Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230136

A robust VGG network combined with Denoising Autoencoder module for human emotion recognition

Xiaotian Li

School of Computing and Information Systems, The University of Melbourne, North Melbourne, 3052, Australia

xiaotian.li5@student.unimelb.edu.au

Abstract. Human emotion can be divided into multiple categories, which makes it possible to recognize emotions automatically. One critical approach for automated emotion recognition is applying the convolutional neural network to classify emotions on human expression images, but the performance decreases if input distortions occur. This paper introduced a hybrid neural network architecture to make the automated emotion recognition robust towards distorted input images and perform similarly to prediction on clean images. This hybrid neural network combines the Denoise Autoencoder (DAE) network with the Visual Geometry Group (VGG) network. Multiple standalone VGG and Hybrid network experiments were conducted with the control variables method. FER-2013 data set from Kaggle was used as the experimental data set. Distorted input images were generated by adding random noise to clean images. As a result, the research raised a valid hybrid network architecture. The hybrid network improved the emotion classification accuracy on the distorted data set from 16.70% to 57.73%, and the accuracy is similar to the classification result on the clean data set.

Keywords: Automated emotion recognition, Denoise autoencoder, VGG network.

1. Introduction

With the rapid development of technologies, the demand for automated emotion recognition is increasing and being employed extensively in human-machine interaction, financial technologies, industry, etc. [1]. To achieve emotion recognition, an approach to classify emotions is vital. According to the discrete emotion theory, which is widely accepted, emotions can be categorized into several core classes. Silvan Tomkins' research concluded that limited core effects produce human emotions, including anger, fear, enjoyment, surprise, and disgust [2]. Carroll Izard from the University of Delaware revealed ten primary emotions that can be labeled: interest, joy, surprise, sadness, anger, disgust, contempt, fear, shame, and guilt [3]. Moreover, a more structured emotion model with primary and secondary emotions was introduced by Robert Plutchik [4].

Since human emotions can be divided into discrete classes, it is possible to achieve automated emotion recognition by performing classification tasks. To achieve emotion classification, an accurate and effective approach to evaluate emotions is essential. *Electroencephalography* (EEG) method can directly detect emotional-related signals of the nervous system to reveal the human emotional state in live [5]. However, EEG methods have disadvantages, including low precision and inefficient signal

DOI: 10.54254/2755-2721/27/20230136

processing. Mara Mather indicated that *Heart Rate variability* (HRV) is not simply random but correlated with emotional responding [6]. Among correlational studies, facial expression becomes one of the most critical approaches in emotion recognition because facial expression delivers rich emotional information and reflects mental state [7].

The deep-learning-based technique is frequently used for emotion recognition on facial expressions. *Convolutional Neural Network* (CNN) is a suitable choice because of its high accuracy, good performance, and fast recognition [8]. The V*isual Geometry Group* (VGG) network was invented by Simonyan and Zisserman [9], which won first prize in ImageNet Large Scale Visual Recognition Competition. VGG network can be a reasonable choice for facial expression classification tasks due to its extraordinary performance in image and object classification tasks [10].

However, deep-learning-based emotion recognition is not always ideal in all conditions. The neural network requires facial expression images to classify and recognize facial expressions, so the quality of input images can influence the classification performance. Computers and people execute classification tasks equally well when the images are not deformed, but human performance tends to be more robust when the images are distorted [11]. Image distortions usually accumulate in extreme weather conditions, cosmic radiation, erroneous input devices, transmission loss, and storage failure [12]. Also, Image distortions take several common forms: random brightness color variation, blur, and data loss. Random image distortions can disguise critical features of the input image. Since Neural Networks for emotion recognition are trained by fixed facial expression data sets, randomly distorted image input can deduct classification accuracy of the emotion recognition procedure. As a result, image distortions bring challenges to automated emotion recognition with facial expressions.

