Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230248

The study of performance related to classical convolutional neural networks in the field of facial emotion recognition

Yanxiao Liu

Electrical Engineering Machine Automation, Chongqing university, 400044, Chongqing, China

Liu5y8@mail.uc.edu

Abstract. Facial expression recognition is a challenging task that has received much attention in the past decade. This paper presents a rigorous examination of state-of-the-art Convolutional Neural Network (CNN) architectures, encompassing VGG, ResNet, MobileNet, and DenseNet, applied to the task of facial expression recognition using the FER2013 dataset. The study delineates the distinct advantages and disadvantages of each architecture with respect to performance, parameter efficiency, and computational complexity, thereby providing valuable insights into their suitability for specific applications. The selection of an appropriate CNN architecture for facial expression recognition is contingent upon the particular requirements and constraints inherent to the application. As such, the paper advocates for a nuanced approach in determining the most suitable architecture, taking into consideration factors such as computational resources, model complexity, and desired accuracy. In conclusion, the paper calls for future research endeavors to concentrate on exploring innovative architectural designs, optimizing training methodologies, and addressing the unique challenges that facial expression recognition poses. By advancing the state-of-the-art in CNN architectures and training techniques, researchers can contribute to further improvements in model performance and expand the range of their real-world applicability. Such advancements would not only benefit the field of computer vision, but also impact numerous practical applications, ranging from human-computer interaction to emotion analysis in various domains.

Keywords: convolutional neural network, facial emotion recognition, FER2013.

1. Introduction

Facial expressions are the result of the action of human facial muscles. Varied emotional states result in different expressions of it. Facial expressions have become a crucial component of non-verbal communication and a means of conveying information, making them a crucial tool for assessing the emotions of others. Facial information has an important position in human socialization [1]. It is also noted in human evolution that the human visual system has evidence of neurons dedicated to the analysis of faces [2], which would help humans better recognize the emotions of others for interaction and expression. Therefore, the use of computer technology to model the recognition of human facial expressions has become particularly important in the process of human emotional expression and computer interaction.

DOI: 10.54254/2755-2721/8/20230248

Currently, building models using image-based facial recognition through neural networks attracted much attention. The technique of determining the identity and even the facial emotion of a person in an image by determining the face detection within the image (i.e., based on the process of facial organ detection and alignment) is a very critical feature.

The results of facial recognition research play an important role in application areas such as surveillance technology and human-computer interaction, such as identity matching for traffic travel and smartphone facial recognition payments. Traditional facial recognition is limited to shallow learning and simple image features, which is more difficult to apply to the field of recognizing facial expressions and has low learning efficiency. Deep learning in the field of machine vision has made many remarkable achievements in research, and face recognition is one of them. Andre et al. studied the video group emotion recognition task by using VGAF dataset and proposed a Convolutional Neural Network (CNN) model that classifies by Support vector classifier (SVC) with Radial Basis Function (RBF) kernel with an accuracy of more than 60% [2]. Gheyath et al. proposed a model for static facial expression classification using the Fully Connected layer (FC) of CNN with an accuracy of 69.85% [3]. The features extracted from a deep CNN-based architecture can provide application-specific learning, applying it to perform well in both dynamic and static datasets.

Deep learning has a powerful ability to learn image facial features when applied to Facial Expression Recognition (FER), but there are still some problems in practical applications, such as the need for large amounts of noise-free training data during training, higher bias [4]. Recently, several models have proposed the use of synthetic facial images generated by Generative Adversarial Network (GAN) techniques to augment the training data and aid in the corresponding recognition task [5]. Traian et al. proposed a new system for emotion classification based on a GAN classifier that uses the TensorFlow machine learning framework [6]. Although most of the training data came from the FER 2013 dataset (85%), the system was also able to correctly classify images from other used datasets with high accuracy (e.g., 94.8% on the JAFFE dataset).

Transfer learning provides a new approach to building CNN models by transferring parameters that the model has already learned and trained to help train the new model. Kennedy Chengeta evaluated deep learning algorithms using pre-trained models against local feature-based algorithms and showed that transfer learning methods can provide higher accuracy and shorter execution times [7]. In addition, Ramalingam investigated the transfer of VGG19-based learning to a small dataset (i.e. FER2013 dataset) that achieved 10-fold cross-validation [8].

