Journal of Systemics, Cybernetics and Informatics (2024) 22(4), 1-5
https://doi.org/10.54808/JSCI.22.04.1

Transfer Learning for Facial Emotion Recognition on Small Datasets

Paolo BARILE
Dipartimento of Computer Science (DI)
Università degli Studi di Salerno
Salerno, Italia

Clara BASSANO
Department of Pharmacy (DIFARMA)
Università degli Studi di Salerno
Salerno, Italia

Paolo PICIOCCHI
Department of Social Science and Communication (DSPC)
Università degli Studi di Salerno
Salerno, Italia

ABSTRACT

In the context of human interactions, complexity science (CS) provides a way to understand the dynamics that arise from the interplay of different individuals. Recently, the possibility of applying the theory of CS to computer science, has shifted the focus to the research of machine learning methods for studying human behaviors and relational dynamics. Among the various existing AI techniques, facial emotion recognition (FER) has proven to be the best-performing and easiest to use human-computer interaction (HCI) tool for emotion detection. Despite the numerous existing approaches, the task of FER is not trivial, mainly due to the absence of large enough datasets to train deep learning (DL) models. A widely used solution is transfer learning (TL), which allows a model, pre-trained on enough data, to be used for a specific task where there is much less data available. The aim of our work is to test the effectiveness of TL for FER on an extremely small dataset, to understand which parameters need to be optimised to obtain the best outcomes. The results showed an overall accuracy of 85.54% for our model and revealed the concrete possibility of applying computer science to complex systems typical of the human psyche.

Keywords: Complexity science, Human-computer interaction, Emotion detection, Facial emotion recognition, Transfer learning.

1. INTRODUCTION

The technological development of the last few decades has marked an epoch-making change in the lifestyle and habits of human beings, on all social levels. With the spread of the Internet of things (IoT) and 5G communications, not only is technology within everyone's reach, but its use is becoming necessary in everyone's daily chores. It is no coincidence, given the current scenario, that numerous efforts in the field of artificial intelligence (AI) have recently been devoted to the field of affective computing (AC); the primary aim of AC is to facilitate the human-machine interaction by endowing the machine itself with a kind of emotional intelligence, so that the computer can succeed in simulating the typically human concept of empathy.

Emotion Recognition (ER) is the part of AC dealing with the identification of the user's emotional state through the study of data collected by sensors, cameras, microphones, and other devices. The collected information generally concerns physical parameters (such as heart rate, blood pressure and galvanic resistance) or a combination of data related to facial expressions, body posture and speech. In this context, we refer also to complexity science (CS) to look for a way to understand human behaviors and relational dynamics [1]. What emerges is that managerial practice focuses on the simplification of cognitive uncertainty, seeking paths different from linear and rational models [2]. Starting from this consideration, it is relevant also for future research to look for models that could minimize the decisional uncertainty induced by the complexity of socio-economic scenarios [3,4,5].

In particular, the conceptual categories of the Viable Systems Approach (VSA) [6,7], and Neuro-linguistic programming (NLP) appropriately allow the qualification of an intuitive model capable of reducing cognitive chaos and encouraging the making of effective decisions [8]. The logic of the model named Virtuous Circle of "Syntropy" (VSC) is found in the established ability of syntropic processes to guarantee the dynamic homeostasis (homeoresis) of organizational systems in contexts with high variability and, therefore, characterized by high decisional uncertainty. In this sense, the transduction from environmental Chaos with strong "entropy" to the Cosmos of the context because of syntropic processes represents a managerial must to which the (VSC) model contributes to providing an answer.

In terms of interdisciplinary education, the application of (CS) in computer science, particularly in studying human behavior and relational dynamics, as well as the use of machine learning methods to link AI and CS, is found in our future proposal of testing the (VSC) model of the process of "simplification" and "sustainability" of the paths of reduction of complexity in decision-making complication, enriching the instrumentation already established and commonly shared in management studies, in general, and in corporate decisions (decision making) under conditions of uncertainty [9].

Going in depth to our purpose, which is to emphasize the significance of emotions in human interactions, of all the ER techniques, the most widespread and widely used for a variety of applications is facial emotion recognition (FER) [10]. The idea

of identifying any movement of the human face in a purely descriptive manner, free from any possible interpretative inference, is far from being an easy task, since the human face can assume an infinity of different expressions, many of which have no particular meaning but are mere muscular actions devoid of any correlation; however, facial expressiveness is a very subjective concept that may come to depend on actor characteristics such as age, ethnicity, gender, facial conformation. A formal and widely used categorisation of the emotionality that transpires from facial expressions is provided by the Facial Action Coding System (FACS) [11]. According to FACS, the contraction and relaxation of facial muscles can generate more than 7000 different combinations of facial expressions, each of which can be traced to one of the 7 universally recognised primary emotions: neutral, fear, anger, happiness, sadness, disgust, surprise.