This paper introduces a pre-processing image scheme into an existing convolutional neural network to deduct input image noise in advance. The neural network will gain consistent classification accuracy in clean and noisy facial expression images by denoising input images. This paper uses VGG based network to achieve image-based emotion classification [9], and uses a DAE network to process distorted input data into clean data [13]. As an extension of a normal autoencoder, the DAE network can capture essential features of an image and reconstruct distorted images by ignoring noises. A hybrid model that combines the VGG model and DAE model is created, which is supposed to make the emotion classification robust towards noisy input images. To validate the performance of the hybrid model, controlled trials are performed. The trial is divided into two groups. The first group uses the VGG network to predict clean image data and distorted data, while the other uses the hybrid VGG and DAE network. As a result, the standalone VGG network got 58.78% accuracy on clean image input and 16.70% accuracy on distorted data input. In comparison, the hybrid VGG + DAE network got 58.74% accuracy on clean input and 57.73% on distorted data.

The paper is organized as follows: Section 2 discusses the input data set and input data pre-process for distorted data. Section 3 explains detailed neural network architectures and configurations for experiments. The experiment outcome will be discussed in section 4. Section 5 focuses on research conclusions and future works.

2. Method

2.1. Input data set and pre-processing

The project used the FER-2013 dataset from Kaggle because the dataset has automatically registered human faces with a central location, and each face occupies a similar space [14], making it convenient for data pre-handling. The dataset contains grayscale images of human facial expressions, and all images are converted into 48x48 pixels in size. All 35, 887 sample images have been divided into train data (28, 709 images) and test data (3, 589 images). Also, the train data categorized into 7 primary emotions: Neutral (4, 965), disgust (436), happy (7, 215), fear (4, 097), surprise (3, 171), angry (3, 995) and sad (4, 830).

For dataset pre-processing, this study cloned original sample data. Then the study generated random noise on the cloned data to conduct check tests between the VGG and hybrid networks. The

DOI: 10.54254/2755-2721/27/20230136

random noise is generated by randomly converting the gray level in the 48x48 pixel zone to simulate data distortions caused by erroneous input devices or extreme operating conditions.

Moreover, to improve the training performance of the neural network, data augmentation is performed for the training dataset. Data augmentation can avoid overfitting caused by irrelevant features so that neural networks can perform remarkably well [15]. The data augmentation in the study includes: 1) Image gray level rescaling: Convert pixels' gray level from [0, 255] range into [0, 1]. By converting feature data into a common range, the data bias caused different numerical contributions will be minimized, which is a practical statistical method in machine learning [16]; 2) Image rotation: Randomly rotate images from 0 degree to 20 degree, because image angle is not a valuable factor for facial expressions. Performing rotations can make classification robust towards rotated images; 3) Horizontal flip: Randomly flipping some images because mirror images of human facial expressions are not vital for emotion recognition.

Also, the training data of the study were split into a validation set with a rate of 20%. To sum up, all training data is divided into 7 emotion classes, 22,968 of them belong to the training set, and 5,741 images form up the validation set. As for testing data, 7,178 images are used to calculate the model's accuracy.

2.2. Proposed model

2.2.1. VGG-based model. In the study, the emotion recognition neural network was built on the VGG model because the VGG model tends to perform well in image classification tasks [10]. Extra layers are attached to the original VGG model to gain stable experiment results. The network consists of 3 dense layers, 13 convolutional layers, and some other layers, including batch normalization, max-pooling, and dropout layers. The whole network contains 33,644,103 trainable parameters and 19,328 non-trainable parameters. Input grayscale image data with 48x48 pixel size is ingested by the model, producing prediction results in 7 classes corresponding to 7 emotions. The model structure and essential parameters are shown in Figure 1.

