1. Souha AYADI[1], 2. Zied LACHIRI[2]

University of Tunis el Manar,Tunisia (1)(2), Signal Image and Information Technology Laboratory , SITI, National Engineering school of Tunis

doi:10.15199/48.2024.08.31

Visual Emotion Recognition based on transfer learning technique using VGG16

Abstract. *Visual emotion recognition is one of the active topics nowadays. Recognizing emotions from a sequence of moving images still shows some difficulty in correctly detecting the exact features due to facial movement in the first place. Especially the movement of the mouth when pronouncing the sentence while producing emotions, which mainly affects the appearance of facial features. Thus, in this work, we focus on emotion recognition from facial expressions expressing speech. The deep neural network used in this work is VGG16 which is considered to be an effective neural network for detection and classification tasks, and can mainly be adaptable with transfer learning, technique. The presented method is conducted on the Video-speech category where we work on the detection of six classes of emotions which are: neutral, calm, happy, sad, angry and fearful, where the precision obtained is 78.12%.*

Streszczenie. *Wizualne rozpoznawanie emocji jest obecnie jednym z aktywnych tematów. Rozpoznawanie emocji na podstawie sekwencji ruchomych obrazów nadal wiąże się z pewnymi trudnościami w prawidłowym wykryciu dokładnych cech, przede wszystkim na podstawie ruchu twarzy. Zwłaszcza ruch ust podczas wymawiania zdania podczas wywoływania emocji, który wpływa głównie na wygląd rysów twarzy. Dlatego w tej pracy skupiamy się na rozpoznawaniu emocji na podstawie mimiki wyrażającej mowę. Głęboka sieć neuronowa wykorzystana w tej pracy to VGG16, która jest uważana za skuteczną sieć neuronową do zadań wykrywania i klasyfikacji i może być dostosowywana głównie do techniki transferu uczenia się. Prezentowana metoda opiera się na kategorii Wideo-mowa, gdzie pracujemy nad detekcją sześciu klas emocji, którymi są: neutralna, spokojna, szczęśliwa, smutna, zła i pełna strachu, gdzie uzyskana precyzja wynosi 78,12%. (**Rozpoznawanie emocji wizualnych w oparciu o technikę uczenia transferowego z wykorzystaniem VGG16**)*

Keywords: Visual-Speech emotion recognition, transfer learning, VG16.
Słowa kluczowe: Wizualne rozpoznawanie emocji, uczenie się transferowe, VG16.

Introduction

Human ability of recognizing emotions has been investigated from different point of views [5]. Especially with the development of existence technology, many views are presented in the field of human-machine intelligence, trying to develop intelligent techniques to automatically imitate this ability. The emotion recognition task is divided into two main subtasks based on where the emotion can be more detected and recognizable. Mainly auditory expressions and facial expressions.

In this article, we are interested in visual emotion recognition, which is based on the appearance of visual facial features while producing emotions. The detection of these characteristics is carried out either from a fixed frame (image) [10] or from sequential images (video) [12]. The production of different emotions always appears by combining different modality with the visual appearance, such as making a sound when laughing or simply speaking. All appearance forms can be a complementary form that complements the interested modality and gives meaning to the extracted features [4]. Which inspired on working on the recognition of emotions from the facial expressions of a spoken person. This category of visual speech type data is provided by the RAVDESS database [6].

Deep learning algorithms [1] are designed to mimic human capabilities and have the ability to learn from large amounts of data. This shows remarkable success for visual emotion recognition tasks [2], especially since most deep neural network architectures are built based on two-dimensional convolutional neural network layers, mainly dedicated to visual tasks. Adopting one of the most efficient neural network architectures is very delicate work, especially when it comes to adapting an existence architecture to new data with new classes to learn. This technique is called transfer learning. Essentially takes advantage of the knowledge of a successful neural network and adds new knowledge. Which works perfectly saving time by learning the basics already learned.

This work includes several subtasks, the first task is data preparation where we try to extract frames from the original video data. This way the training process will go faster and all the data will be processed without losing any information. By providing a high number of frames, there is no need to increase the data, so all relays will be on the choice of the neural network. The second task concerns the implementation of the choice neural network where we adapt the VGG16 [11] architecture considered as one of the deep neural network models.

VGG16 model

VGG16 stands for Visual Geometry Group [7], is one of the best deep neural network architectures for computer vision tasks. Mainly known as a type of convolutional neural network [3] because it is based on sixteen conv2D layers. It is characterized by very small convolutional kernels (3x3). Its ability to classify a thousand images among a thousand categories proves that this architecture is gentle for learning different classes other than objects as in our case different classes of emotions. Which is perfect for adopting the transfer learning method to readapt the neural network to the database. Each block of convolutional layers ends with a max-pooling layer, arranged consistently throughout the architecture.

