감정 인식을 위해 CNN을 사용한 최적화된 패치 특징 추출

하이더 이르판[1], 김애라[2],이귀상[3], 김수형[1]
[1]인공지능융합학과 전남대학교
[2]인공지능융합학과 전남대학교
[3]인공지능융합학과 전남대학교
irfan_haider99@hotmail.com, kimar3660@naver.com, gslee@jnu.ac.kr, shkim@jnu.ac.kr

Optimized patch feature extraction using CNN for emotion recognition

Irfan Haider[1], Aera kim[2], Guee-Sang Lee[3], Soo-Hyung Kim[1]
[1] [2] [3]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

In order to enhance a model's capability for detecting facial expressions, this research suggests a pipeline that makes use of the GradCAM component. The patching module and the pseudo-labeling module make up the pipeline. The patching component takes the original face image and divides it into four equal parts. These parts are then each input into a 2D convolutional layer to produce a feature vector. Each picture segment is assigned a weight token using GradCAM in the pseudo-labeling module, and this token is then merged with the feature vector using principal component analysis. A convolutional neural network based on transfer learning technique is then utilized to extract the deep features. This technique applied on a public dataset MMI and achieved a validation accuracy of 96.06% which is showing the effectiveness of our method.

Keywords: Emotion Recognition, Pseudo label, GradCAM, Image patches

1. Introduction

Research into emotion recognition is vital in the pursuit of creating machines capable of perceiving and responding to human emotions. Emotions can be conveyed in a variety of ways, from facial expressions and vocal tone to body language, making this a challenging task. Feature extraction and patch label analysis is one method currently in use for emotion recognition[1]. In order to classify feelings, it is necessary to first identify and then extract features from the input data. Facial expression analysis, for instance, can take advantage of feature extraction to determine things like eyebrow position, mouth shape, and eye movement. Separating the input data into smaller pieces, or "patches," and then labeling each patch according to its emotional content is what patch label analysis is all about. After assigning these labels, the underlying sentiment of the data can be categorized [2].

In this paper we have used the Convolutional NeuralNetworks (CNNs) and the patch based feature extraction[3]. When training and inferring, neural networks can selectively focus on different areas of the input data thanks to their reliance on attention mechanisms. In addition, a customized Resnet18 model is trained and deployed with a triplet loss function. Because of these features, CNN has proven to be highly effective for various issues in natural language processing, including language translation and sentiment analysis. Yet, when dealing with high-resolution images in computer vision tasks, the quadratic rise in processing complexity caused by attention in Transformers can become a bottleneck[1]. As Convolutional Neural Networks (CNNs) are able to process massive amounts of image data with significantly reduced computational complexity, they have become the de facto standard in computer vision.

2. Related Work

2.1 Convolutional Neural Networks (CNNs)

When it comes to studying the characteristics of data, deep learning is a machine learning technique that uses numerous processing layers. Because of its excellent performance in fields like visual object recognition and categorization, its popularity has recently skyrocketed [4].

When compared to traditional neural networks, CNN stands out due to its ability to extract and classify convolutional features. There are several layers involved in training a CNN, including convolutional, pooling, normalizing, activation, and softmax[5]. Forward and backpropagation form the backbone of a convolutional neural network. Back-propagation algorithms are used to optimize the settings of forward propagation algorithms. Several convolutional and link layers are used in forward propagation. Convolutional layers are distributed across layers to extract various features from the input data. First, simple features are extracted in the first convolutional layer, and then, as the layer's progress, increasingly complex features are extracted[6]. In addition, convolutional layers assign data-specific feature maps with local linkages of features obtained in precedinglayers[7].

2.2 Transfer Learning

When it comes to image classification tasks in computer vision, convolutional neural networks (CNNs) have typically been the go-to solution. Basically In general, transfer learning is the acquisition of prior expertise to gain novel expertise. In other terms, it is the search for similarities between two types of knowledge. The current knowledge is referred to as the source domain, and the new knowledge to be acquired is referred to as the target domain. Because the source domain differs from the one that is the target domain, we have to restrict the pattern of distribution distinction between the two domains and transfer knowledge to accomplish data validation[1]. New approaches of transfer learning can apply learned expertise in one area of emotional study to another [8].

2.3 GradCAM and PCA

GradCam and PCA have been the subject of numerous studies looking into emotion recognition. For instance, [9] used GradCam to see which parts of the face were being used by a CNN-based emotion identification model. They discovered that the model was preoccupied with the face, especially the eyes, eyebrows, and mouth, all of which play a significant role in expressing different emotions. In another study[10], employed principal component analysis to zero in on the key facial traits for spotting the six most universal expressions of emotion. Anger and despair were found to be best identified in the eyebrows and forehead, whereas happiness was shown to be best identified in the mouth region.

3. Methodology

First, we introduce the suggested pipeline that includes the GradCAM module to improve the model's face expression detection ability. To begin, we quadrupled the original face image. In order to extract feature vectors, a 2D convolutional layer will be fed a series of face fragments, each of which comprises a subset of the original face. The pseudo-labeling pipeline then takes each component and runsit through the GradCAM module to generate a weight token for the relevant feature vector. After using the principal component analysis(PCA) to optimize the features, the feature vectors for each of thefour components will be combined with their associated weight token to produce the final feature vector. Integrating a convolutional neural network block to the technique to extract the feature over these four feature vectors defines the facial expression in the original images. Figure 1 depicts the proposed pipeline.