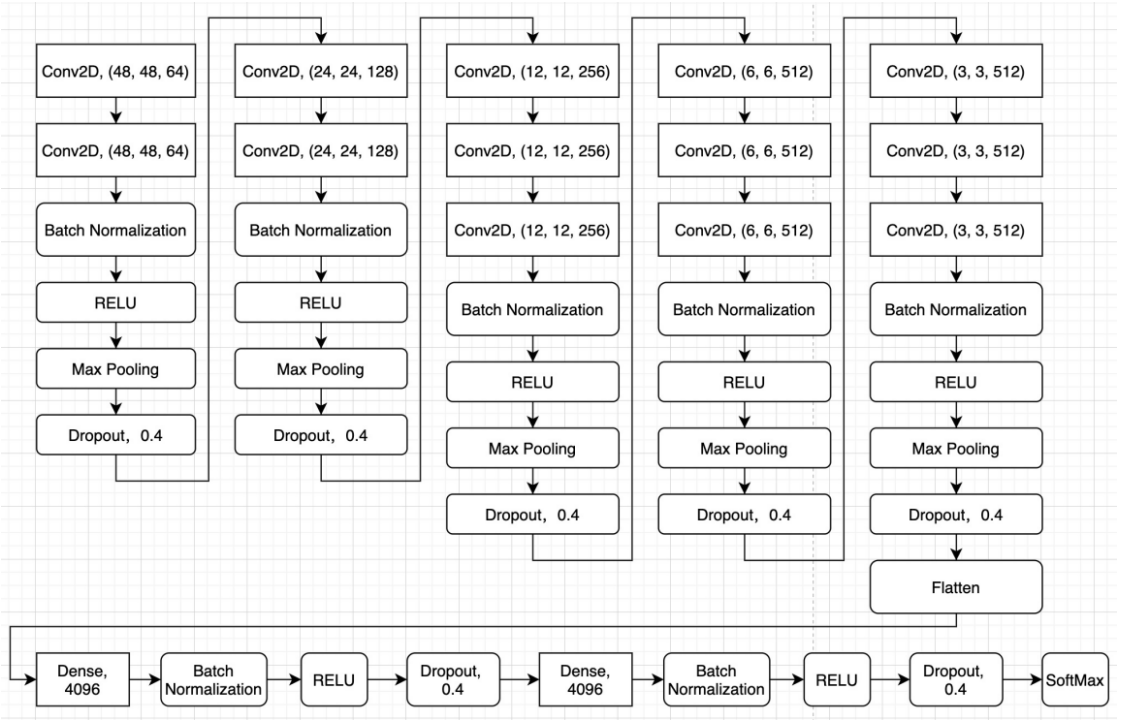

Figure 1. The structure of VGG based model in the study.

DOI: 10.54254/2755-2721/27/20230136

As is shown in Figure 1, all dropout layers have 0.4 as the dropout rate. In convolution 2D layers (Conv2D), 3x3 kernels, 1x1 strides, and zeros padding are applied. Moreover, the input size of Conv2D is diverse from 48x48 pixels to 3x3 pixels, while filter numbers are from 64 to 512. Full connect dense layers have 4096 neural units. In the output layer, a dense layer with 7 units and a Softmax activation function is applied to achieve the classification task. The result will be the code name of one emotion class from seven classes in total.

2.2.2. Denoise autoencoder model.

In the study, denoise autoencoder (DAE) was used for distorted image pre-processing because DAE is competent in image denoise performance [13], and DAE is also relatively easy to be implemented in the study. The architecture of the DAE contains 4 layers as the encoder and 5 layers as the decoder. The model has 5 convolutional layers, 2 max-pooling layers, and 2 up-sampling layers. The DAE network takes 48x48 grayscale distorted images as the input and generates denoised images with the sigmoid activation function at the output layer. The generated images will have less noisy pixels and positively affect subsequent emotion classification. The model structure and parameters are shown in Figure 2.

