Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230222

Interpretability study of transfer learning for emotion classification

Yuting Shao

Department of Computing and Information Sciences, Rochester Institute of Technology, 14623, NY, US

ys2884@g.rit.edu

Abstract. Transfer learning is a powerful technique for improving the performance of deep neural networks on a target task with limited labeled data. However, it is not always clear why pre-trained models on unrelated tasks, such as ImageNet, can be effective for emotion recognition tasks. This paper evaluates the performance of transfer learning for facial expression recognition tasks and investigates the influence of network depth on performance using three different typical pre-trained models: VGG16, MobileNet, and ResNet50. The layers of the pre-trained models are frozen, and their output layer is substituted with a new output layer that corresponds to the seven-class emotion classes in the target dataset. The study demonstrates that transfer learning with pre-trained models, even though with limited labelled dataset regarding facial expressions, can be an effective approach for improving the performance of deep neural networks on emotion recognition tasks. The findings provide insights into the interpretability of transfer learning for facial expression recognition tasks and highlight the importance of fine-tuning pre-trained models for optimal performance. Additionally, the study also offers ideas for future model design in facial expression recognition tasks.

Keywords: Transfer Learning, Convolutional neural network (CNN), Emotion Classification, ImageNet.

1. Introduction

Facial expressions allow humans to convey a range of emotions and nonverbal cues that are essential for effective communication, social coordination, and interaction [1]. Among the elements related to emotion recognition, facial expression analysis has been a crucial research area. Studies have identified a range of facial expressions that are associated with specific emotions, such as happiness, sadness, anger, fear, surprise, and disgust [2]. In recent times, facial expression analysis has garnered significant attention in various applications, such as human-computer interaction [3], virtual reality, gaming [4], and social robots [5], to enhance the communication and emotional interaction between humans and machines. Furthermore, it also has exceptional performance in application in the healthcare, education, and marketing industries [6]. The technology of emotion recognition using facial expressions has been widely used. However, the complexity and types of facial expressions make recognizing facial expressions a challenging topic. Traditional methods for emotion recognition have used handcrafted features or shallow learning but generally with low efficiency and recognition limitations [7]. In contrast,

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230222

deep learning based on the Convolutional Neural Network (CNN) has presented great potential in the study of face and object recognition [8].

Training a CNN from scratch can be challenging, this typically requires a large amount of labeled dataset to train the network to recognize patterns in the data, and a significant number of computational resources [9]. Given the broad range of scenarios encountered in practical applications of computer-based facial expression recognition, collecting and labeling such a dataset is time-consuming and expensive [10]. Furthermore, as different emotional-image domains exhibit different image styles, CNN scalability is limited. Transfer learning based on large dataset such as ImageNet provide a power solution for a variety of computer vision tasks, including emotion classification [2]. Akhand et al. proposed a method based on transfer learning of pre-trained AlexNet architecture on the ImageNet dataset for FER, which reached an accuracy of 70.52% [11]. Similarly, Khan et al. applied MobileNetV2-based transfer learning to improve the performance of automatic facial emotion recognition systems, which achieved 98.70% accuracy [12]. While ImageNet is not specifically designed for emotion classification, transfer learning from it to expression recognition can achieve good accuracy with limited data and parameter modifications. However, in these studies, the interpretability of transfer learning in facial expression classification tasks has not been sufficiently analysed.

This paper aims to evaluate the performance of the transfer learning for emotion classification tasks and the influence of network depth, to further provide a design idea for future model construction for facial expression recognition. FER2013 dataset [13], a popular benchmark dataset for evaluating facial expression recognition algorithms that involve seven emotional states be used in this work to retrain and evaluate all models under the same parameter settings. This paper analyses the performance between the model pre-trained on ImageNet and the model without pretraining under three network models with different depths, VGG16 [14], MobileNet [15], and ResNet50 [9]. The result shows that transfer learning performed good ability for facial expression recognition by leveraging the knowledge learned from pre-trained models in other fields. In addition, the models fine-tuned by re-adjusting the training of the target classification can significantly improve their performance, even surpassing models trained specifically on emotion datasets, and more apparent the improvement with the deeper the layers.

