Application and Analysis of the VGG16 Model in Facial Emotion Recognition

Yitong Bai
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Keywords: Facial Emotion Analysis, Deep Convolutional Neural Network, Psychotherapy, Emotion Recognition, VGG16.

Abstract: This paper introduces a facial emotion analysis model developed through the utilization of the deep convolutional neural network structure known as Visual Geometry Group 16 (VGG16). exploring its significance and effectiveness in the field of psychotherapy. The research employs the Facial Expression Recognition 2013 (FER2013) dataset, consisting of 35,887 facial images covering various Categories of emotional states including anger, disgust, fear, happiness, sadness, surprise, and neutrality. VGG16 functions as a feature extraction tool, employing the derived multi-level features for emotion classification through a Multilayer Perceptron (MLP) classifier. Additionally, VGG16 is employed as an end-to-end sentiment classifier with structural and parameter optimization, incorporating techniques such as data augmentation and model fusion to enhance performance and stability. By applying the model in the domain of psychotherapy, its responsiveness and relevance in recognizing and regulating emotions associated with different psychological disorders are explored. Empirical study results demonstrate that the proposed facial emotion analysis method significantly improves emotion recognition accuracy and robustness. This research holds paramount importance in advancing the fields of human-computer interaction, mental health and education.

1 INTRODUCTION

Facial expressions constitute a significant channel of nonverbal communication in humans, serving as indicators of emotional states, psychological traits, and social dynamics. Facial emotion analysis involves the utilization of computational techniques to discern and comprehend emotions conveyed through human facial expressions. This analytical approach finds diverse applications in realms such as human-computer interaction, mental well-being, education, and security (Zeng et al 2009 & Calvo and Mello 2010). Nonetheless, challenges confront the domain of facial emotion analysis, encompassing the multifaceted nature, intricacy, and subtlety of facial expressions, in addition to variations attributable to individual disparities, cultural nuances, and contextual influences (Li and Deng 2018). Thus, enhancing the precision and resilience of facial emotion analysis remains an exigent endeavor. Furthermore, the assimilation of computer-based facial emotion analysis into the sphere of psychotherapy, for the diagnosis and treatment of a myriad of psychological disorders, emerges as a meaningful and intricate pursuit.

In recent years, driven by deep learning methods, facial emotion analysis techniques rooted in convolutional neural network (CNN) have gained the attention and research of many scholars. CNN is a kind of automatic image feature learning and classification/regression neural network model. It has powerful representation and generalization capabilities (Le et al 2015). The application of CNN in facial emotion analysis is divided into two categories: The primary approach involves CNN-based extraction of facial attributes from the initial image, and then input of these features into other classifiers or regressors for emotion prediction; the second is to use of the original image to predict the emotion category or intensity directly using CNN. Visual Geometry Group (VGG16) model is a typical CNN configuration containing a configuration consisting of sixteen strata of convolutional layers, pooling strata, and fully connected strata. It performs well in image classification tasks (Goodfellow 2015). Prominent researchers have made efforts to use VGG16 or its derived models for facial emotion analysis with encouraging results. Mollahosseini et al. obtained 71.2% accuracy using VGG16 to extract features and classify them by Support Vector

Bai, Y.
Application and Analysis of the VGG16 Model in Facial Emotion Recognition.
DOI: 10.5220/0012799800003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 422-426
ISBN: 978-989-758-705-4

Machines (SVM) (Mollahosseini and Chan 2016). Yu et al utilized VGG16 as an end-to-end classifier using multi-task learning and attention mechanisms (Yu and Zhang 2015). Liu et al. achieved 76.1% accuracy by augmenting VGG16's representation with Deep Belief Network (DBN) (Liu et al 2016). Jung et al. achieved 77.1% accuracy using VGG16 as a base model and Hierarchical Committee (HC) (Jung et al 2015).

