Effect of the deep residual networks at different depths on expression recognition

Yufeng Pan

Institute of International Education, Hebei University of Technology, Tianjin, 300000, China

Yufeng.Pan@student.lut.fi

Abstract. Facial expression recognition plays a critical role in numerous applications like emotion analysis, human-computer interaction, and surveillance systems. Given the importance of this task, this study aims to investigate the effectiveness of different depths of Residual Networks (ResNet). The primary objective is to scrutinize and compare these ResNet models in terms of their training and validation losses and performance metrics like accuracy, recall, and F1 scores. In this research, a thorough comparative analysis is conducted by setting up exhaustive experiments using these models. The experiment is carried out on a popular facial expression dataset. Despite the depth differences, ResNet101 emerged as the model demonstrating superior performance. It struck the most effective balance between model complexity and generalization capacity, leading to the lowest validation loss and better performance. Experimental results show that a more complex model does not necessarily yield optimal results. The optimal balance between model complexity and generalisation needs to be investigated. These findings can provide essential guidance in the design of deep learning models for facial expression recognition and other similar tasks.

Keywords: residual networks, facial expression recognition, model complexity, generalization, deep learning.

1. Introduction

Facial expression recognition in AI is crucial for applications like interaction, social media, surveillance, and mental health [1]. It encapsulates the essence of non-verbal communication, playing a crucial role in understanding human emotions and intentions. Facial expressions are dynamic, multi-modal signals that inherently carry rich emotional information. Effectively extracting and interpreting these signals is a complex task, given the substantial variability in facial expressions among individuals and across cultures [2].

Conventional approaches to recognizing facial expressions were predominantly based on the extraction of manually engineered features [3,4]. Although these methods are groundbreaking when introduced, they often struggle with complex and non-linear facial expressions. Challenges rise due to their limited accuracy and insufficient generalizability, which curtails their effectiveness in delivering robust and reliable results [5]. These methods require extensive feature engineering and typically struggle to capture complex interactions between different facial regions. Following this, deep l Earning was suggested as a solution for facial expression recognition [6]. Central to deep learning are

Convolutional Neural Networks (CNNs), which are proficient at autonomously extracting high-level features from datasets [7]. However, conventional CNNs are not devoid of shortcuts. As the depth of the network increases, they often encounter predicaments like vanishing or exploding gradients. These issues pose significant hindrances to the learning process, limiting the ability to train very deep networks and thereby restricting performance improvements [8]. Moreover, they often require training examples and are prone to overfitting when trained on limited data. A major breakthrough in overcoming these limitations was the introduction of deep residual networks (ResNets) by He et al. in 2015 [9]. Residual blocks mitigate gradient disappearance or explosion by learning the residual function through reference layer in puts [10]. Yet, within the context of facial expression identification tasks, the relationship between the depth of ResNets and their effectiveness continues to be an unresolved issue. Some studies suggest that increasing network depth can improve facial expression recognition performance, but others have found that overly deep networks can result in overfitting [11]. Overfitting occurs when the model becomes too specialized in learning the training data, which can reduce its ability to generalize to new, unseen data.

Therefore, this study develops into the relationship between the depth of ResNets and their performance in facial expression recognition. The central research question involves around identifying the optimal depth for ResNets, which results in the most effective facial expression recognition. F urther, the benefits and potential Drawbacks of ResNets are analyzed. This study explores two questions: first, whether increasing the depth of ResNets improves performance. Second, whether there is an optimal depth that saturates the model, and thus discusses the reasons for the degradation of model per form beyond the optimal depth. To Address them, first, three distinct ResNet variants with differing depths are used to building facial expression recognition models including ResNet-50, ResNet-101, and ResNet-152. Second, the FER2013 dataset is used for training and evaluation ation [12]. The dataset is a publicly available, comprehensive collection of human facial expressions captured in real-world conditions. Third, the performance of ResNets of different depths on the dataset is evaluated and compared to previous neural network models. In conclusion, the st udy yields Crucial insights into the application of ResNets in facial expression recognition tasks. The experiment guides the selection of an optimal ResNet variant for specific real-world applications. Ultimately, the findings aim to enhance the design and application of more effective deep learning mod els for facial expression recognition.