2. Methodology

2.1. Data description and preprocessing

This study employs the FER2013 dataset, which is a vast, openly accessible database of facial expressions. The dataset contains unconstrained images collected through the Google image search API, which have been registered and resized to 48 × 48 pixels. The dataset comprises 35,887 grayscale images, with 28,709 images allocated for training, 3, 589 for validation, and 3, 589 for testing purposes. The images are classified into seven categories based on the emotions shown in Figure 1 (left) with their corresponding distribution (Figure 1. right).



Figure 1. Seven facial expressions sample (left) and distribution (right) of the FER2013 dataset.

DOI: 10.54254/2755-2721/8/20230222

The input size of the model is 48×48×3 RGB facial expression images, and this paper have pre-processed all training and validation images by normalizing the pixel values to [0, 1]. By normalizing the pixel values, the model can be trained to better recognize patterns and features in the images, also reduce the impact of noise and improve the robustness of the model's predictions. This can help to improve the accuracy and consistency of the model's predictions.

2.2. CNN Models

Transfer learning serves as an approach that utilizes the knowledge and skills acquired from solving a particular problem and applying them to address another related problem [16]. By using the learned parameters from a pre-trained model, the new model can start with a better initialization than random initialization, which can result in faster convergence during training and better performance on the new task as the pre-trained model has already learned rich feature representations from a large dataset such as ImageNet [17]. While facial expression recognition and classification is a challenging task in computer vision due to variations in facial features and lighting conditions [13], transfer learning has demonstrated tremendous potential to overcome the problem in this domain [11, 12, 18, 19]. By leveraging the learned features and parameters of pre-trained models, transfer learning can generate the models with acceptable accuracy via training through limited facial expression images. To evaluate the performance of studied models, this paper compares the accuracy and efficiency between the pre-trained networks utilizing ImageNet weight parameters and the network models trained on FER2013 without transfer learning.

ImageNet is a large and diverse dataset, containing over a million images with 224 × 224 × 3 resolution across 1,000 categories [17]. While ImageNet is not specifically designed for facial expression recognition or emotion classification tasks, it does contain some images of facial expressions in its dataset [20]. However, the number of such images is limited compared to emotion-specific datasets such as FER2013, CK+, and JAFFE. Moreover, the facial expression images in ImageNet are often not labeled with the specific emotion or expression depicted in the image [20], which can make it challenging to use these images for training and evaluation of facial expression recognition models. To verify the effectiveness of transfer learning models based on small dataset, this paper used feature extraction ability learned from ImageNet as a starting point for facial expression classification task. Then retrained the pre-trained models on specific facial emotion dataset enable them can adapt to the new task. The process of conducting transfer learning utilizing pre-trained deep convolutional neural networks for emotion classification is depicted in Figure 2.

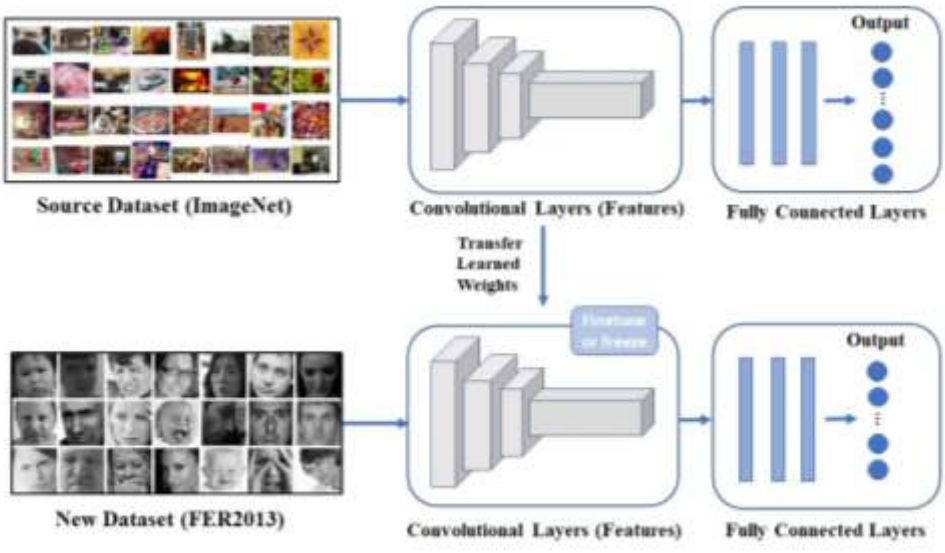


Figure 2. Transfer learning process form ImageNet to emotion classification.

