Herag Arabian*, Verena Wagner-Hartl, and Knut Moeller

Network Architecture Influence on Facial Emotion Recognition

https://doi.org/10.1515/cdbme-2022-1134

Abstract: Artificial Intelligence has been blending into daily life by means of many useful applications from voice command to facial recognition. One therapeutic application to be supported by AI solutions is treatment of people with Autism Spectrum Disorder. A closed loop feedback system is planned in conjunction with a novel reward system that will encourage the user to express emotions and be rewarded for it in a virtual environment. In this work five popular neural network architectures of VGG16, ResNet50, GoogleNet, ShuffleNet and EfficientNetb0 are studied and compared, with the aim of finding a relation between accuracy and developed features based on the architecture, for the application in Facial Emotion Recognition (FER). Three datasets were used, the OULU-CASIA for training and validation, alongside FACES and JAFFE for robustness analysis. The images were first pre-processed to eliminate background noise. The performance of the model was based on the true positive predictions with Grad-CAM prediction visualizations to visualize the focus of the networks in making decisions for classification. Results showed that deep network architectures with high parameter space performed best, with architecture design showing more influence on the region of focus than on classification results. This is attributed to the different layer combinations as well as parameters used for feature extraction. Shallow depth networks with high parameter space performed better than deep networks with low parameter space for FER application.

Keywords: Autism Spectrum Disorder (ASD), Deep Learning, Facial Emotion Recognition (FER), Therapeutic Application.

***Corresponding author: Herag Arabian:** Institute of Technical Medicine (ITeM), Hochschule Furtwangen University, Jakob Kienzle Str. 17, VS-Schwenningen 78054, Germany, E-Mail: H.Arabian@hs-furtwangen.de
Verena Wagner-Hartl: Departement of Industrial Technologies, Campus Tuttlingen Furtwangen University, 78532 Tuttlingen, Germany
Knut Moeller: Institute of Technical Medicine (ITeM), Hochschule Furtwangen University, VS-Schwenningen 78054, Germany

1 Introduction

Deep learning has become a popular topic over the recent years. The use of deep learning has been adapted into daily life through many useful applications i.e., voice command recognition and transcription. The popularity of intelligent systems coupled with the willingness of people to adopt this technology, has paved the way for researchers to implement deep learning algorithms for different applications and industries i.e., financial institutions, security applications and medical domain.

One such application of deep learning in the medical field is for the recognition of human emotions [1]. Emotion recognition is currently being studied as a method to help treat patients with Autism Spectrum Disorder (ASD), a developmental brain disorder that affects the social interactions and communications of individuals [2]. A closed-loop feedback system is being developed which immerses the subject in a virtual reality environment i.e., a game, in which the subject is presented with different scenarios ranging from social skills building to emotional reaction inducing stimuli. To encourage the subject to interact more and express emotions, a novel reward system is in development. The system is designed such that the user will receive incentives as rewards inside the virtual environment whenever they are able to express an emotion to the stimuli in the game [3].

The concept of using emotion recognition to help people with ASD has been studied in different works with promising results. A clinical trial performed in [4] showed that the use of such a treatment method resulted in better social skills in children with ASD. In [5] a pilot study was performed, and the virtual reality treatment system administered yielded encouraging results. The implementation of a facial emotion recognition (FER) system studied in [6] provided a positive outlook into the improvement of ASD children's interactions and behavioural monitoring. These studies show that such treatment interventions assist the physicians work in helping alleviate the effects of ASD.

As a main component of the system, expressed emotions must be efficiently recognised. The use of facial expressions for emotion recognition is being studied for this application.

The facial expressions were selected as 55% of a person's feelings and attitudes are conveyed by facial gestures [7]. To translate images of facial expressions into an emotion class, a Convolution Neural Network (CNN) was adopted. Different CNN architectures were trained on one and tested against two other databases to further evaluate the robustness of the chosen models.