The total number of convolutional layers in VGG16 is thirteen, with five maxpooling layers and three dense layers. The input image is been resized to (90,90) with 3 RGB channel. The total number of layers is twenty-one but only sixteen layers have learnable parameters, as shown in figure 1. The first block consists of two conv2Ds with a kernel of size 64. The second block consists of two conv2Ds with a kernel of size 128. The third block consists of three conv2Ds with a kernel of size 256. The fourth and fifth blocks are made up of three conv2Ds with a kernel of size 512. The fully connected layer consists of three dense layers with 4096 channels each. Between each one of them we replace a dropout layer to decrease the complexity of the neural network and minimise the overfitting problem. The final layer is the softmax layer for the classification result.

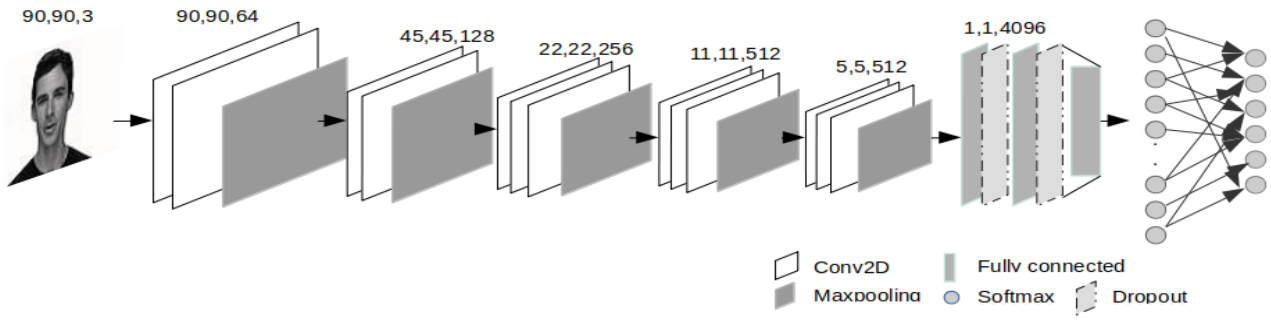

Fig.1. VGG16 architecture

Transfer Learning

Transfer learning [9] is a very popular technique used to adapt a pre-trained model on new data to solve a new problem. By taking advantage of the problem already learned and avoided, to continue improving the neural network's ability to solve a different problem. So, basically, the knowledge of the same expressions of emotions could be improved in terms of precision level and better avoid training problems. Technically, the weight of the previously trained information will be transformed and adapted to the new weight of the new data.

Despite the original application of VGG16, changing application type means changing domains, which means either having different feature spaces or different marginal distributions. Where the domain can be expressed mathematically as:

(1) $$D = \{X, P(X)\}$$

where it is based on two components:
X represent feature space.
P (X) represent a marginal probability distribution.
For each specific domain D, a task can be expressed as :

(2) $$T = \{Y, f()\}$$

Which is based on two elements, where Y refers to label space, and f () a predictive function. The predictive function f() can be learned from the training data (x_i , y_i) □ 1, 2...N , where x_i □ X and y_i □ Y .
The task can therefore also be written in the form :

(3) $$T = \{Y, P(Y|X)\}$$

Due the the fact that f (x_i) can be written $p(y_i|x_i)$. Which means the prediction function is based on the marginal distribution of labels based on the feature space. Thus, for different tasks, either the label spaces are different (Y_a # Y_b) or the conditional probability distributions are different (P (Y_a |X_a) # P (Y_b |X_b)). The task presented in this article is based on visual emotion recognition while the original task presented for VGG16 involved different images of different objects. So the difference between them lies in the labels.

From a technical point of view, transfer learning is applied by keeping the initial and intermediate layers and we only retrain the last layers. This technique is also known as sleep mode technique where we put the first previous layers into sleep mode and only use and modify the last layers, in this way we only approve an exact number of layers on which to work. In this case, it is recommended to only edit the last six layers. And the modification could be changing the kernel size inside each layer, as well as adding a dropout layer to avoid the overfitting problem.

Dropout

Deep learning neural networks are susceptible to rapid overfitting when using a transfer learning method to learn new models based on previously learned features. Different methods are known to reduce overfitting, most require more calculations and the creation of new models. An existing model can be used by randomly removing nodes from arbitrary layers during the training phase. This technique called dropout, which is a regularization technique that aims to minimize the complexity of neural network architecture without losing basic information, by forcing nodes in a layer to probabilistically take on more responsibilities to properly maintain the best functionality learned. Since the chosen architecture is already trained, only the weighted layers can be adjusted to the new classes. So that the dropout [8] cannot be placed arbitrarily, especially the neural network could not respond if placed between the main core of the architecture. So it is prudent to place it before the decision-making layers, which are the fully connected layers. In this case, we place the dropout layer between the dense layers. The value of the first dropout is 0.3, which means removing 30% of the nodes, and the value of the second dropout is 0.5, which means removing 50% of the nodes.