DOI: 10.54254/2755-2721/8/20230222

This paper uses three convolutional neural networks with different network depths and pretrained on ImageNet dataset as based models to investigate the performance of models with and without transfer learning approach. The adopted CNN models are shown as following:

- VGG16: A convolutional neural network architecture that was introduced by the Visual Geometry Group (VGG) at the University of Oxford in 2014. It consists of 16 layers including 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers [14]. The convolutional layers have tuneable parameters that allow the network to learn rich feature representations, and the max-pooling layers help to reduce the dimensionality of the feature maps.
- MobileNet: A convolutional neural network architecture that was introduced by Google in 2017. The network consists of 28 convolutional layers and one fully connected layer followed by a softmax layer. In addition, a batch normalization and ReLU is applied after convolution [15].
- ResNet50: A convolutional neural network architecture that is 50 layers deep has 48 Convolution layers along with one MaxPool and one Average Pool layer, which introduced by Microsoft Research in 2015 as part of the ResNet family of models. The ResNet-50 architecture consists of multiple layers of convolutional and pooling layers, with skip connections that allow the network to learn residual functions [9].

2.3. Implementation details

The networks used in this paper invoke from TensorFlow via Keras API. Keras provides several pre-trained models that can be used for transfer learning and allows users to customize various aspects of the network, such as the optimizer, loss function, and activation functions. By modifying the parameter of the model weights, models can load the ImageNet weights (transfer learning) or use the weights that be initialized randomly (without transfer learning). In the context of facial expression recognition, the output layer of all three models is replaced with a new output layer that corresponds to the specific emotion classes being classified, which is a new seven-class SoftMax layer. In this study, transfer learning was performed on VGG16, MobileNet, and ResNet50 by freezing all trainable layers except the output layer. This means that the gradients are not propagated through these layers during backpropagation, and their weights are not updated during training. Instead, only the weights of the output layer are trained on the emotion dataset. During the training process, models are trained with 128 batch size and 250 epochs, using rectified Adam optimizer (learning rate = 0.001) and categorical crossentropy as loss function. An early stopping mechanism utilizes training accuracy as a monitoring indicator also implemented while training the proposed model to prevent the model from overfitting. The training process will automatically stop if the training accuracy does not increase for five consecutive epochs.

3. Results and Discussion

Table 1. Performance of pretrained model (freeze), pretrained model (finetune), and unpretrained models with three different depths

Performance	Pretrained Model (freeze)			Pretrained Model (finetune)			Unpretrained Model		
	VGG16	Mobile Net	ResNet 50	VGG16	Mobile Net	ResNet 50	VGG16	Mobile Net	ResNet 50
Training Accuracy	0.4293	0.3595	0.5882	0. 9086	0.9164	0.9077	0.9151	0.9034	0.8469
Testing Accuracy	0.4289	0.3650	0.4532	0.6382	0.6230	0.6401	0.6500	0.6166	0.5933
Training Loss	1.4778	1.6207	1.0926	0. 2666	0.2296	0.2169	0.2424	0.272	0.4168
Testing Loss	1.4987	1.6155	1.6826	0.8066	0.8245	0.7966	0.6871	0.8837	0.8120
Early Stop (epoch)	39	22	19	116	114	87	187	228	155

According to the results of Table 1, all three un-pretrained models (VGG16, MobileNet, and ResNet50) achieved considerable performance in facial expression recognition when trained on the FER2013 dataset with random initialization. Specifically, VGG16 had the highest testing accuracy of 65% and the lowest testing loss of 68.71%, despite having the fewest number of layers among the three

DOI: 10.54254/2755-2721/8/20230222

models. MobileNet and ResNet50 had similar performance in emotion classification, achieving around 60% testing accuracy, but ResNet50 achieved this in fewer epochs. The good performance of the un-pretrained models may be due to the FER2013 is the facial specialized datasets that have been specifically annotated with emotional labels, which means these models have learned to extract features that are more relevant to the emotion classification task.