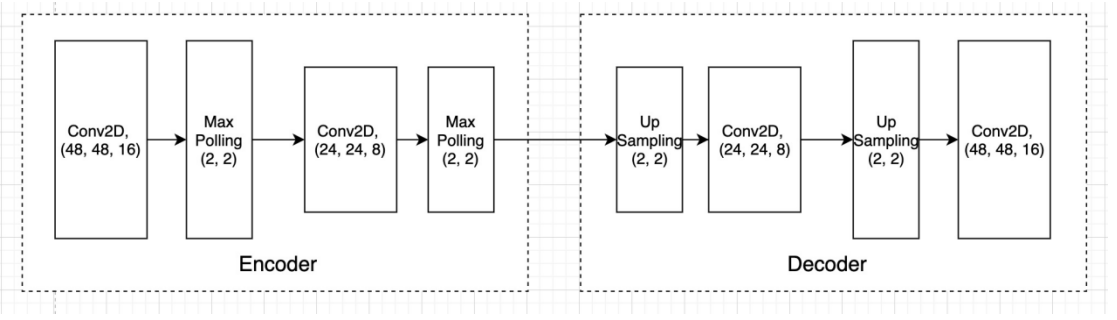

Figure 2. The structure of DAE model in the study.

As is shown in Figure 2, in Conv2D layers, 16 filters, and 24 filters are applied. All Conv2D layers use 3x3 kernel, 1x1 strides, zero paddings, and ReLU activation functions. For max polling layers and up-sampling layers, the pool size is set to 2x2. Conv2D layers with 1 filter and sigmoid activation function are configured in the output layer. The output will be the denoised image with the same shape as the original input image.

2.2.3. Hybrid model

The study introduced a hybrid neural network model with VGG and DAE network to make the emotion classification perform robust in both distorted and clean images. The study proposed two different hybrid schemes, shown in Figure 3 and Figure 4.

DOI: 10.54254/2755-2721/27/20230136

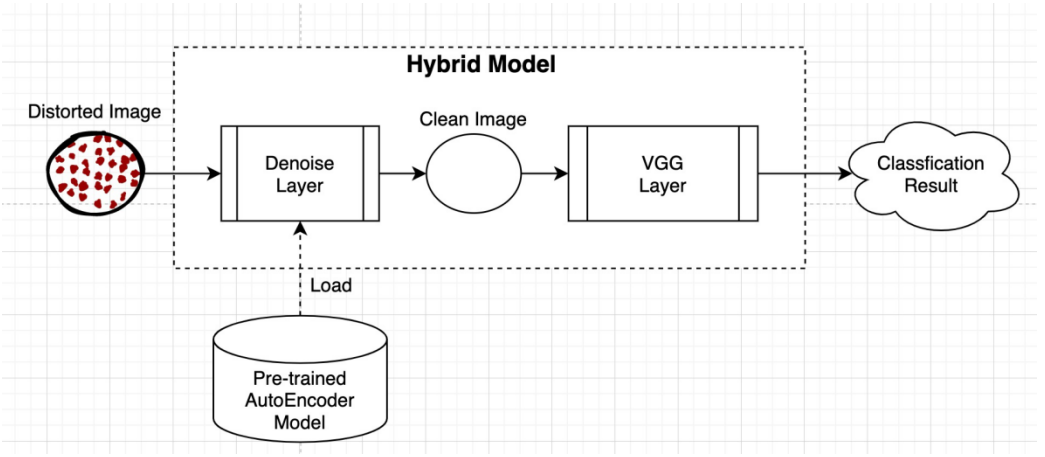

Figure 3. Hybrid model variant A.

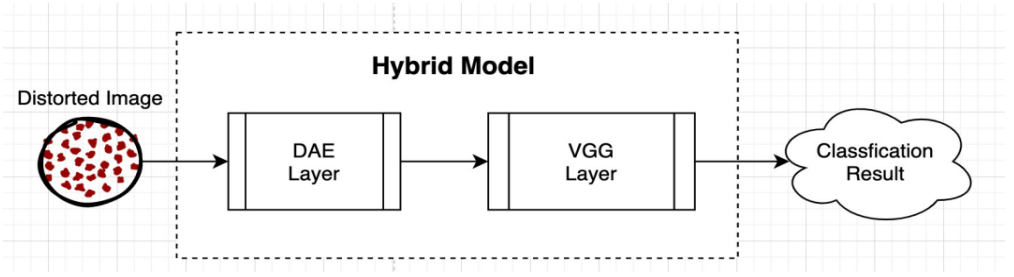

Figure 4. Hybrid model variant B.