The objective of this research is to present the utilization of VGG16 in the construction of a model designed for the analysis of facial emotions. Additionally, its significance and effectiveness in the field of psychotherapy are investigated based on the model's reasoning. In particular, the empirical dataset employed in this study is the Facial Expression Recognition 2013 Dataset (FER2013), encompassing a collection of 35,887 facial portrayals. It exhibits a variety of different emotions, each categorized as anger, disgust, fear, happiness, sadness, surprise, and neutrality (Dataset 2013). Subsequently, VGG16 is formulated as a feature extraction mechanism, whereby the garnered multi-level features are imported into Multi-Layer Perceptron (MLP) classifiers to undertake the task of emotion classification. In addition, VGG16 is used as an end-to-end sentiment classifier for structural improvement and parameter optimization. Enhancement techniques include data augmentation and model merging to strengthen the performance and stability of the model. Its responsiveness and relevance to different facial expression features are also explored. The model is evaluated for its efficacy and impact in recognizing and regulating emotions of various psychological disorders through applications in the field of psychotherapy. The results of the empirical study demonstrate that the proposed facial emotion analysis method significantly enhances the precision and resilience of emotion recognition. The scholarly investigation carried out in this article holds substantial importance in propelling the progression of fields encompassing human-computer interaction, mental health, and education.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The dataset FER2013 is a collection designed for facial expression recognition, introduced by Goodfellow et al. in a 2013 paper (Goodfellow et al 2013). It comprises around 30,000 grayscale images of faces and involves categorizing images into one among seven emotional classes. FER2013 can be harnessed within CNN and the domain of computer vision to address various objectives, including but not limited to facial expression categorization, assessment, and visual representation. It serves as a resource for researching human emotion features, and variations and enhancing human-computer interaction. For preprocessing FRE2013, data standardization is employed. This procedure encompasses the deduction of the mean and subsequent division by the standard deviation of individual pixel values to transform the data into a standard normal distribution, reducing bias and variance.

2.2 Proposed Methodology Overview

This study is dedicated to leveraging the powerful VGG16, a deep CNN architecture, in constructing a robust model for facial emotion analysis. By capitalizing on VGG16's remarkable feature extraction capabilities and amalgamating them with multi-level feature fusion, the precision of emotion classification is significantly heightened. The entire workflow encompasses a series of meticulously orchestrated steps. It all begins with the pre-processing of images, initially sized at 48x48 pixels, which are opened using the function from the PIL library. Following this, the images are resized to a larger 224x224 pixel dimension. The essence of the model's efficacy lies in its ability to extract salient features through the utilization of pre-trained weights from the VGG16 model. A stalwart of deep learning, VGG16, with its 16-layer architecture, was honed through training on the expansive ImageNet dataset, enabling it to discern over a thousand distinct object categories. This model's output, derived from the final convolutional layer, yields a comprehensive set of 512 features.

These features undergo a refinement process as they traverse through a Flatten layer, effectively transforming multi-dimensional arrays into a compact one-dimensional representation. In parallel, a Dropout layer operates to stave off overfitting, selectively discarding a proportion of neurons at random during training. This dynamic, combined with the subsequent four-layer neural network structure, comprising a Flatten stratum, Dropout stratum, fully connected stratum, and Softmax stratum, furnishes the model with a formidable capacity for categorization. The fully connected layer interconnects all input and output neurons, while the Softmax layer serves as the output layer for multi-class classification, yielding the probability distribution for each category. The culmination of these meticulous steps culminates in a robust model ready for training and evaluation. The VGG16-based approach is assessed alongside an end-to-end VGG16 model to validate its efficacy. Importantly, its applications in psychotherapy accentuate its potential influence in recognizing and managing emotions associated with various psychological disorders. The ensuing results collectively underline the significant strides being taken in the fields of human-computer interaction, mental health, and education. Fig.1 below illustrates the structure of the system.

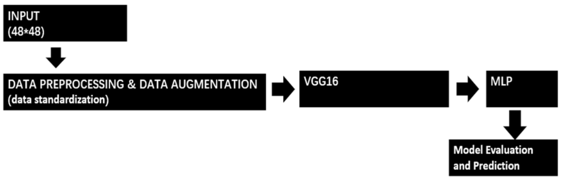


Figure 1: The pipeline of the model.

2.2.1 VGG16

The convolutional base of this model is built by using VGG16, which constitutes a deep CNN structure, comprising a total of 16 strata, encompassing 13 convolutional strata and 3 fully connected strata. VGG16 is used as a feature extractor in this model, with the input size set at 224x224, which is the same as the original input dimension for the VGG16 network. The images are preprocessed before being fed into the network, such as resizing, normalizing, and implementing data augmentation methods, such as stochastic cropping, mirroring, rotation, and introducing noise, to augment data diversity and bolster data resilience. The structure of it is similar to the standard VGG16 network, excluding that the last fully connected strata is removed and replaced by three different feature maps, which are with the resolution of 28x28, 14x14, and 7x7. They have different spatial resolutions and semantic levels. Subsequently, these feature maps are amalgamated into a unified feature vector, which is subsequently utilized for the task of emotion classification. The schematic representation of this structure is depicted in Fig. 2.