The task of matching a given facial expression to one of the emotional categories is far from trivial; several approaches in this field seem to favour deep learning (DL) techniques that can guarantee acceptable levels of accuracy and correctly classify expressions even when facial features are quite similar and difficult to attribute to one class rather than another. The main problem related to the use of neural networks in the context of FER is the lack of datasets capable of enabling optimal training for the DL models. A neural network of any type requires very time-consuming operations to train in terms of execution time and computational resources; furthermore, training a neural network from scratch requires a very large amount of data in order not to run into the very common problem of overfitting.

To solve these problems, several studies in the literature are oriented towards the technique of transfer learning (TL) [12]. TL consists of tackling a specific machine learning task using a model that has been pre-trained for a different purpose on a large amount of data. The success of the process is linked to the generalisation capacity of the selected model: where the features extracted in the training phase are not easily reusable in the new context, or the source task is poorly adapted to the target task, it becomes complicated to obtain satisfactory results. The great advantage in using the TL is, therefore, that of saving time and resources by using the features and weights derived from a previous training, carried out on a very extensive dataset, for the readaptation of a new model that can work on a more specific task with much less available data.

The aim of our work is to investigate the possibility of using a very restricted dataset for FER with the help of TL, and to understand the determining factors for achieving an optimal result with limited computational resources. The rest of the paper is structured as follows: in section 2 the literature related to our work will be analysed; in section 3 the experimental setup will be described, going into the characteristics of the dataset and the TL model; in section 4 the results of the study will be presented and discussed; finally, in section 5 the conclusions will be drawn.

2. RELATED WORKS

The analysis of facial expressions for FER has been undertaken in recent times with a wide variety of ML techniques. The traditional approach, regardless of the type of algorithm being used, involves a first phase of facial feature extraction and a second phase of classification (generally based on FACS criteria) [13]. The main advantage in using DL is the possibility of condensing the two phases, thus guaranteeing greater adaptability of the model to any type of image. The use of DL for FER has proven to be a very popular tool, especially in the last few decades; convolutional neural networks (CNN) have found wide exploitation due to their ability to reduce the high dimensionality of images without losing its information, which makes them one of the best tools for image classification [14]. Among the various works cited in the literature, very few address the task of FER by training the CNN from scratch (sometimes with poor results). Pranav et al. [15], for example, used a very light model of CNN with only two convolutional layers on a restricted dataset of self-collected facial images, managing to obtain a good accuracy (78.04%); however, this is an anomalous case, since in most similar works the accuracy level is between 40% and 60% [16].

The absence of a sufficiently large FER-specific dataset to allow for the optimal training of a neural network has led many experts in the field to switch to TL. In [17], for example, a two-stage fine-tuning process of the TL model for FER in the wild is proposed, which can increase the level of classification accuracy by up to 16% compared to the baseline method. Bentoumi et al. [18], on the other hand, addressed the problem of overfitting by introducing an innovative early stopping criterion during model training, managing to use TL with accuracies greater than 96% on several databases. Finally, Akhand et al. [19] reaffirmed the validity of TL for the realisation of a FER system by successfully testing 8 different pre-trained models on some of the best-known facial image datasets.

Furthermore, several studies use a DL approach based on TL in combination with other machine learning techniques; in this case, the neural network performs the feature extraction task while other ML methods are used for the classification. An example is in the work of Shaees et al. [20], who applied a hybrid model in which, after feature extraction by the CNN, classification is carried out using a support vector machine (SVM) algorithm.

3. EXPERIMENTAL SETUP

3.1 Dataset

The choice of dataset stems from the need to avoid the problem of overfitting, which is very frequent when attempting to train a complex machine learning model with a limited amount of data. For image classification algorithms the main risk is that the different classes of interest are not represented in a balanced manner within the dataset: in these cases, the model finds great difficulty in predicting the minority classes, thus invalidating the accuracy of the measurement. In order to avoid this, we chose to use the ”Japanese Female Facial Expression (JAFFE)” dataset [21], in which each of the classes of interest, corresponding to the 7 primary emotions identified by Ekman, is represented (to a very good approximation) by the same number of samples. JAFFE is made up of facial images of 10 Japanese women, each of whom was asked to interpret different facial poses for each of the 7 emotions to be classified; the dataset contains a total of 213 high-resolution grayscale images (256x256 pixels), all taken frontally with respect to the camera and with the same neutral background (Figure 1).