In hybrid-model variant A, a pre-trained DAE model is imported in the denoise layer, and the denoise layer is frozen from training. Then the whole hybrid model will be trained and perform prediction later. In another scheme, as hybrid-model variant B shows, both DAE and VGG are treated as a whole. Both layers are trained together and then perform classification tasks. The reason for testing two variants of the hybrid-model in the project is to determine which one has a better performance. Both hybrid models have the same variables as configured in section 2.2.1 and section 2.2.2.

2.3. Implementation details

TensorFlow 2.5.0 is used in the experiment device with Python runtime version 3.8 on the Ubuntu 18.04 platform. As for GPU, NVIDIA RTX3090 with 24 GB memory accelerates the training procedure.

To control variables, identical training hyperparameters are defined in the experiments, as shown in Table 1.

Table 1. Common hyperparameters used in experiments.

Hyper-parameter	Value
Epoch number	150
Optimizer	Adam
Batch Size	16
Learning rate	0.05
Adam beta_1	0.9
Adam beta_2	0.99
Early Stopping patience	20
Early Stopping mode	Min
Reduce learning rate on Plateau factor	0.5

DOI: 10.54254/2755-2721/27/20230136

As for loss functions, the DAE model uses Binary Cross Entropy, while the VGG model uses Categorical Cross Entropy. Accuracy is used as the model evaluation metric.

3. Experimental results and discussion

3.1. Experiment overview

Multiple experiments were conducted to determine whether the hybrid model brings more accuracy than the standalone VGG model in emotion classification based on distorted input images.

Firstly, the DAE model was trained with a distorted image dataset as the input and a clean image dataset as the expected output. Then, the pre-trained model was exported for the upcoming experiment. After that, the standalone VGG-based model mentioned was trained, then predicted with clean images and noisy mages for comparison. Afterward, two different hybrid models were trained by the same input dataset and hyperparameters. The first hybrid model's DAE layer was imported from the previous pre-trained model, whereas the other hybrid model was not and trained together with the VGG layer in a whole.

The reasons for conducting these experiments are: 1) To prove that the hybrid model has better accuracy of distorted images than standalone VGG model does; 2) To make sure that the hybrid model should not have a major performance deduction in predicting clean images, compared to pure VGG one; 3) To prove that the hybrid model has consistent classification accuracy in both clean input images and noisy images.

3.2. Result of the DAE network

The DAE network was trained in the research with 150 epochs. The training procedure stopped in the last epoch with a loss value of 0.5528. Figure 5 illustrates the comparison between noisy input images and generated denoised images.

Figure 5. Denoising result of DAE network.

As shown in Figure 5, the first row is about the distorted input images with massive random noise, while the second is about the reconstructed images created by the DAE network. It can be concluded that the trained DAE model can effectively reconstruct distorted images into clean images. This model can effectively handle the random noise from erroneous input devices or extreme working conditions. Moreover, this trained model was exported into the persistent storage for further usage.

3.3. Result of the standalone VGG based model

A standalone VGG-based model was trained in the research. The training result, including loss value and accuracy value, is shown in Figure 6.

DOI: 10.54254/2755-2721/27/20230136

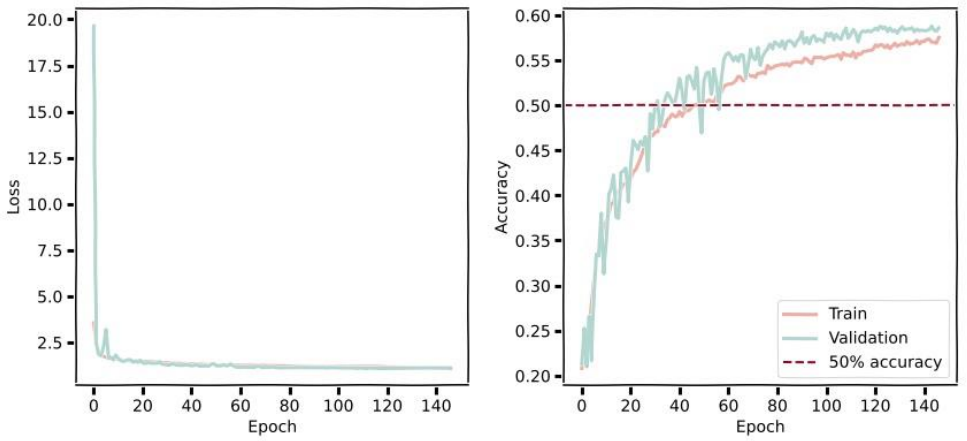

Figure 6. VGG-based model's Loss Curve and Accuracy Curve.