2. Methodology

2.1. Dataset description and preprocessing

The experiment utilizes the FER-2013 dataset, a popular open-source dataset frequently used for training machine learning and deep learning models in facial expression recognition [12]. This dataset comprises grayscale images of faces, each 48x48 pixels. The task involves categorizing each face into one of seven emotional expression categories based on the facial expression. In total, there are 35,887 examples. During the preprocessing stage, several tasks are performed to ensure the images are ready to be utilized in the deep learning models. First, image normalisation is performed. Secondly, data augmentation methods are employed to amplify the dataset's diversity and enhance the model's capability to generalize. Specifically, variations are introduced to the original images by altering their width and height parameters, effectively generating different versions of the same image. These alterations enable the models to learn from a more diverse range of data, ultimately improving their performance on unfamiliar or unseen data. Overall, the preprocessing steps include loading the data, normalizing it, and then augmenting it for better generalization.

2.2. Proposed approach

The approach towards facial emotion recognition hinges on leveraging the power of the Residual Network (ResNet) architecture. ResNet's unique strength in mitigating the pervasive vanishing/exploding gradient problem in deep neural networks makes it a robust choice for this task. Three variations of ResNet, namely ResNet50, ResNet101, and ResNet152, are employed to compare their performances and ascertain the effect of network depth on emotion recognition accuracy. The methodology follows a well-structured sequence of steps: data preprocessing, model creation, training, and finally, evaluation.

Data Preprocessing: The dataset used for this project, FER2013, consists of grayscale images of human faces displaying various emotions. Each image is initially read from storage and converted into a proper tensor format. They are then grouped into batches for more efficient processing. To improve the model's learning capabilities, the image tensor values are rescaled from a range of 0-255 to a range of 0-1, which is a more favorable input range for neural networks. Furthermore, Data augmentation techniques, such as altering the width and height of the images, are introduced to increase data diversity and enhance model generalization.

Model Creation: Deep learning models are constructed next. Utilizing the Keras library, three distinct models are created based on the ResNet50, ResNet101, and ResNet152 architectures.Each model follows the same basic structure but differs in the number of layers, with ResNet50 containing 50 layers, ResNet101 containing 101 layers, and ResNet152 containing 152 layers.

Model Training: Every model is trained using the preprocessed dataset. Training occurs over 50 epochs. We've integrated features like early stopping and model checkpointing during the training process. Model checkpointing, on the other hand, safeguards the best performing model - the one with the lowest validation loss - by saving it.

Evaluation: Finally, each model's performance is evaluated based on its ability to recognize and categorize the various facial expressions. Their accuracy and loss on the validation set are assessed to understand their effectiveness.

2.2.1. ResNet. The innovation of Residual Networks, or ResNets, dramatically redefines the perspective on deep learning models.The concept of ResNet is the introduction of "skip connections" (or "shortcut connections"), which permit the gradients to be backpropagated to earlier layers without any modifications. Hence, the layers learn a residual mapping in reference to the layer inputs. This unique approach has led to the networks being designated as Residual Networks. A Residual Block is the standard unit in ResNet, which comprises a sequence of operations: Convolution,As depicted in Figure 1, the outcome of a residual block is achieved by incorporating the results of Batch Normalization and ReLU activation to the initial shortcut input.This structure empowers ResNet to bypass the vanishing/exploding gradient issue and learn complex representations effectively.

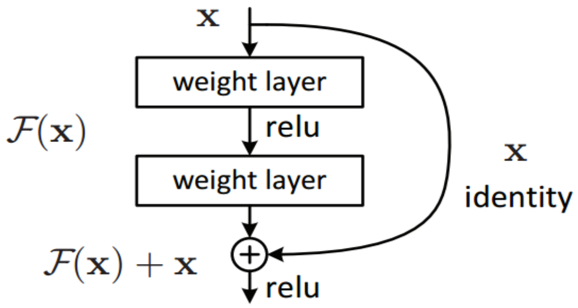
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1. Basic structure of a residual block [9].

The numbers '50', '101', and '152' in ResNet50, ResNet101, and ResNet152 signify the count of layers within each of these network versions. Although the core architecture and underlying principles remain consistent across these versions, the depth of the networks varies. A deeper network, theoretically, can capture more intricate features and thus is expected to perform better for complex tasks.

2.2.2. Linear regression. A linear regression model is employed to classify the extracted features into different emotion categories. Linear regression estimates the relationship between the input (extracted features) and output (emotion category) by fitting a linear equation to observed data. The steps to obtain the linear equation parameters are an iterative optimization process, refining the predictions at each step.