Figure 1: Sample images from JAFFE dataset.

All these qualities, together with the absence of disturbing elements on the images such as the presence of beards, spectacles, or other accessories, will help to significantly lighten the pre-processing phases and preserve the accuracy of the model.

3.2 Transfer Learning model

The model chosen for the use of transfer learning on the FER specific task is the MobileNetV2. It is a 53-layered deep convolutional neural network (DCNN) for image classification, pretrained on approximately one million images from the ImageNet database and counting 1000 output classes. MobileNetV2 uses a special type of key building block called Depthwise Separable Convolutional Layer (DWise) which guarantees a reduction of the computational cost by up to 9 times compared to traditional CNNs [22] and an increase in execution speed by up to 7 times for the specific task of image classification [23]. The aim is to lighten the convolutional process by performing the filter operation at each convolutional layer in two separate steps, splitting the three-dimensional kernel in two parts: a two-dimensional filter of unit depth and a mono-dimensional one (1x1 size). Another peculiarity of MobileNetV2 is the use of the Inverted Residual Linear Bottleneck, a particular skip-connection model [24] that prevents an excessive information loss and bypasses the vanishing gradient problem while preserving the model's accuracy. The main advantage of such a lightweight architecture is a model requiring very limited computational resources to run, making it available on most portable and non-portable devices. Transfer learning was realised by replacing the last dense layer of the original architecture in such a way as to decrease the number of possible outputs from 1000 to 7, corresponding to the classes of interest for the specific FER task to be achieved; this was only possible because of the great adaptability of the model.

3.3 Data augmentation and model training

The use of transfer learning cannot disregard the adaptation of the pre-trained model to the task to be fulfilled; for this reason, it is necessary for the CNN to be retrained on the selected dataset in such a way as to correctly re-use the features of the initial model for the specific problem to be ad- dressed. The possibility of retraining the model on a very restricted dataset such as the JAFFE represents, however, a fruitless attempt to achieve a reliable result: where the number of parameters of the neural network is much greater than the number of data, the accuracy of the classification is at great risk of being invalidated by overfitting. Data augmentation was performed precisely to overcome this problem, which would otherwise be insurmountable with a dataset consisting of only 213 images. The data augmentation operation created new "artificial" images to increase the sample size by randomly rotating, stretching, zooming, and shifting the individual images in the dataset (Figure 2). Through data augmentation, it was possible to increase the number of images to 79450, 90% of which made up the training set and the remaining 10% the test set used for validation. The data augmentation process obviously did not change the balance of the dataset, so that each class in both the training and test set remained represented by the same number of samples.

For the training procedure, the Adam [25] algorithm was used as optimiser and the sparse categorical cross entropy as loss function.

During the fitting phase, different values for the model parameters (number of epochs, batch size, learning rate) were tested, looking for the optimal settings to ensure the highest level of classification accuracy.

Figure 2: Examples of images from JAFFE dataset modified through data augmentation processes.

4. RESULTS AND DISCUSSION

4.1 Model evaluation

The model evaluations performed on the test set showed 85.54% as the best achieved result in terms of accuracy.

This is an important outcome, since, referring to the literature, it shows a significant increase in accuracy both in comparison to similar transfer learning models used on different datasets, and to trained from scratch examples of DL [26]. This certifies, firstly, the possibility of obtaining appreciable results with TL even using limited data and computational resources; secondly, it demonstrates the im- portance of choosing the right dataset for the investigation.

In identifying the optimal model configuration, we observed that certain parameters influenced the final accuracy of the classification more than others. The optimal batch size value, for example, was found to be 32 samples per gradient upgrade: a higher value caused an excessive burden on the model in terms of execution time and required memory, a lower value led to undesirable fluctuations in validation accuracy. Another fundamental parameter proved to be the learning rate; the choice of 0.001 as initial value allowed the model to converge to the optimum result in acceptable times, whereas the use of a higher learning rate (0.01) ended up assigning random weights to the model, invalidating the study.