Because of the Early Stop configuration, the model training was stopped at epoch 147 with the loss value of 1.1361, accuracy of 57.55%. Two test image data sets were used to evaluate the trained VGG-based model. The result is shown in Table 2.

Table 2. Standalone VGG-based model evaluation result of 2 input image sets.

	Loss	Accuracy
Clean input images	1.0874	58.78%
Distorted input images	3.2104	16.70%

According to the model evaluation result, VGG based network performs better in the emotion classification of clean images with an accuracy of 58.78% than distorted images with an accuracy of 16.7%. This result indicates that the standalone VGG-based model performs well in predicting clean image input, but the model is not robust in noisy image input.

3.4. Result of the hybrid model

Experiments on two hybrid model architecture were conducted, so as to pick up the one with higher accuracy in emotion recognition tasks.

3.4.1. Hybrid model variant A. In the Hybrid Model Variant A, the pre-trained DAE model was imported as the first layer of the hybrid model, and the subsequent layer is the VGG-based model. After 150-epoch training with the same train dataset as the standalone VGG-based model, the result is shown in Figure 7.

DOI: 10.54254/2755-2721/27/20230136

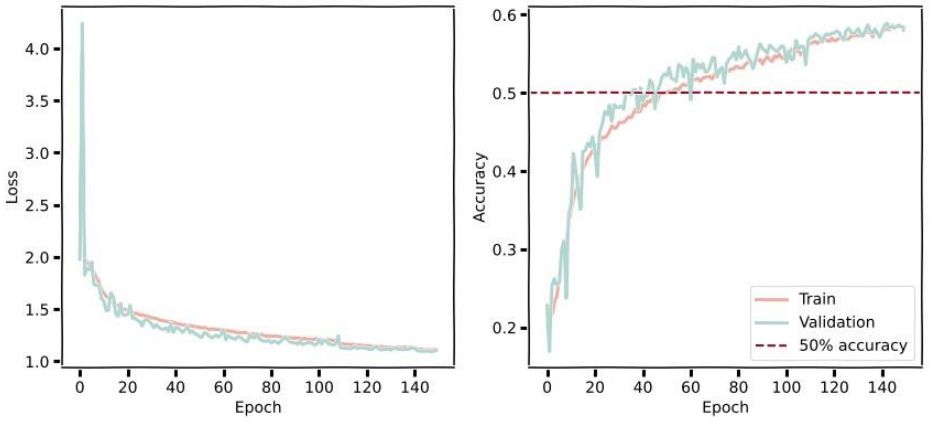

Figure 7. Hybrid model variant A's Loss Curve and Accuracy Curve.

In Figure 7, the loss curve and accuracy curve were generated with 150 training epochs. At the end of the training procedure, the loss value was 1.1108, the classification accuracy value reached 0.5840. Similar to the experiment on standalone VGG based model, two data set were used to evaluate the trained hybrid model, and the result is shown in Table 3.

Table 3. Hybrid model variant A evaluation result of 2 input image sets.

	Loss	Accuracy
Clean input images	1.0838	58.74%
Distorted input images	1.1033	57.73%

Table 3 reveals that the DAE and VGG-based network combination performs better than the standalone VGG-based network in the automated emotion recognition task with distorted image input. However, for classification with clean input images, the standalone VGG-based network has 58.78%, which is better than Hybrid Network Variant A's 58.74%. This result shows that introducing a DEA network ahead of VGG based network can make the whole neural network robust to image distortion to make the automated emotion recognition gain higher accuracy in the extreme working condition. But due to the DEA network's reconstruction effect, data loss may occur. As a result, Hybrid Model Variant A will have a slightly different performance on clean images from the standalone VGG model.