RAVDESS database

The Ryerson Audiovisual Database of Emotional Speech and Songs (RAVDESS) [6] is dedicated to emotion recognition studies for audio and visual data. Where it has three different types of modalities Audio-only, Audio-Video and Video-only. All modalities are available in two different forms: Speech and Song. Performed by 12 female and 12 male actors with a neutral North American accent vocalizing two lexically matched statements. This dataset contains 7,356 files of different numbers of emotions. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions, including neutral emotions for speech and song. Two levels of emotional intensity (normal, strong) are produced in each expression. The modality form that interests us is video only and belongs to the Speech form. The video form is treated as frames rather than a video by transforming each video into a number of frames. All extracted images are grouped into different folders by each class. We only use six classes belonging to speech, including neutral, calm, happy, sad, angry and fearful.

Results and discussion

The results are displayed as an accuracy curve, where the curves are colored with different colors: green, red, blue, and orange to indicate loss, validation loss, accuracy, and validation accuracy, respectively, as shown in figure 2.

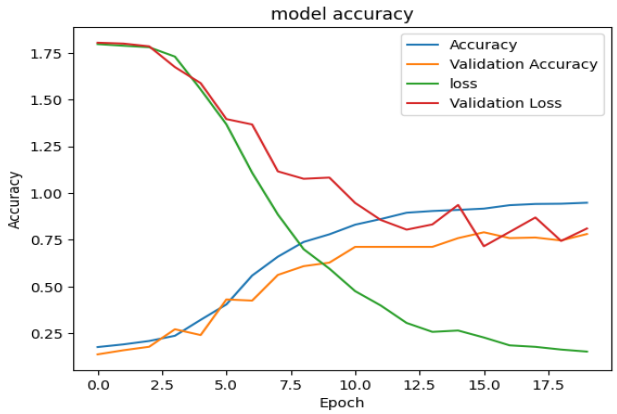

Fig.2. Accuracy results

The model is trained for 20 epochs, where the achieved train accuracy was 94.88%, validation accuracy was 78.12%, train loss was 0.1523, and validation loss was 0.8108, as shown in the table I.

Table 1. The achieved accuracy result

Epoch number	loss	accuracy	Validation loss	Validation accuracy
20	0.1523	0.9488	0.8108	0.7812

Contrary to our presumption, despite the good accuracy results, the neural network model still somehow shows an overfitting problem, where the validation loss is greater than the training loss, as shown in the curve form in figure 2. This problem concerns the high number of data used even if we know that the use of vgg16 is only successful for a smaller number of data, as well as the complexity of the model where the latter is based on sixteen layers in total.

Conclusion

Visual emotion recognition has been widely studied and greatly advanced by addressing the creation of different techniques to solve several still unsolved problems. Different techniques show to improve the recognition ability of neural network and better emotion classification, where transfer learning technique was adopted in this work to leverage the knowledge from previous successful neural network and focuses on the improvement in the level of precision. The deep neural network adopted is VGG16, based on sixteen two-dimensional convolutional layers which showed greater success for a small number of data. In our case, we tried to use the total number of data transformed from video to image to facilitate feature extraction and speed up the training process. The accuracy result obtained is 78.12%, which is considered good accuracy after using a large number of data.

Authors: *Souha AYADI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering, National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: souha.ayadi@enit.utm.tn; Zied LACHIRI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering,National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: zied.lachiri@enit.utm.tn .*

REFERENCES

[1] Souha Ayadi and Zied Lachiri. Deep neural network for visual emotion recognition based on resnet50 using song-speech characteristics. In 2022 5th International Conference on Advanced Systems and Emergent Technologies (IC ASET), pages 363–368. IEEE, 2022.

[2] Souha Ayadi and Zied Lachiri. Visual emotion sensing using convolutional neural network. Przeglad Elektrotechniczny, 98(3), 2022.

[3] Shuo Cheng and Guohui Zhou. Facial expression recognition method based on improved vgg convolutional neural network. International Journal of Pattern Recognition and Artificial Intelligence, 34(07):2056003, 2020.

[4] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. IEEE Transactions on Multimedia, 24:3480–3490, 2021.

[5] Essam H Houssein, Asmaa Hammad, and Abdelmgeid A Ali. Human emotion recognition from eeg-based brain–computer interface using machine learning: a comprehensive review. Neural Computing and Applications, 34(15):12527–12557, 2022.

[6] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one, 13(5):e0196391, 2018.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[8] PVVS Srinivas and Pragnyaban Mishra. An improvised facial emotion recognition system using the optimized convolutional neural network model with dropout. International Journal of Advanced Computer Science and Applications, 12(7), 2021.

[9] D Tamil Priya and J Divya Udayan. Transfer learning techniques for emotion classification on visual features of images in the deep learning network. International Journal of Speech Technology, 23:361–372, 2020.

[10] Hansen Yang, Yangyu Fan, Guoyun Lv, Shiya Liu, and Zhe Guo. Exploiting emotional concepts for image emotion recognition. The Visual Computer, 39(5):2177–2190, 2023.

[11] Haoyan Yang, Jiangong Ni, Jiyue Gao, Zhongzhi Han, and Tao Luan. A novel method for peanut variety identification and classification by improved vgg16. Scientific Reports, 11(1):15756, 2021.

[12] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 32(3):1034–1047, 2021.