Further significant considerations can also be made about the data pre-processing phase. Data normalisation, for example, turned out to be highly counterproductive in our case study, generating unpredictable fluctuations in accuracy and deteriorating it by more than 40% with respect to the optimal result; any face cropping operations, on the other hand, did not show very significant differences in the results with respect to the use of the original images in the dataset.

A possible explanation for the first evidence is that, although the normalisation was expected to improve the model accuracy by giving equal importance to all features, the operation seems to have caused a rescaling of the images that invalidated the data augmentation process and led the model to overfitting. As for face cropping, it would have been essential if the images used for the test phase had been completely different from those used for training; however, the use of images with the same neutral background, for both training and validation of the model, did not make this necessary.

4.2 Confusion matrix and Single Class Accuracy

The confusion matrix in Figure 3 shows an example of validation regarding the highest accuracy value achieved by the CNN (85.54%). From the analysis of the confusion matrix, we can observe that our model correctly recognised all images labelled as “surprise”, while in all other cases it resulted in misclassifications. The evaluation errors were negligible for the classes “neutral”, “fear” and “anger”, for which the Single Class Accuracy (SCA) value remained above 96%; what negatively influenced the overall accuracy of the model were, however, the classes “sadness” (SCA of 63.7%) and “disgust” (SCA of 54.6%).

This result highlights a problem, well known in the literature, related to FER algorithms: some emotions are more easily recognisable than others, and not all facial expressions can be easily traced back to a single one [27]; for this reason, the SCA differs according to the class predicted by the CNN, regardless of the parameters used in the training phase and the architecture of the selected model.

A possible interpretation for the problem is related to the strong subjective component that links emotionality to the expression that represents it: in other words, everyone is led to express his/her state of mind with different facial expressions and it is not always easy to find common traits; in the specific case of the JAFFE dataset, where facial expressions are not spontaneous but simulated, contributes to complicate the situation even further. A second explanation is related to the physiological reactions that characterise our face in the presence of certain emotions; these can be strongly distinctive in some cases (wide-open eyes, wide-open mouth, curved eyebrows are unambiguous indicators of surprise), not so much in others where the distinguishing features are less delineated.

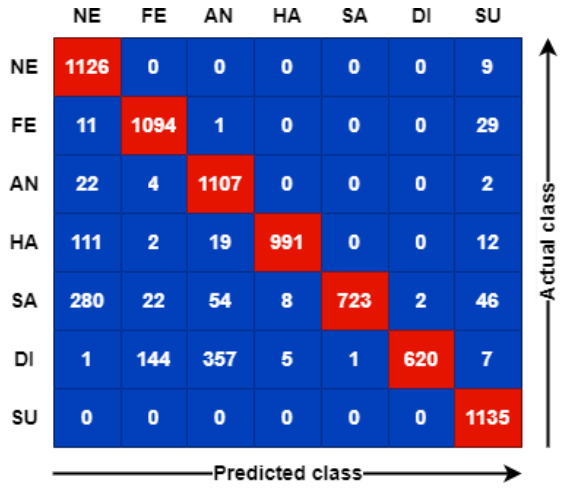

Figure 3: Confusion matrix for a 85.54% accuracy validation test. Each of the 7 classes is indicated by its own abbreviation (NE=neutral, FE=fear, AN=anger, HA=happiness, SA=sadness, DI=disgust, SU=surprise).

5. CONCLUSIONS

Our study revealed the possibility of successfully using TL for FER on a very small dataset. The danger of overfitting was properly averted thanks to the use of data augmentation; this demonstrated how, provided the right measures are taken, it is possible to work with a limited amount of data without degrading the accuracy of the results. Undoubtedly, the choice of parameters in the model training phase and the pre-processing operations remain particularly delicate steps, which can determine or compromise the success of the investigation. Ultimately, the TL proved to be a valid alternative for facial emotions classification where limited computational resources and time constraints do not allow satisfactory results to be obtained by training the model from scratch. It remains to be asked, with a look at future developments, to what extent the choice of the JAFFE dataset contributed to the achievement of the result and how the model would have behaved in the presence of data extrapolated from a real context in which it is not always possible to obtain ”clean” images (frontal with respect to the camera and free of disturbing elements).

From a CS perspective, the study of human behavior and relational dynamics through the use of machine learning, represents the perfect link between complexity and computer science. Our study showcases the synergistic relationship between AI and CS in unraveling human behavior and relational dynamics. By employing machine learning techniques, we delved into the nuanced nature of human emotions and the challenges inherent in their recognition. We untied the intricate patterns of human emotions, shedding light on the underlying complexities of interpersonal interactions.

