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

# Transfer Learning in Facial Emotion Recognition: Useful or Misleading?

https://doi.org/10.1515/cdbme-2022-1170

**Abstract:** The use of machine learning in medicine holds a lot of potential in the domains of patient diagnosis, monitoring and treatment. One such application is the use of Emotion intelligence to aid in the treatment of people suffering from autism spectrum disorder (ASD). As training a robust network model requires large datasets, transfer learning is often implemented. The aim of this study is to show if using pre-trained weights, trained on different images, as an initial starting point for training a new model remains biased after training and does not generalize well to unseen data of the new trained model. Image pre-processing was performed and the data trained on three models, the base model of VGG16 architecture and two with attention modules of SE and CBAM. The OULU-CASIA database was used for training with 10-fold cross validation for evaluating the performance of the model and the robustness was tested against two other emotion datasets of FACES and JAFFE. The results showed the training from scratch had better adherence to the regions of importance in the image. This validates the hypothesis that prior knowledge, i.e. weights from pre-trained models of large datasets, may not be usable for special applications.

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

## 1 Introduction

Artificial intelligence (AI) has been a topic that has driven researchers for many years, and what was once fiction has now become part of mainstream reality. The use of AI has been incorporated into many daily routines from the simplest use in leisure of highlighting the face in photo capturing, to advanced financial predictions and security applications. The use of machine learning or AI in medicine holds a lot of potential in the domains of patient diagnosis, monitoring and in treatment. The implementation of AI in the medical field has not been widely adopted because of a number of reasons being of ethical concerns to reliance [1].

One such application of AI in medicine is the use of emotion intelligence to aid in the treatment of people suffering from autism spectrum disorder (ASD). ASD is a developmental brain disorder which affects the social skills of an individual by hindering their interaction, communication, behaviours, and interests [2]. Estimates reveal that 1 out of 59 people are affected by ASD [3]. Children suffering from ASD are accustomed to a certain routine and any deviation from the normalcy can cause psychological and emotional challenges to the child as well as an increased stress levels for the caregiver [4]. An individually adjusted virtual world combined with a reward system in the form of a gaming platform and technical affinity of most ASD individuals creates a suitable atmosphere for treatment [1], [5].

Similar treatment modalities have shown the improvement and acceptance of subjects to the technology. In [6] the use of such a treatment method was tested in a clinical trial showing improvement in the socialization of children with ASD. In [7] facial emotion recognition (FER), was used in a closed loop system that provided positive perspective towards ASD children's interactions and behavioural monitoring. A small pilot study performed in [8] implemented a closed loop virtual reality environment with encouraging results. This suggests that such a system is beneficial for the support of ASD patients in improving social skills.

In order for the smooth adoption of the treatment modality the real time data acquisition for emotion recognition should not deter the use of the technology. This combined with the study that 55% of a person's emotions are recognized through facial expressions [9] is the grounds to use FER by capturing the image of a subject via camera during system use. The convolutional neural network (CNN) was adopted, because of the popularity and ability of neural networks to outperform

***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:** Department of Industrial Technologies, Campus Tuttlingen Furtwangen University, 78532 Tuttlingen, Germany
**Knut Moeller:** Institute of Technical Medicine (ITeM), Hochschule Furtwangen University, VS-Schwenningen 78054, Germany

traditional machine learning algorithms [10], in order to classify the image captured into a relevant emotion class.

To train a CNN model to provide robust results, a large dataset of images is needed. Large data training requires many resources and takes a significant amount of time, to bypass this hindrance many researches have opted for transfer learning i.e. the use of a trained networks parameters as the initial start to their study [11]. The transfer learning has shown faster convergence and sometimes better results because of previously learned patterns. The results however may be misleading as pre-trained networks are widely trained on ImageNet [12] datasets, which is regarded as the benchmark for comparing the quality of network designs, and transferring this learned pattern data to another study is often irrelevant.

In this study the use of transfer learning is compared with the training from scratch for FER. The VGG16 [13] CNN architecture is used as the base model and combined with two different attention modules, the Squeeze and Excitation (SE) [14] and Convolution Block Attention Module (CBAM) [15]. The Oulu-CASIA [16] database is used to train and validate the performance of the different models and two other emotion datasets of FACES [17] and Japanese Female Facial Expressions (JAFFE) [18] are used for testing generalization capabilities. The generated data is statistically analysed to assess classification accuracy and a similarity metric was used to evaluate the regions of focus of the models decision process.

The aim of this study is to show that using pre-trained weights as an initial starting point for training a new model remains biased after training and does not generalize well to unseen data of the new trained model.

# 2 System Description

As a first step image pre-processing was performed on the datasets according to previous work in [5], in order to reduce background noise and focus on the face of the individual. The OULU-CASIA [16] database was used for training and validation. The training was performed with 10-fold cross validation scheme so that all the images can be used for training and validation. Some images from FACES [17] and JAFFE [18] were excluded from the study as they had an extra emotion class not available in OULU-CASIA [16].