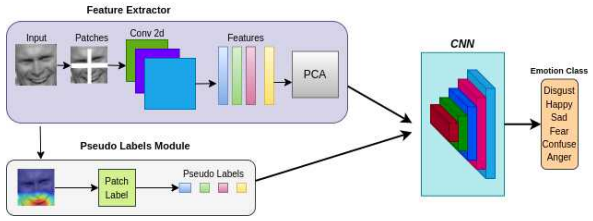

Figure 1: Proposed Pipeline

Each module of the proposed pipeline will be discussed in detail in the following subsections.

3.1 Patching Module

The patching module only performs the function of dividing images into smaller pieces. In this work, we take images with a 40x40 grid and split them into four sections, each of which is 20x20. After that, a convolutional layer with a 20x20 kernel size and stride will be used to generate a feature vector from each component.

3.2 Pseudo-Labeling Module

GradCAM is used in the proposed model to generate the weight token for each image region. After the original image has been passed into the GradCAM module, a CAM map will be generated. Then, we

applied a global pooling layer to the CAM map's representation of each image segment to generate weight tokens for those segments individually.

4. Results

We adapted a frame-extracting method to extract the most 20 representative frames in each video of MMI dataset, which resulted in 3145 static images. Then we use a random split using 0.7 ratio to divide the dataset into train and val set. We setup a training with batch size 256 and run for 250 epochs. To optimize the model, we use SGD algorithm with learning rate 0.01 and momentum 0.9. Training and testing curves on MMI dataset are shown in Figure 2.

Conclusion & Discussion

In conclusion, this paper presents a pipeline that, using the GradCAM module, boosts a model's capacity to recognize facial expressions. The patching module and

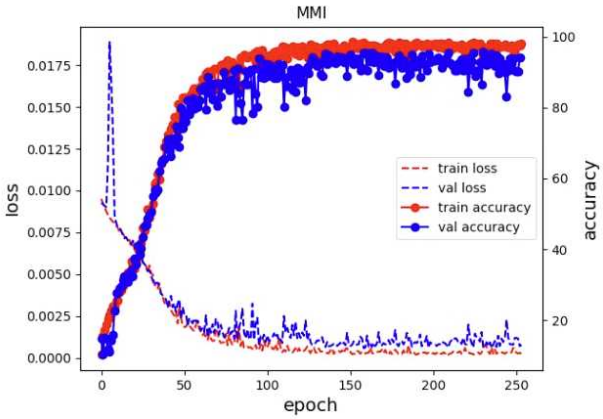

Figure 2: Training Curve of MMI Dataset

the pseudo-labeling module makeup the pipeline. Feature vectors are extracted from the original face image by the patching module, and weight tokens are created by the pseudo-labeling module with the help of GradCAM. A convolutional neural network is then utilized to deduce the subject's emotional state from the final feature vector. Improved accuracy in facial expression detection was seen after applying the proposed pipeline to a dataset of 3145 static photos. The pipeline can be used in numerous contexts, including emotion analysis and computer-generated facial animation. Researchers and practitioners in the fields of computer vision and machine learning may find the proposed pipeline to be a helpful tool.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408).

References

[1] S. Park, "ConvMixer: Patches Are All You Need? Overview andthoughts," CodeX, Nov. 02, 2021

[2] M. Aslan, "CNN based efficient approach for emotion recognition," Journal of King Saud University - Computer and Information Sciences, Aug. 2021

[3] G. Cao, Y. Ma, X. Meng, Y. Gao, and M. Meng, "Emotion Recognition Based On CNN," 2019 Chinese Control Conference (CCC),Jul. 2019

[4] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T.Shimamura, "Facial Emotion Recognition Using Transfer Learning inthe Deep CNN," Electronics, vol. 10, no. 9, p. 1036, Apr. 2021

[5] M. Sharma, A. S. Jalal, and A. Khan, "Emotion recognition using facial expression by fusing key points descriptor and texture features," Multimedia Tools and Applications, vol. 78, no. 12, pp.16195 - 16219, Dec. 2018

[6] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," IEEETransactions on Affective Computing, vol. 6, no. 1, pp. 1 - 12, Jan.2015

[7] E. S. Salama, R. A.El-Khoribi, M. E.Shoman, and M. A.Wahby,"EEG-Based Emotion Recognition using 3D Convolutional Neural Networks," International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, 2018

[8] Z. Huang, W. Xue, and Q. Mao, "Speech emotion recognition with unsupervised feature learning," Frontiers of Information Technology& Electronic Engineering, vol. 16, no. 5, pp. 358 - 366, May 2015

[9] F. Wang et al., "Emotion recognition with convolutional neural network and EEG-based EFDMs," Neuropsychologia, vol. 146, p.107506, Sep. 2020

[10] M. Marsot et al., "An adaptive pig face recognition approach using Convolutional Neural Networks," Computers and Electronics in Agriculture, vol. 173, p. 105386, Jun. 2020