2.2.3. Loss function. In the present investigation, considering the task's character as a problem of multi-category classification, categorical cross-entropy is utilized as the loss function. Owing to its proficient capacity in quantifying the deviation between the forecasted and actual labels, categorical cross-entropy is an optimal selection for problems of multi-class classification.The categorical cross-entropy loss can be calculated using the following formula:

$$L = -\sum(y_i \times \log(y_i^l)) \tag{1}$$

Total cross-entropy loss L is what's aimed to minimize. The summation Σ runs over all classes. y_i represents the true label for class i. This would be 1 for the correct class and 0 for all other classes, one-hot encoded class labels are dealt with here. y_i^1 represents the predicted probability for class i. This value comes from the output of the model, which should ideally be close to 1 for the correct class and close to 0 for all other classes. The log function is the natural logarithm, and the multiplication and the subsequent summation over all classes are done element-wise.

2.3. Implementation details

The implementation sequence began with the creation of the deep learning models, each based on a ResNet variant - ResNet50, ResNet101, and ResNet152. These models, created using the Keras library, followed the same foundational architecture but varied in the total count of layers. In the optimization process, the capabilities of the Adam (Adaptive Moment Estimation) optimizer are leveraged. This proves advantageous in situations involving large datasets or a significant number of parameters. For the training phase, Google Colaboratory (Google Colab) is utilized for its GPU capabilities. Each model underwent a training regimen of 50 epochs with a batch size of 32, taking roughly 30 minutes per model. The use of Google Colab's resources significantly reduced the computational load, enabling a more efficient training process. In this phase, model checkpointing and an early stopping callback are employed to ensure the highest-quality model is retained and to prevent overfitting., respectively. The models are subsequently evaluated on their ability to accurately categorize facial expressions. Focus is put on observing their accuracy and loss performance on the validation set.

3. Results and discussion

This chapter analyzes the impact of different depths of residual networks on facial recognition performance from two aspects: loss curves and test set recognition performance.

3.1. Training and validation loss

The initial phase included an evaluation of the training and validation loss for each model, as outlined in Table 1.

Table 1. Training and validation loss for ResNet models.

model	Training loss	Validation loss
ResNet50	1.700592279434204	1.7599059343338013
ResNet101	1.7494916915893555	1.7231714725494385
ResNet152	1.7400091886520386	1.750921607017517

Figure 2 illustrates the training and validation loss progression for ResNet50, ResNet101, and ResNet152.

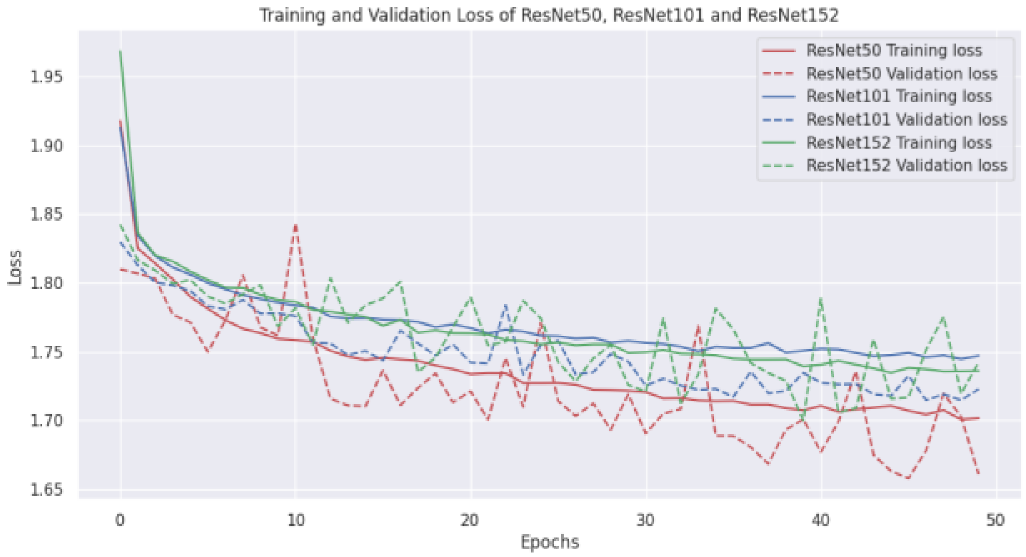

Figure 2. training and validation loss curves of ResNet50, ResNet101, and ResNet152 (Picture credit: Original).