The processed images were then trained on three models. The first was the base model of VGG16 [13] architecture, since a deep representation depth is considered important in classification results [13]. The second was the attention module of SE [14] in combination with the base model, and the third was the attention module of CBAM [15] in combination with the base model. Each of the three models were trained once using the pre-trained parameters of the base model, and once with an initial start for the parameters described further in the next section. The transfer learning utilized the pre-trained weights of the model trained on ImageNet dataset.

## 2.1 Attention Modules Implementation

The attention modules have shown improvement in CNN representations by acquiring spatial correlations between the features [14]. The SE block was introduced in [14] to improve the performance of neural networks by computing the interdependencies between channels. The CBAM was developed by [15] to extend the work of SE blocks to focus on the Spatial as well as Channel information, stating that the Spatial attention is important in deciding where the network must focus [15].

The attention modules were introduced into the base model after the second, third and fourth convolutional blocks of the VGG16 [13] architecture. The parameters of the models were randomly initialized with the “Glorot” [19] method. The models were trained with 150 epochs each, a mini batch size of 50, at a constant learning rate of 1e-4, shuffling after every epoch and a stochastic gradient descent with momentum (SGDM) optimization function with 0.9 as momentum.

## 2.2 Performance Criteria

The models were trained on the OULU-CASIA database with a 10-fold cross validation. The accuracy data collected from the True Positive (TP) predictions of each model were tabulated and the statistical results of mean, and standard deviation (SD) analysed. The data was also plotted as a Boxplot to showcase any outliers that are 1.5 times greater than that of the inter quartile range. The mean of the TP predictions from the validation sets were used as the performance metric of the model [5].

In order to visualize where the network was focusing its decisions on, the visualization technique of Gradient-weighted Class Activation Mapping (Grad-CAM) [20] was used. The visualizations of each of the images from the OULU-CASIA validation sets, FACES and JAFFE were extracted and then averaged along each class. The mean of the DICE [21] and IoU [22] metrics, across all classes, was used for comparing the generalization performance.

## 2.3 Database Description

The Oulu-CASIA database consists of image sequences from 80 different subjects expressing 6 basic emotions of Anger, Disgust, Fear, Happiness, Sadness and Surprise [16]. The image sequences of the original RGB, of visible light with

strong illumination lighting were selected for this study. The dataset selected consisted of 10,379 images in total [5].

The Japanese Female Facial Expressions (JAFFE) [18] database is composed of 213 facial portrait images of 10 different Japanese female students in grey scale expressing 7 emotions (the six basic emotions plus the Neutral). The FACES [17] dataset is made of images of facial portraits from varying subject ages. It contains a total of 2,052 images expressing 6 emotion classes of 5 of the six basic emotions excluding Surprise class and an additional Neutral class.

# 3 Results & Discussions

## 3.1 Image Data Selection and Analysis

**Table 1:** Image Distribution of each Dataset into Classes before and after 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 |

Table 1 shows the distribution of the images of each dataset into each class before and after image pre-processing. After performing the image pre-processing, the process excluded 10.41%, 8.77% and 1.41% of the images in OULU-CASIA, FACES, and JAFFE datasets respectively from further processing due to the inability of the algorithm to correctly segment the regions of interest. After which the images of the Neutral class were removed to have a fair comparison of the results. The remaining images in the datasets were 9299, 1542, and 180 for OULU-CASIA, FACES and JAFFE datasets respectively.

## 3.2 Model Performance

The training time and convergence for the different models varied between the transfer learning and the scratch training as well as between the base and attention models. The time for convergence of the transfer learning was an average 12 hours per training fold. While that of the training from scratch was an average 24 hours per fold. The models converged at around the 120 epoch mark for the training from scratch while that of the transfer learning started converging at around 30 epochs.

**Table 2:** Performance Results of the three models of Base, Squeeze & Excitation (SE) and Convolution Block Attention Module (CBAM) at each training method.

| | Transfer Learning | | | Train from Scratch | | |
|---|---|---|---|---|---|---|
| **Class** | **SE** | **CBAM** | **Base** | **SE** | **CBAM** | **Base** |
| OULU-CASIA | 99.11 ± 0.42 | 99.27 ± 0.16 | 99.49 ± 0.17 | 98.39 ± 0.27 | 98.60 ± 0.34 | 98.96 ± 0.24 |
| FACES | 42.84 ± 6.40 | 47.82 ± 7.98 | 43.93 ± 5.47 | 52.43 ± 3.87 | 50.24 ± 6.74 | 54.75 ± 4.43 |
| JAFFE | 26.33 ± 5.59 | 30.06 ± 6.23 | 32.94 ± 4.82 | 30.44 ± 1.41 | 33.56 ± 4.26 | 29.72 ± 6.84 |

Table 2 shows the performance results of the OULU-CASIA, FACES and JAFFE datasets. As can be seen the differences in terms of validation accuracy of each model were minimal with less than 0.40% difference between the models in OULU-CASIA. The training from scratch showed slightly lower performance than that of the transfer learning which was to be expected, since the training set is small compared to that of the one used in training the initial weighs from Transfer learning, therefore this reduces the ability of generating adequate filters for feature extraction.