In this study, transfer learning was performed on VGG16, MobileNet, and ResNet50 by freezing all layers except the output layer. As shown in Table 1, the performance of these models was good to some extent, with ResNet50 achieving the highest accuracy of 45.32%. However, the performance of the pre-trained models was still not as good as the models trained directly on the FER2013 dataset. As shown in the Figure 3, the training accuracy and training loss of the un-pretrained models that trained on FER2013 dataset are steadily increased and decreased respectively, means all modes are continuously learning. However, the performance of transfer leaning models based on ImageNet remains the same after the initial learning. One possible reason for this is that the pre-trained models may not have learned features that are optimal for facial expression recognition on the FER2013 dataset. The ImageNet dataset used for pretraining is not specifically designed for facial expression recognition and may not contain enough emotion-related images to provide useful pre-trained weights for this task. Additionally, freezing all layers except the output layer and any newly added layers during transfer learning may not be sufficient to adapt the pre-trained weights to the specific characteristics of the FER2013 dataset.

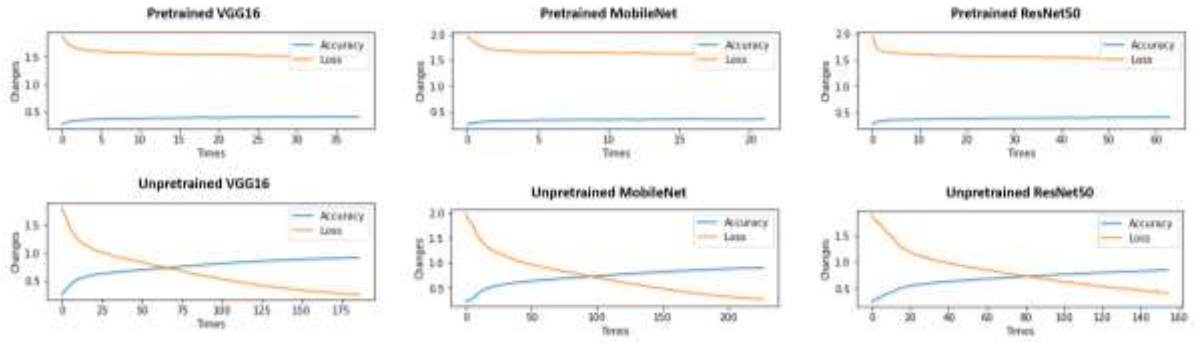


Figure 3. Training accuracy and loss curves for the un-pretrained models trained on the FER2013 dataset and the models trained using transfer learning from ImageNet.

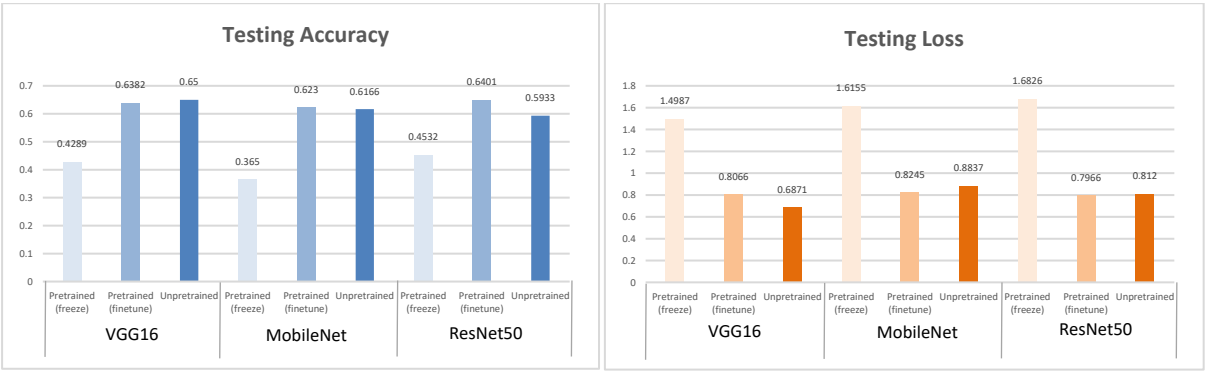


Figure 4. Performance comparison among transfer learning models with and without finetune and the models without transfer learning.