In this study three databases of OULU-CASIA [8], FACES [9], and Japanese Female Facial Expressions (JAFFE) [10] were selected for model analysis. The images from the datasets were first pre-processed to remove background noise and highlight the face of the subject in the image as per previous work [11]. After which each model is trained and validated on the OULU-CASIA [8] database and tested against the FACES [9] and JAFFE [10] datasets. The performance of the model is based on prediction accuracy along with a quantitative analysis of the regions of focus of the model for classification, extracted by the Gradient-Weighted Class Activation Mapping (Grad-CAM) technique [12].

The aim of this study is to evaluate and compare the performance of different network architectures for FER.

2 Methods

Images from the three datasets of OULU-CASIA [8], FACES [9], and JAFFE [10] were first pre-processed by a segmentation algorithm developed in previous work [11] to remove the background noise and focus on the face of the subject. After which five different CNN model architectures were selected, the VGG16 [13], GoogleNet [14], ResNet50 [15], ShuffleNet [16], and EfficientNetb0 [17]. Transfer learning was used as the initial weights for the FER training. The pre-trained weights of the different models are the results of training the given CNN architecture with ImageNet datasets. The models were then re-trained and validated on the OULU-CASIA [8] database.

2.1 Model Selection & Training Options

When it comes to deep learning the idea of going 'deeper' for a better performance has stood out [14]. However, some studies have argued that with deeper models the possibility of overfitting increases and at the same time a drawback of computational inefficiency rises. The deeper the network the more computational time and demand [14]. To this end different architectures of varying depth and parameter calculations were studied.

Different CNN model architectures were reviewed and five were selected based on factors such as depth, number of parameters, same input dimension, as well as computational concept. The chosen architectures were VGG16 [13], GoogleNet [14], ResNet50 [15], ShuffleNet [16], and EfficientNetb0 [17]. Table 1 describes the differences between model architectures. The models were trained according to the training options of Stochastic Gradient Descent with Momentum (SGDM) as solver, with 100 Epochs and batch size of 50, shuffling at every Epoch, at a constant learning rate of $1e^{-4}$ and 0.9 as momentum.

Table 1: Architecture design comparison.

Architecture	Depth	Parameters	Layers
VGG16	16	138 M	41
GoogleNet	22	7 M	144
ShuffleNet	50	1.4 M	172
ResNet50	50	25.6 M	177
EfficientNetb0	82	5.3 M	290

The concept of the GoogleNet [14] is a stacked inception-based approach, this method balances the increase in depth with lower computational complexity. The ResNet50 [15] is designed as a bottleneck to enhance the performance of the training by using residual mapping to tackle the saturation degradation problem. ShuffleNet [16] is designed with needs of computational limited resources in mind. The architecture promotes grouped convolution and involves more feature map channels which encode more information, crucial for small network performance. The VGG16 [13] is modelled based on the notion that depth is critical to performance accuracy and visual representations. In EfficientNets [17] the concept of scaling up a network architecture by means of a compound scaling method was addressed. The architecture shows that carefully balancing network width, depth, and resolution is important for improved accuracy and efficiency.

2.2 Database Description

The Oulu-CASIA database is a collection of image sequences captured from 80 different subjects expressing six basic emotions of Anger, Disgust, Fear, Happiness, Sadness and Surprise. The database was built with three different lighting situations and two image capturing techniques [8]. Images of original RGB, visible light with strong illumination lighting were selected with a total of 10,379 images.

The Japanese Female Facial Expressions (JAFFE) database is composed of 10 different Japanese female students expressing seven emotions (six basic plus Neutral) totalling 213 facial portrait images portrayed in grey scale [10]. The FACES dataset is made of images from varying subject ages.

It has a total of 2,052 images expressing six emotion classes of Anger, Disgust, Fear, Happiness, Neutral and Sadness [9].