The difference is noticed between the accuracies of the OULU-CASIA validation set and the FACES and JAFFE dataset predictions. This implies that the model is not generalizing well to unseen data. The main focus of this study is the comparison between training from scratch and transfer

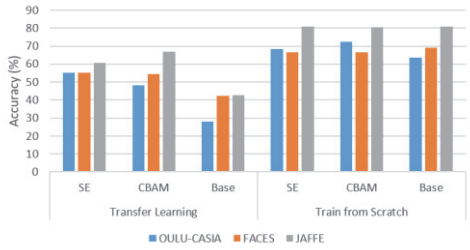


**Figure 1:** Mean DICE similarity accuracies of the three different datasets of OULU-CASIA, FACES and JAFFE.

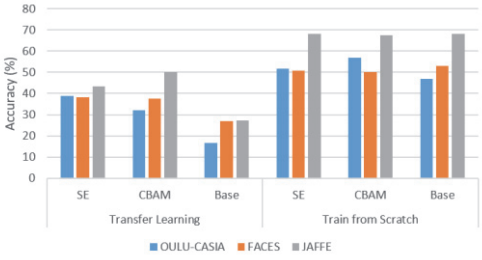


**Figure 2:** Mean IoU similarity accuracies of the three different datasets of OULU-CASIA, FACES and JAFFE.

learning on the outcome. As can be seen, the models trained from scratch had better robustness than the transfer learning with an average of 7.61% and 1.47% for the FACES and JAFFE datasets respectively. To get a clear perspective into the areas of the decision process, the similarity scores were evaluated according to the visualizations computed.

Figs. 1 and 2 represent the accuracies for the different datasets of the mean of the DICE and IoU metrics respectively. As can be seen from Fig. 1 the accuracies of the models trained from scratch outperform those that were trained using transfer learning in all the datasets with a difference in some models greater than 15%. The same analysis can be seen in Fig. 2 of the IoU score with values greater than 10% in most models. The data from the similarity metric coefficients strengthen the statement that transfer learning results are misleading.

# 4 Conclusion

This study highlights the vulnerability of accepting results without exploring further the decisions made by a neural network. The training process is shrouded behind a cloak of darkness that makes decisions without having the ability to follow or recreate the exact process. The findings showed that transfer learning was able to extract features that produced accuracies of up to 99% and performed better than the scratch training. However, the method did not generalize well to unseen data from different sources and was unable to influence the weights for the new data in the early layers of the network, suggesting that it was still biased to the previously trained data. This leads to the conclusion that although prior knowledge, i.e. pre-trained weights, is important in network training, the information will still be biased and may be unusable in special applications.

### 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.

## References

[1] K. Grifantini, ‘Detecting Faces, Saving Lives’, IEEE Pulse, vol. 11, no. 2, pp. 2–7, Mar. 2020

[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, May 2001

[3] L. Rylaarsdam and A. Guemez-Gamboa, ‘Genetic Causes and Modifiers of Autism Spectrum Disorder’, Front. Cell. Neurosci., vol. 13, p. 385, 2019

[4] J. Lugo-Marín et al., ‘COVID-19 pandemic effects in people with Autism Spectrum Disorder and their caregivers: Evaluation of social distancing and lockdown impact on mental health and general status’, Res. Autism Spectr. Disord., vol. 83, p. 101757, May 2021

[5] H. Arabian, V. Wagner-Hartl, J. Geoffrey Chase, and K. Möller, ‘Facial Emotion Recognition Focused on Descriptive Region Segmentation’, in 2021 43rd Annual Int. Conf. of the IEEE Engineering in Medicine Biology Society (EMBC). 2021

[6] 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

[7] M. Leo et al., ‘Automatic Emotion Recognition in Robot-Children Interaction for ASD Treatment’, in 2015 IEEE Int. Conf. on Computer Vision Workshop (ICCVW), Dec. 2015

[8] 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 2019

[9] A. Mehrabian, ‘Communication without words’, in Communication Theory, C.D.Mortensen, Ed.Routledge, 2017

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[11] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, ‘A Survey on Deep Transfer Learning’, ArXiv180801974 Cs Stat, Aug. 2018, Accessed: Jan. 17, 2022.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ‘ImageNet: A large-scale hierarchical image database’, in 2009 IEEE Conf. on Computer Vision and Pattern Recognition, Jun. 2009, pp. 248–255.

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

[14] J. Hu, L. Shen, and G. Sun, ‘Squeeze-and-Excitation Networks’, p. 10.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ‘CBAM: Convolutional Block Attention Module’, in Computer Vision – ECCV 2018, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19

[16] 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,

[17] 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

[18] M. J. Lyons, M. Kamachi, and J. Gyoba, ‘Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)’, 2020

[19] X. Glorot and Y. Bengio, ‘Understanding the difficulty of training deep feedforward neural networks’, p. 8.

[20] 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

[21] L. R. Dice, ‘Measures of the Amount of Ecologic Association Between Species’, Ecology, vol. 26, no. 3, pp. 297–302, 1945

[22] P. Jaccard, ‘The Distribution of the Flora in the Alpine Zone.1’, New Phytol., vol. 11, no. 2, pp. 37–50, 1912