3.2. Model performance

Performance of each model on the facial expression recognition task was evaluated next, including metrics such as accuracy. The results for these metrics can be seen in Table 2.

Table 2. Performance metrics for ResNet models.

Model	Accuracy	Macro Avg (Precision, Recall, F1-Score)	Weighted Avg (Precision, Recall, F1-Score)
ResNet50	0.18	(0.14, 0.14, 0.12)	(0.17, 0.18, 0.16)
ResNet101	0.19	(0.27, 0.14, 0.11)	(0.29, 0.19, 0.15)
ResNet152	0.22	(0.14, 0.14, 0.10)	(0.17, 0.22, 0.15)

The findings indicate that ResNet50 displayed the lowest training loss, while ResNet101 demonstrated the best generalization capacity, reflected in the lowest validation loss. This suggests that ResNet101, and not ResNet50 as might have been expected, strikes an optimal balance between model complexity and generalization. These results can be traced back to the intrinsic differences in the model architectures. For instance, ResNet50, despite being less deep, performed effectively during training, potentially due to reduced opportunities for overfitting. On the other hand, ResNet101, with its slightly larger capacity, seems to capture a more generalized representation of the features, translating to superior performance on unseen data. ResNet152, despite being the deepest network, did not necessarily guarantee superior performance. This aligns with the research question concerning the optimal network depth, beyond which performance may plateau or even deteriorate. In this context, the added depth of ResNet152 might have led to overfitting or encountered other limiting factors, causing marginally inferior performance. Considering these findings, it can be concluded that ResNet101 provides superior performance among the evaluated models on the facial expression recognition task. This conclusion underscores the significance of finding an appropriate balance between model complexity and generalization. Importantly, a highly complex model might not always ensure superior results due to risks of overfitting and the potential inability to effectively learn from the data.

4. Conclusion

This research intends to assess the efficiency of three distinct depths of Residual Networks in the realm of a task related to facial expression recognition. A detailed comparison and analysis of the models is performed. This included assessing parameters such as training and validation loss, along with performance metrics like accuracy, recall, and F1 scores. Exhaustive experiments are carried out to scrutinize the proposed method. Experimental results indicated that ResNet101, despite not being the deepest model, outperformed both ResNet50 and ResNet152 on the facial expression recognition task. ResNet101 showed the best balance between model complexity and generalization. The experimental results show that a highly complex model may not always yield optimal results. A balance between model complexity and generalisation is necessary. The nuances of ResNet architectures and their adaptability to various recognition tasks will be considered as the main research objective. In the future, the research will concentrate on analyzing and improving the overall robustness and versatility of ResNet models, with an intention to further optimize performance on a broad spectrum of complex tasks.

References

[1] Martinez A Du S 2012 A model of the perception of facial expressions of emotion by humans: research overview and perspectives Journal of Machine Learning Research 13: pp 1589-1608

[2] Zafeiriou S Kollias D Nicolaou M Papaioannou A Zhao G Kotsia I 2017 Facial affect in-the-wild Proceedings of the IEEE 105(5): pp 922-944

[3] Ekman P Rosenberg E 1997 What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System Oxford University Press USA

[4] Huang G Learned-Miller E 2014 Labeled faces in the wild: Updates and new reporting procedures Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, Tech. Rep, 14(003)

[5] Lucey P Cohn J Kanade T Saragih J Ambadar Z Matthews I 2010 The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops IEEE pp 94-101

[6] LeCun Y Bengio Y Hinton G 2015 Deep learning Nature 521(7553): pp 436-444

[7] Krizhevsky A Sutskever I Hinton G 2012 Imagenet classification with deep convolutional neural networks Advances in neural information processing systems 25: pp 1097-1105

[8] Bengio Y 2009 Learning deep architectures for AI Foundations and trends® in Machine Learning 2(1): pp 1-127

[9] He K Zhang X Ren S Sun J 2016 Deep residual learning for image recognition In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) IEEE pp 770-778

[10] Huang G Liu Z Maaten L Weinberger K 2017 Densely connected convolutional networks In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) IEEE pp 4700-4708

[11] Simonyan K Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556

[12] Zhang K Zhang Z Li Z et al 2016 Joint face detection and alignment using multitask cascaded convolutional networks IEEE signal processing letters 23(10): pp 1499-1503