessary parameters and strengthen the reuse of expression features between networks. Meanwhile, a convolutional block attention module (CBAM) was integrated into the model, which can help the networks to focus on effective information by assigning different weights to the facial expression features of different channels. By testing with 217 video clips of 33 students in an online course, the proposed model has shown promising effects in improving the accuracy of facial expression recognition. The proposed approach has a high potential to help teachers to accurately recognize students' emotional states and provide real-time adjustment in online teaching and learning environments.

Author Statement

The authors declare that there is no conflict of interest.

Acknowledgements

The authors would thank Professor Haijing Jiang for his valuable guidance and support for this study.

ORCID

Heng Zhang  <https://orcid.org/0009-0002-8451-9952>

Minhong Wang  <https://orcid.org/0000-0002-1084-6814>

References

- Alawamleh, M., Al-Twait, L. M., & Al-Saht, G. R. (2020). The effect of online learning on communication between instructors and students during Covid-19 pandemic. *Asian Education and Development Studies*, 11(2), 380–400. <https://doi.org/10.1108/AEDS-06-2020-0131>
- Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., & Courville, A. (2016). Dynamic capacity networks. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 2549–2558). Retrieved from <https://proceedings.mlr.press/v48/almahairi16.html>
- Ashwin, T. S., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, 25(2), 1387–1415. <https://doi.org/10.1007/s10639-019-10004-6>
- Bender, D. M., & Vredevoogd, J. D. (2006). Using online education technologies to support studio instruction. *Educational Technology & Society*, 9(4), 114–122.
- Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., & Tong, Y. (2018). Island loss for learning discriminative features in facial expression recognition. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 302–309). <https://doi.org/10.1109/FG.2018.00051>
- Cárdenas-Robledo, L. A., & Peña-Ayala, A. (2019). A holistic self-regulated learning model: A proposal and application in ubiquitous-learning. *Expert Systems with Applications*, 123, 299–314. <https://doi.org/10.1016/j.eswa.2019.01.007>
- Chiu, T. K. F. (2022). Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education*, 54(Sup1), S14–S30. <https://doi.org/10.1080/15391523.2021.1891998>
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893). <https://doi.org/10.1109/CVPR.2005.177>
- Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems*, 49(1), 5–22. <https://doi.org/10.1177/0047239520934018>
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129. <https://doi.org/10.1037/h0030377>
- Fard, A. P., & Mahoor, M. H. (2022). Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10, 26756–26768. <https://doi.org/10.1109/ACCESS.2022.3156598>

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144. <https://doi.org/10.1145/3422622>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Retrieved from https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4700–4708).
- Kazemitabar, M., Lajoie, S. P., & Doleck, T. (2019). Examining changes in medical students' emotion regulation in an online PBL session. *Knowledge Management & E-Learning*, 11(2), 129–157. <https://doi.org/10.34105/j.kmel.2019.11.008>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Retrieved from https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
- Lasri, I., Riadsolh, A., & Elbelkacemi, M. (2023). Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. *Education and Information Technologies*, 28(4), 4069–4092. <https://doi.org/10.1007/s10639-022-11370-4>
- Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
- Lindeberg, T. (2012). Scale invariant feature transform. *Scholarpedia*, 7(5): 10491. <https://doi.org/10.4249/scholarpedia.10491>
- Manikowska, M., Sadowski, D., Sowinski, A., & Wrobel, M. R. (2023). DevEmo—Software developers' facial expression dataset. *Applied Sciences*, 13(6): 3839. <https://doi.org/10.3390/app13063839>
- Mehrabian, A. (2008). Communication without words. In A. Mehrabian (Ed.), *Communication Theory* (2nd ed.) (pp. 193–200). Routledge.
- Mukhtar, K., Javed, K., Arooj, M., & Sethi, A. (2020). Advantages, limitations and recommendations for online learning during COVID-19 pandemic era. *Pakistan Journal of Medical Sciences*, 36(COVID19-S4), S27–S31. <https://doi.org/10.12669/pjms.36.COVID19-S4.2785>
- Nayak, H. D., & Sarvaiya, A. K. (2022). Facial expression recognition based on feature enhancement and improved alexnet. *ICTACT Journal on Soft Computing*, 12(3), 2589–2600.
- Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>
- Park, J., Li, S., Wen, W., Tang, P. T. P., Li, H., Chen, Y., & Dubey, P. (2017). Faster CNNs with direct sparse convolutions and guided pruning. *arXiv preprint arXiv:1608.01409*. <https://doi.org/10.48550/arXiv.1608.01409>
- Patil, S., & Patil, Y. M. (2022). Face expression recognition using SVM and KNN classifier with HOG features. In B. Iyer, T. Crick, & S.-L. Peng (Eds.), *Applied Computational Technologies* (pp. 416–424). Springer Nature. https://doi.org/10.1007/978-981-19-2719-5_39
- Revina, I. M., & Emmanuel, W. R. S. (2021). A survey on human face expression

- recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 619–628. <https://doi.org/10.1016/j.jksuci.2018.09.002>
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). <https://doi.org/10.1109/CVPR.2018.00474>
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815–823). <https://doi.org/10.1109/CVPR.2015.7298682>
- Shao, J., Qu, C., Li, J., & Peng, S. (2018). A lightweight convolutional neural network based on visual attention for SAR image target classification. *Sensors*, 18(9): 3039. <https://doi.org/10.3390/s18093039>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. <http://arxiv.org/abs/1409.1556>
- Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Retrieved from <https://proceedings.neurips.cc/paper/2014/hash/e5e63da79fcd2bebbd7cb8b1c1d0274-Abstract.html>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*. <https://doi.org/10.48550/arXiv.1409.4842>
- Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). Retrieved from <https://proceedings.mlr.press/v97/tan19a.html>
- Tang, T., Abuhmaid, A. M., Olaimat, M., Oudat, D. M., Aldhaeabi, M., & Bamanger, E. (2023). Efficiency of flipped classroom with online-based teaching under COVID-19. *Interactive Learning Environments*, 31(2), 1077–1088. <https://doi.org/10.1080/10494820.2020.1817761>
- Tonguç, G., & Ozaydin Ozkara, B. (2020). Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education*, 148, 103797. <https://doi.org/10.1016/j.compedu.2019.103797>
- Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7), 926–930. <https://doi.org/10.1109/LSP.2018.2822810>
- Wei, H.-C., & Chou, C. (2020). Online learning performance and satisfaction: Do perceptions and readiness matter? *Distance Education*, 41(1), 48–69. <https://doi.org/10.1080/01587919.2020.1724768>
- Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A Discriminative feature learning approach for deep face recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 499–515). Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_31
- Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G.-J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human-Computer Interaction*, 35(4/5), 356–373. <https://doi.org/10.1080/10447318.2018.1543084>
- Wong, R. (2023). When no one can go to school: Does online learning meet students' basic learning needs? *Interactive Learning Environments*, 31(1), 434–450. <https://doi.org/10.1080/10494820.2020.1789672>

- Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the 2018 European Conference on Computer Vision (ECCV)* (pp. 3–19). Retrieved from https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html
- Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning*, 15(2), 133–152. <https://doi.org/10.34105/j.kmel.2023.15.008>
- Zou, D., Xie, H., & Wang, F. L. (2018). Future trends and research issues of technology-enhanced language learning: A technological perspective. *Knowledge Management & E-Learning*, 10(4), 426–440. <https://doi.org/10.34105/j.kmel.2018.10.026>