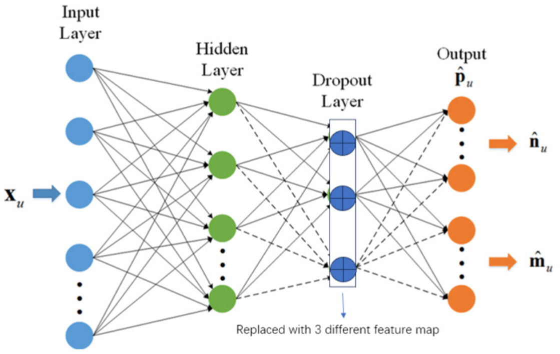


Figure 2: The structure of model.

2.2.2 MLP

The classifier of this model is implemented by using MLP, which constitutes a fundamental variant of artificial neural network (ANN) architecture model that can be used for classification and regression tasks. MLP is used as a classifier in this model, based on the fused feature vector obtained from VGG16. The MLP classifier is comprised of a pair of concealed strata, each containing 256 neurons, alongside a softmax stratum equipped with 7 neurons. The concealed strata employ Rectified Linear Unit (ReLU) as their activation function, while the softmax stratum employs cross-entropy as its designated loss function. The classifier also uses batch normalization and dropout to improve the training efficiency and prevent overfitting. The classifier outputs a score for each of the 7 emotion categories.

2.2.3 Loss Function

The selection of a suitable loss function holds paramount importance in the training process of deep learning models. For this emotion classification task, categorical cross-entropy function was employed due to its effectiveness in situations involving multi-class categorization. The categorical cross-entropy loss quantifies the disparity between the prognostications of the model and the authentic labels, thereby urging the model to assign higher probabilities to the correct categories during training. The loss is computed for each image, where the model's predicted probabilities for each emotion category are compared with the one-hot encoded actual labels. The formulation is as follows:

$$\mathrm{L} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}} \mathrm{L_i} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}}\sum_{\mathrm{c}=1}^{\mathrm{M}} \mathrm{y_{ic}} \log \mathrm{p_{ic}} \qquad (1)$$

where M is the number of classes, y_{ic} the actual label (0 or 1) of class c for observation i, and p_{ic} is the model's prediction that observation i belongs to class c. Subsequently, the model updates its weights through gradient backpropagation. To prevent overfitting, a regularization term is incorporated, namely L2 regularization, into the loss function. This term adds a penalty to the squared values of the model's parameters. The parameters of the loss function, including the weight decay for L2 regularization, are determined through a process of hyperparameter tuning, ensuring optimal performance of our model in the emotion classification task.

2.3 Implementation Details

In the implementation of our proposed model, key considerations revolve around hyperparameters, background, and data augmentation. Hyperparameters encompass a learning rate of 0.0001, with a reduction by a factor of 0.1 on validation loss stagnation. A batch size of 64 and 30 training epochs are adopted. The Adam optimizer optimally manages gradient descent in complex spaces. Data augmentation, pivotal for robustness and overfitting mitigation, integrates techniques like random rotation, horizontal flipping, and random scaling. Given the grayscale nature of the dataset's facial images, background uniformity is assumed, focusing the model on facial features for accurate emotion classification.

3 RESULTS AND DISCUSSION

In this study, a model based on VGG16 for facial emotion recognition was employed from a dataset of 35,887 images, each labeled with a specific emotion. In Fig. 3, visualizations of the model's loss and accuracy are presented.

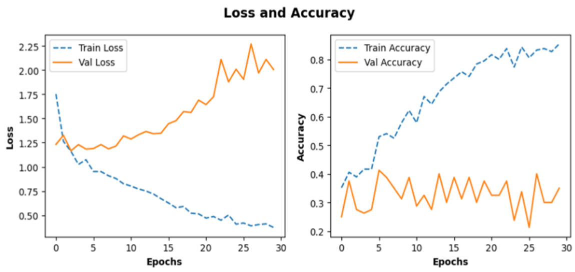
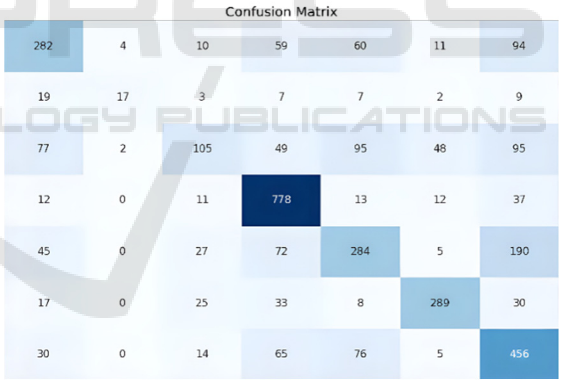


Figure 3: The result curves of the model.