To further explore the performance of transfer learning, this paper finetunes the pre-trained models in transfer learning and retraining them on the new dataset. As a result, in Figure 4, the finetuned models displayed significant improvement compared to original transfer learning models and showed a similar and even better performance compared to un-pretrained models but with faster training times. As shown

DOI: 10.54254/2755-2721/8/20230222

in Figure 5, all finetuned transfer learning models cost almost half the time of their relevant models trained from scratch. In addition, the higher the number of layers, the more pronounced the improvement.

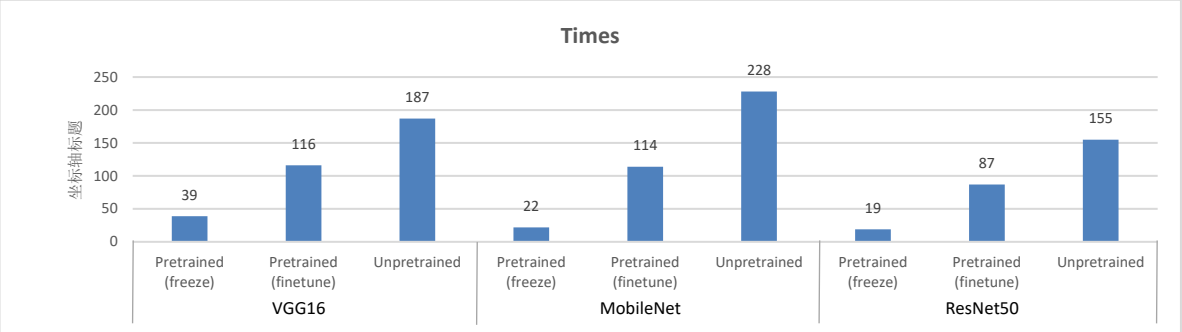


Figure 5. Training time comparison among transfer learning models with and without finetune and the models without transfer learning.

These results suggest that even without pretraining on a large dataset such as ImageNet, these deep neural network architectures can still learn effective representations for facial expression recognition when trained on the FER2013 dataset. In addition. it appears that fine-tuning the pre-trained models resulted in significant performance improvements compared to the original transfer learning models. Furthermore, the fine-tuned models demonstrated similar or even better performance compared to models trained from scratch on the target dataset. This demonstrates the effectiveness of transfer learning and fine-tuning pre-trained models in improving the performance of machine learning models on specific tasks. In addition to the performance improvements, another advantage of fine-tuning pre-trained models in transfer learning is that it can often result in faster training times compared to models without transfer learning, whcih may contribute to the pre-trained model that has already learned valuable representations of the data, which can be leveraged to speed up the learning process for the new task.

4. Conclusion

This paper systematically analyses the interpretability of transfer learning for expression classification with three selected neural network models in three different depths. This study compared the performance among three transfer learning models with and without finetune and the models without transfer learning under the same parameters and facial expression dataset. The results suggests pre-trained models on ImageNet can get considerable performance on emotion classification task, it is often more effective to use models that are specifically trained on emotion datasets. However, fine-tuning pre-trained models, such as those pre-trained on ImageNet, can still lead to significant performance improvements in emotion classification tasks. Fine-tuning allows the pre-trained model to adapt its learned representations to the specific task of emotion classification, while still leveraging the valuable features learned during pre-training. In some cases, fine-tuning a pre-trained model may even outperform models specifically trained on emotion datasets and generally have higher training efficiency, which influenced by the depths of the model layers. Overall, pre-trained models on ImageNet can be a powerful tool in emotion analysis tasks, as they are able to extract rich features from images that are useful for identifying emotions. However, it is important to note that the effectiveness of these models may depend on various factors such as the size of the available labeled data, the similarity of the new task to the pre-training task, and the specific requirements of the application. Researchers and practitioners should carefully consider these factors when choosing a model for emotion classification tasks.