2.3 Performance Criteria

To analyse and compare between the different models, performance criteria were set. The OULU-CASIA [8] was first partitioned into 90% training and 10% validation set, with a randomised selection. The same images were used for each of the five models to have a fair comparison. The FACES [9] and JAFFE [10] datasets were used for the model robustness performance analysis to unseen data from different sources.

The model performances were based on true positive prediction accuracies of the validation set, while robustness on the testing datasets. The visualization technique of Grad-CAM [12] was used as a performance metric to observe if the models were focusing on regions relative to emotion classification.

The visualizations were calculated between the 'SoftMax' probability layer and the final layer before either the Global Average Pooling (GAP) or Fully Connected (FC) layer of each of the different models. The visualization feature maps were computed for each image; after which they were averaged across the class. The class feature maps were then compared with a class agnostic mask, generated in [11], that highlighted the regions of importance for emotion classification. The Dice [18] similarity coefficient was used as a metric for region of focus analysis.

Table 2: Distribution of images into the respective classes of each of the datasets of OULU-CASIA, FACES and JAFFE before and after image pre-processing.

	OULU-CASIA		FACES		JAFFE	
Class	**Orig.**	**Proces**	**Orig.**	**Proces**	**Orig.**	**Proces**
Anger	1790	1538	342	292	30	30
Disgust	1633	1425	342	292	29	29
Fear	1796	1734	342	292	32	32
Happy	1791	1725	342	292	31	31
Sad	1668	1445	342	292	31	31
Surprise	1701	1432	0	0	30	27
Total	10379	9299	1710	1542	183	180

3 Results & Discussion

The image pre-processing phase is analysed and results shown in Table 2, along with the distribution of the images for each of the three datasets into their respective classes. The image pre-processing algorithm excluded 10.41%, 8.77% and 1.41% of the images of OULU-CASIA, FACES and JAFFE respectively, from further processing due to the failure of the method to segment the prescribed regions. The Neutral class was removed from the analysis. The class distribution was near equivalent presenting no bias towards a specific class.

3.1 Model Performance

The performance of the different models on the validation and testing sets are represented in table 3. As can be seen from the results the VGG16 architecture achieved the best performance on the validation set and this complemented the robustness performance achieving 52.46% for the FACES and 38.89% for the JAFFE dataset. The EfficientNet architecture achieved the lowest performance in the validation set with ResNet50 having the worst robustness accuracy among all models. To get a clearer perspective a look at the Grad-CAM prediction visualizations and similarity coefficient was taken.

Table 3: Performance results of the true positive predictions of the five different models on the validation and testing sets.

%	OULU-CASIA	FACES	JAFFE
VGG16	**99.46**	**52.46**	**38.89**
ResNet50	94.73	17.25	17.22
GoogleNet	96.88	43.64	35.00
EfficientNetb0	91.71	22.76	18.89
ShuffleNet	96.99	36.77	29.44

Figure 1 represents the Dice similarity coefficient results for the validation and testing sets for each of the five trained models. A variable performance was noticed between the different datasets, with ResNet50 having highest values in the testing sets at 57.11% and 63.55% for the FACES and JAFFE respectively. This comes as a contrast to what was observed in the accuracy results (Table 3), where the Resnet50 showed the worst performance. This anomaly is difficult to interpret with

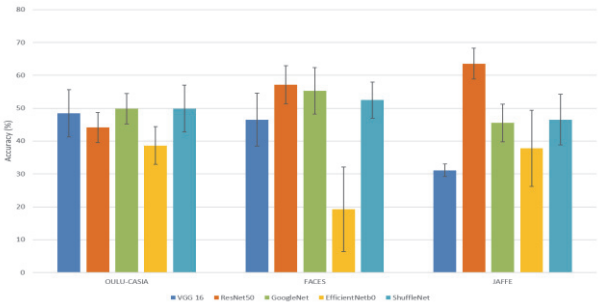

Figure 1: Dice similarity coefficient results for each of the datasets and model architectures. Average and Standard Deviation over the five trained models are shown.

the given measures. A visual inspection of the Grad-CAMs showed the differences in informative features focus area; however, a quantifiable measure was not performed and was left for development in future work. The design of the architecture was considered as the main influence factor for such observations.