3.4.2. Hybrid Model Variant B. In Hybrid Model Variant B, the DAE and VGG networks were treated as a whole and trained together with the same input and hyper-parameters as Hybrid Model Variant A. The training result is shown in Figure 8.

DOI: 10.54254/2755-2721/27/20230136

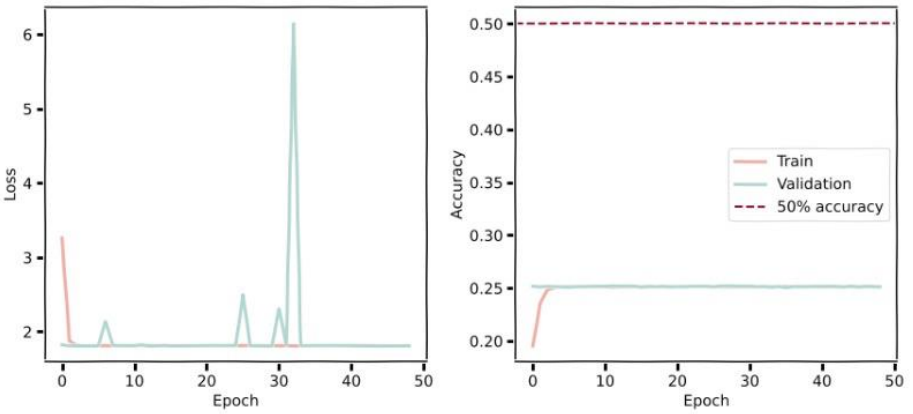

Figure 8. Hybrid model variant B's Loss Curve and Accuracy Curve.

Figure 8 indicated that Hybrid Model Variant B is not a feasible neural network architecture because the loss value was not decreased, and the accuracy value was never beyond 0.5. The training stopped at 49/150 epoch due to the early stopping policy, where the loss value is 1.8101, and the accuracy is 0.2514. In addition, the model evaluation result is shown in Table 4. According to Table 4, Hybrid Model Variant B does not work well in both input images.

Table 4. Hybrid model variant B evaluation result of 2 input image sets.

	Loss	Accuracy
Clean input images	1.8136	24.71%
Distorted input images	1.8136	24.71%

Based on the result, there is no performance difference between clean input images and distorted input images. All accuracy results are below 50%, meaning the model cannot classify emotions. This result indicated that the hybrid model should not be trained as a whole. The reason is that the DAE network needs distorted images as training input, while the clean images as training expected output to make the DAE network reconstruct the image correctly. However, this training procedure did not train the DAE network correctly, so down-sampling operations in encoder layers will discard useful features from clean and distorted input images. Consequently, Hybrid Model Variant B is deemed inappropriate for the emotion classification task.

To sum up, Hybrid Model Variant B structure is not a suitable neural network implementation for automated emotion recognition because the wrong-trained DAE network will discard features during the down-sampling stage. On the other hand, Hybrid Model Variant A structure is the feasible approach to build a robust neural network for emotion classification, which can handle both clean and distorted image input. Although Hybrid Model Variant A may experience a slight decrease in accuracy when predicting clean images compared to a standalone VGG-based model, such a tiny drop in accuracy is negligible based on the experiment result.

3.5. Discussion

From the results above, the hybrid model, which combines the DAE and VGG-based models, performs better than the standalone VGG based model in emotion recognition tasks with distorted image input. The reason for the experiment result is that distorted images lack many feature points for the neural network to catch. Moreover, the noise can also be identified as feature points by mistake. As a result, the standalone VGG-based model classifies emotions with lower accuracy while using distorted images.

DOI: 10.54254/2755-2721/27/20230136

On the other hand, by introducing the DAE network, distorted input images will be reconstructed to mitigate the noise. Therefore, the successor VGG network can classify emotions with higher-quality images. Then the whole hybrid network will be robust in distorted input images. With the same input data and hyperparameter set, the prediction accuracy on distorted images was increased from 16.70% to 57.73%.