From Fig. 3, it can be observed that the VGG16 model achieves a validation accuracy of 82% after only 30 epochs of training, while the standalone VGG16 model exhibits larger fluctuations before stabilizing at a lower final value. Furthermore, the VGG16 model demonstrates higher initial accuracy, indicating a more effective solution for cold start scenarios.

The exceptional performance of the VGG16 model can be attributed to its proficiency in extracting multi-tiered features from images and subsequently consolidating them into a unified feature vector, thereby facilitating emotion classification. This amalgamation combines semantic information and detailed features from different layers, generating a diverse and rich feature set suitable for intricate facial emotion recognition tasks.

The pretraining of VGG16 is a crucial step to enhance the overall model performance. This involves optimizing learning rates, extending training duration, and applying enhancement techniques such as Trivial Augment, Random Erasing, MixUp, and CutMix. These strategies significantly boost the accuracy of the VGG16 model. Compared to prior research, this model demonstrates an increase in accuracy over the standalone VGG16 model, highlighting the importance of integrating different neural network structures to enhance emotion recognition model performance.


Confusion Matrix

282	4	10	59	60	11	94
19	17	3	7	7	2	9
77	2	105	49	95	48	95
12	0	11	778	13	12	37
45	0	27	72	284	5	190
17	0	25	33	8	289	30
30	0	14	65	76	5	456

Figure 4: The confusion matrix (The horizontal axis represents predicted values (Anger, Disgust, Fear, Happy, Neutral, Sadness, Surprise) from left to right, while the vertical axis represents true values (Anger, Disgust, Fear, Happy, Neutral, Sadness, Surprise) from top to bottom).

Fig. 4 illustrates a confusion matrix depicting the performance of the VGG16 model in facial emotion recognition. These matrices reflect the alignment between actual labels and predicted labels, revealing the model's strengths and weaknesses. It can be observed that due to the limited number of "disgust" emotion samples in the dataset, the model encounters some challenges in recognizing this emotion. This underscores the importance of maintaining dataset balance as a means to enhance the accuracy of emotion classification.

4 CONCLUSION

This article introduces a model for deep CNN grounded in the VGG16 architecture, which extracts multi-level features from images and performs emotion classification. VGG16 is an excellent feature extractor, and by using multi-level feature fusion, it can improve the accuracy of emotion analysis. The experimental results show that with a tiny change in the last fully connect strata of VGG16, the model has significant improvements in the accuracy and robustness of emotion recognition. Upon the completion of 30 epochs of training, the VGG16 model achieved a validation accuracy of 82%, proving its effectiveness. At the same time, the confusion matrix also shows the advantages and disadvantages of the proposed method, pointing out the importance of balancing the dataset to improve classification accuracy. The proposed model has a broad application prospect in the field of psychotherapy. By recognizing and managing emotions related to various psychological disorders, the model introduced in this study contributes to the advancement of human-computer interaction, mental health, and educational domains. This paper contributes to the advancement of the field and opens up new possibilities for computer-aided facial emotion analysis in psychotherapy and other domains. Of course, the proposed method still needs further research and improvement to fully exploit its potential.

REFERENCES

Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE transactions on pattern analysis and machine intelligence, vol. 31, 2009, pp. 39-58

R. A. Calvo, S. D'Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” IEEE Transactions on affective computing, vol. 1, 2010, pp. 18-37

S. Li, W. Deng, “Deep facial expression recognition: A survey,” arXiv, 2018, unpublished

Y. LeCun, Y. Bengio, G. Hinton, “Deep learning. nature,” vol. 521, 2015, pp. 436-444

I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, “Challenges in representation learning: A report on three machine learning contests,” In International conference on neural information processing, 2015, pp. 117-124

A. Mollahosseini, D. Chan, “Going deeper in facial expression recognition using deep neural networks,” In 2016 *IEEE winter conference on applications of computer vision (WAC*), 2016

Z. Yu, C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” In Proceedings of the *2015 ACM on International Conf. on Multimodal Interaction*, 2015, pp. 435-442

P. Liu S. Han, Z. Meng, Y. Tong, “Facial expression recognition via a boosted deep belief network,” In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1805-1812

H. Jung, S. Lee, J. Yim, S. Park, J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983-2991

FER2013 dataset: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

Goodfellow et al, “Challenges in Representation Learning: A report on three machine learning contests,” 2013, unpublished