Future research should continue to explore the potential of AI and TL in understanding human emotions in real-world contexts. By integrating insights from CS, we can further enhance our understanding of human behavior, paving the way for more empathetic and effective human-computer interactions in an increasingly interconnected world.

Acknowledgments

We would like to thank Francesco Caputo from the University of Naples “Federico II” for the useful suggestions and comments in order to improve the quality of our work.

6. REFERENCES

[1] S. Barile, “The dynamic of information varieties in the processes of decision making”, Proceeding of the 13th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2009.

[2] S. Beer, “Brain of the firm. The managerial cybernetics of organization”, The Penguin Press, 1972.

[3] J.H. Holland, Emergence. From Chaos to Order, Addison-Wesley, Reading, Ma. 1998.

[4] S. Kauffman, “The origins of order. Self-Organization and selection in evolution”, Oxford University Press, New York, 1993.

[5] A. Korzybski, Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics, Institute of General Semantics, New York, 1933.

[6] S. Barile, Management sistemico vitale. Parte prima decidere in contesti complessi. Giappichelli Editore, Torino, 2009.

[7] G.M. Golinelli, Viable Systems Approach (VSA). Governing Business Dynamics, Cedam, Kluwer, 2010.

[8] P. Piciocchi, et al. “The Virtuous circle of Syntropy (VCS). An interpretative chaos vs cosmos model for managing complexity”, Proceedings Act of the 2nd Annual Euromed Conference, University of Salerno, Fisciano, Italy, October, 26-28th, DOI: 10.3292

[9] H. Simon, “The sciences of the artificial”, (1st ed.), MIT Press, Cambridge, MA., 1969.

[10] C. Busso, et al. "Analysis of emotion recognition using facial expressions, speech and multimodal information."

Proceedings of the 6th international conference on Multimodal interfaces. 2004.
[11] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Palo Alto, Consulting Psychologists Press, 1978.
[12] JC. Hung, et al."Recognizing learning emotion based on convolutional neural networks and transfer learning." Applied Soft Computing 84 (2019): 105724.
[13] Y. Tan, et al. "A multimodal emotion recognition method based on facial expressions and electroencephalography." Biomedical Signal Processing and Control 70 (2021): 103029.
[14] A. Singh, et al. "Facial emotion recognition using convolutional neural network." 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM). IEEE, 2021.
[15] E. Pranav, et al. "Facial emotion recognition using deep convolutional neural network." 2020 6th International conference on advanced computing and communication Systems (ICACCS). IEEE, 2020.
[16] HW. Ng, et al. "Deep learning for emotion recognition on small datasets using transfer learning." Proceedings of the 2015 ACM on international conference on multimodal interaction. 2015.
[17] D. Nguyen, et al. "Meta transfer learning for facial emotion recognition." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.
[18] M. Bentoumi, et al. "Improvement of emotion recognition from facial images using deep learning and early stopping cross validation." Multimedia Tools and applications 81.21 (2022): 29887-29917.
[19] M. A. H. Akhand, et al. "Facial emotion recognition using transfer learning in the deep CNN." Electronics 10.9 (2021): 1036.
[20] S. Shaees, et al. Facial emotion recognition using transfer learning. In: 2020 International Conference on Computing and Information Technology (ICCIT-1441). IEEE, 2020. p. 1-5.
[21] M. Lyons, (2021). “Excavating AI” Re-Excavated: Debunking a Fallacious Account of the Jaffe Dataset. SSRN Electronic Journal.
[22] L. Bai, et al. "A CNN Accelerator on FPGA Using Depthwise Separable Convolution," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 65, no. 10, pp. 1415-1419, Oct. 2018.
[23] B. Sun, et al. "LRPRNet: Lightweight deep network by low-rank pointwise residual convolution." IEEE Transactions on Neural Networks and Learning Systems (2021).
[24] M. Sandler, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[25] R. Poojary, A. Pai. "Comparative study of model optimization techniques in fine-tuned CNN models." 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA). IEEE, 2019.
[26] M. Rescigno, et al. "Personalized models for facial emotion recognition through transfer learning." Multimedia Tools and Applications 79 (2020): 35811-35828.
[27] L. Liu, "Human Face Expression Recognition Based on Deep Learning-Deep Convolutional Neural Network," 2019 International Conference on Smart Grid and Electrical automation (ICSGEA), Xiangtan, China, 2019, pp. 221-224.