References

[1] Katembu S Xu Q Rostami H N Recio G and Sommer W 2022 Effects of Social Context on

DOI: 10.54254/2755-2721/8/20230222

Deliberate Facial Expressions: Evidence from a Stroop-like Task Journal of Nonverbal Behavior 46 3 (Sep. 2022) 247–267
[2] Deng J Dong W Socher R Li L.-J Li K and Fei-Fei L 2009 ImageNet: A large-scale hierarchical image database 2009 IEEE Conference on Computer Vision and Pattern Recognition (Jun. 2009) 248–255
[3] Wang H Gaddy V Beveridge J R and Ortega F R 2021 Building an Emotionally Responsive Avatar with Dynamic Facial Expressions in Human—Computer Interactions. Multimodal Technologies and Interaction 5 3 13
[4] Peeters D 2019 Virtual reality: A game-changing method for the language sciences Psychonomic Bulletin & Review 26 3 (Jun. 2019) 894–900
[5] Ramis S Buades J M and Perales F J 2020 Using a Social Robot to Evaluate Facial Expressions in the Wild Sensors 20 23 6716
[6] Bălan O Moise G Petrescu L Moldoveanu A Leordeanu M and Moldoveanu F 2020 Emotion Classification Based on Biophysical Signals and Machine Learning Techniques Symmetry. 12 1 (Jan. 2020) 21
[7] Step by step VGG16 implementation in Keras for beginners: 2020 https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c. Accessed: 2023-01-13.
[8] Appasaheb Borgalli R and Surve S 2022 Deep Learning Framework for Facial Emotion Recognition using CNN Architectures 2022 International Conference on Electronics and Renewable Systems (ICEARS) 1777–1784
[9] He K Zhang X Ren S and Sun J 2016 Deep Residual Learning for Image Recognition 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778
[10] McDuff D Kaliouby R Senechal T Amr M Cohn J and Picard R 2013 Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected 881–888
[11] Akhand M. a. H Roy S Siddique N Kamal M.A.S and Shimamura T 2021 Facial Emotion Recognition Using Transfer Learning in the Deep CNN Electronics 10, 9 (Jan. 2021) 1036
[12] Khan M U Abbasi MA Saeed Z Asif M Raza A and Urooj U 2021 Deep learning based Intelligent Emotion Recognition and Classification System 2021 International Conference on Frontiers of Information Technology (FIT) 25–30
[13] Li S and Deng W 2022 Deep Facial Expression Recognition: A Survey IEEE Transactions on Affective Computing 13 3 2022 1195–1215
[14] Simonyan K and Zisserman A 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv
[15] Howard A G Zhu M Chen B Kalenichenko D Wang W Weyand T Andreetto M and Adam H 2017 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications arXiv
[16] Savchenko A V Savchenko L V and Makarov I 2022 Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network IEEE Transactions on Affective Computing 13 4 2132–2143
[17] George Karimpanal T and Bouffanais R 2019 Self-organizing maps for storage and transfer of knowledge in reinforcement learning Adaptive Behavior - Animals Animats Software Agents Robots Adaptive Systems 27 2 (Apr. 2019) 111–126
[18] Qi C Li M Wang Q Zhang H Xing J Gao Z and Zhang H 2018 Facial Expressions Recognition Based on Cognition and Mapped Binary Patterns IEEE Access 6 18795–18803
[19] Ramalingam S and Garzia F 2018 Facial Expression Recognition using Transfer Learning 2018 International Carnahan Conference on Security Technology (ICCST) 1–5
[20] Yang K Yau J Fei-Fei L Deng J and Russakovsky O 2022 A Study of Face Obfuscation in ImageNet arXiv