The results obtained helped in disregarding, at this stage, some network architectures from further consideration. However, a definitive optimal architecture could not be achieved. The main performance metric for any architecture is considered to be the true positive accuracy performance, however, the visualization measures tell a different story.

4 Conclusion

In this study the influence of CNN architectures on FER was analysed. The results showed that the VGG16 model performed best in terms of a quantitative true positive measure. However, this contrasted with the similarity metric analysed which showed best performance, in terms of region of focus, for the ResNet50 model. Therefore, it was deduced that the design of the architecture has more influence on model region of focus for decision making, which is important for classification. The results shown are preliminary; further research is required. Testing on a closed loop system is being planned with a general user interface already established, further tests and analyses will be carried out in real time.

Author Statement

Research funding: Partial support by a grant from the German Federal Ministry of Research and Education (BMBF) under project No. 13FH5I06IA – PersonaMed is gratefully acknowledged. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, 'Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition', in 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, pp. 2983–2991. doi: 10.1109/ICCV.2015.341.

[2] C. on C. W. Disabilities, 'The Pediatrician's Role in the Diagnosis and Management of Autistic Spectrum Disorder in Children', Pediatrics, vol. 107, no. 5, pp. 1221–1226, 2001

[3] H. Arabian, V. Wagner-Hartl, J. Geoffrey Chase, and K. Moeller, 'Image Pre-processing Significance on Regions of Impact in a Trained Network for Facial Emotion Recognition', presented at the IFAC BMS, 2021.

[4] C. Voss et al., 'Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial', JAMA Pediatr., vol. 173, no. 5, pp. 446–454, May 2019.

[5] V. Ravindran, M. Osgood, V. Sazawal, R. Solorzano, and S. Turnacioglu, 'Virtual Reality Support for Joint Attention Using the Floreo Joint Attention Module: Usability and Feasibility Pilot Study', JMIR Pediatr. Parent., vol. 2, no. 2, p. e14429, Sep. 2019, doi: 10.2196/14429.

[6] M. Leo et al., 'Automatic Emotion Recognition in Robot-Children Interaction for ASD Treatment', in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Dec. 2015, pp. 537–545.

[7] A. Mehrabian, 'Communication without words', in Communication Theory, C. D. Mortensen, Ed. Routledge, 2017.

[8] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, 'Facial expression recognition from near-infrared videos', Image Vis. Comput., vol. 29, no. 9, pp. 607–619, Aug. 2011.

[9] N. C. Ebner, M. Riediger, and U. Lindenberger, 'FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation', Behav. Res. Methods, vol. 42, no. 1, pp. 351–362, Feb. 2010.

[10] M. J. Lyons, M. Kamachi, and J. Gyoba, 'Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)', Sep. 2020, doi: 10.5281/zenodo.4029680.

[11] H. Arabian, V. Wagner-Hartl, and K. Moeller, 'Image Pre-processing Effects on Attention Modules in Facial Emotion Recognition', presented at the IUPESM World Congress on Medical Physics and Biomedical Engineering (IUPESM WC2022), In Press.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization', Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, Feb. 2020.

[13] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', ArXiv14091556 Cs, Apr. 2015, Accessed: Aug. 04, 2021.

[14] C. Szegedy et al., 'Going Deeper with Convolutions', ArXiv14094842 Cs, Sep. 2014, Accessed: Aug. 17, 2021.

[15] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', ArXiv151203385 Cs, 2015.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, 'ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices', ArXiv170701083 Cs, Dec. 2017

[17] M. Tan and Q. V. Le, 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', ArXiv190511946 Cs Stat, Sep. 2020

[18] L. R. Dice, 'Measures of the Amount of Ecologic Association Between Species', Ecology, vol. 26, no. 3, pp. 297–302, 1945, doi: 10.2307/1932409.