What's more, the DAE model in the hybrid model should be correctly trained with the distorted images input and clean images as the expected output. DAE model should be inserted into the hybrid model as the frozen pre-trained module instead of being trained with the VGG model as a whole. An incorrect training procedure will cause the DAE model to lose the necessary features, so as to put the hybrid model out of action.

4. Conclusions

In the research, a hybrid neural network for automated emotion recognition was proposed to have a robust performance in distorted input images. This hybrid model can pre-process the input image in advance to deduct the noise and gain better classification accuracy in extreme working conditions. The hybrid model combines a pre-trained denoise autoencoder, and a VGG-based convolutional neural network, which has a solid theoretical foundation, and the model is easily implemented. Several experiments were conducted with the control variable method. The accuracy achieved by the standalone VGG-based model is 58.78% in clean images and 16.70% in distorted images. To improve the accuracy, the hybrid model was implemented, reaching 58.74% in clean images and 57.73% in distorted images. The experiment result proved that the hybrid model of DAE + VGG is robust to the random noise of input images. Also, the hybrid model has no remarkable accuracy deduction on the clean input images compared to the standalone VGG model. As a result, the hybrid model presented by the research is feasible to solve the distorted input image problem for automated emotion recognition. In the future, hyperparameters should be modified so that the hybrid neural network can gain higher accuracy in the automated emotion recognition task.

References

[1] Dzedzickis A Kaklauskas A and Bucinskas V 2020 *Human Emotion Recognition: Review of Sensors and Methods* Sensors 20 592.
[2] Tomkins S 1962 *Affect Imagery Consciousness: Volume I: The Positive Affects* (Springer Publishing Company).
[3] Izard C E 1977 *Human Emotions* (Boston, MA: Springer US).
[4] Plutchik R 1980 *Chapter 1 - A General Psychoevolutionary Theory Of Emotion* Theories of Emotion ed R Plutchik and H Kellerman (Academic Press) pp 3–33.
[5] Yu C and Wang M 2022 *Survey of emotion recognition methods using EEG information* Cognitive Robotics 2 132–46.
[6] Mather M and Thayer J F 2018 *How heart rate variability affects emotion regulation brain networks* Current Opinion in Behavioral Sciences 19 98–104.
[7] Song Z 2021 *Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory* Front. Psychol. 12 759485.
[8] Badrulhisham N A S and Mangshor N N A 2021 *Emotion Recognition Using Convolutional Neural Network (CNN)* J. Phys.: Conf. Ser. 1962 012040.
[9] Simonyan K and Zisserman A 2015 *Very deep convolutional networks for large-scale image recognition* 3rd International Conference on Learning Representations (ICLR 2015).
[10] Atabansi C C Chen T Cao R and Xu X 2021 *Transfer Learning Technique with VGG-16 for Near-Infrared Facial Expression Recognition* J. Phys.: Conf. Ser. 1873 012033.
[11] Lévêque L Villoteau F Sampaio E V B Perreira Da Silva M and Le Callet P 2022 *Comparing the Robustness of Humans and Deep Neural Networks on Facial Expression Recognition* Electronics 11 4030.

DOI: 10.54254/2755-2721/27/20230136

[12] Bajaj K Singh D K and Ansari Mohd A 2020 *Autoencoders Based Deep Learner for Image Denoising* Procedia Computer Science 171 1535–41.
[13] Vincent P Larochelle H Lajoie I Bengio Y and Manzagol P A 2010 *Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion* Journal of Machine Learning Research 11 3371–408.
[14] Manas S 2020 (online) *FER-2013: Learn facial expressions from an image* Kaggle Retrieved from https://www.kaggle.com/datasets/msambare/fer2013?select=test.
[15] Shorten C and Khoshgoftaar T M 2019 A *survey on Image Data Augmentation for Deep Learning* J Big Data 6 60.
[16] Singh D and Singh B 2020 *Investigating the impact of data normalization on classification performance* Applied Soft Computing 97 105524.