The aim of this study is to compare the performance of classical convolutional neural network models in emotion recognition. This study tested various pre-trained models, including MobileNet V1, MobileNet V2, ResNet50, VGG16, VGG32 models. Considering the time complexity and training cost, only the FER-2013 dataset was used for training and testing in this study. By comparing the testing results of the six models, it can be concluded that xxxx has better testing results and is more suitable for facial expression recognition among the classical convolutional neural network models.

2. Method

2.1. Dataset preparation

This study uses the open-source dataset called FER2013 obtain from Kaggle [9]. The Fer2013 facial expression dataset was unveiled at the International Conference on Machine Learning (ICML) in 2013, where it garnered significant attention within the scientific community. This dataset has since been widely adopted as a benchmark for evaluating the performance of facial expression recognition models, and was also employed as the primary dataset for the prestigious 2013 Kaggle face recognition competition. The Fer2013 dataset comprises 35,886 high-quality images of facial expressions, which are divided into three distinct subsets: a training set, a public testing set, and a private testing set. Specifically, Each image is a fixed-size 48×48 grayscale image. The data set shares seven labels: 0 anger; 1 disgust; 2 fear; 3 happy; 4 sad; 5 surprised; 6 normal. To avoid overfitting and obtain satisfactory

DOI: 10.54254/2755-2721/8/20230248

performance, this paper considered using four augmentation methods, namely Translation, Rotation, Flip, and Zoom, to enhance the data set.

2.2. CNN Models

The Visual Geometry Group at the University of Oxford proposed the VGG network architecture which was introduced during the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [3]. The fundamental objective of their work was to demonstrate that an increase in network depth could potentially enhance the network's final performance to a certain degree. This was accomplished through the development of two network structures, VGG16 and VGG19, which have since become prominent models in the field of computer vision. The two have no essential difference, but the network depth is different. In VGG, a more minor 3x3 convolution filter architecture is used, effectively reducing the total amount of parameters compared to 5x5 convolution and 7x7 convolution filters. While deepening the network, it is shown that an improvement in CNN performance can be achieved by pushing the depth to 16-19 weight layers.

Residual networks (ResNets) have emerged as a promising approach to achieve state-of-the-art performance in image classification tasks and enable the training of very deep neural networks [4, 10, 11]. ResNets utilize identity shortcut connections, similar to those employed in highway networks, to facilitate information flow across multiple layers without attenuation, leading to enhanced optimization. However, despite their remarkable empirical performance, ResNets have some limitations, including accumulating multiple levels of feature representations at each layer, which may need to be more helpful in deeper layers. Additionally, the ResNet architecture hypothesizes that learning identity weights is challenging. Learning the additive inverse of identity weights needed to eliminate information from the representation at any given layer is equally challenging. Moreover, the fixed-size layer structure of residual block modules necessitates that residuals are learned by shallow subnetworks, despite evidence that deeper networks offer better expressivity.

MoblieNetV2 and MoblieNetV3 are two notable variations of the MobileNet architecture designed to address image classification challenges in resource-constrained settings [5]. These models incorporate several novel architectural features, includiLike highway networks, and residual networks use identity shortcut connections that enable the flow of information across layers without attenuation y of the network.

Depthwise separable convolutions replace the traditional convolutional layers in MobileNet with depthwise convolutions, followed by pointwise convolutions, which results in significant parameter reduction and computational savings. In addition, linear bottlenecks further enhance the model's efficiency by reducing the number of channels in the intermediate layers. At the same time, inverted residuals facilitate the flow of information across the network, allowing for faster convergence.

MoblieNetV2 and MoblieNetV3 have been extensively evaluated on large-scale datasets, such as ImageNet, and have demonstrated state-of-the-art performance in image classification tasks while being highly optimized for resource-constrained devices.

DenseNet, short for "Dense Convolutional Network," is a deep neural network architecture proposed by Huang et al. in 2017 [6]. The DenseNet architecture differs from traditional neural network architectures in that it employs densely connected layers, which allows information to be shared more efficiently between layers. Specifically, each layer receives input from all previous layers and its previous layer. This dense connectivity pattern results in shorter paths for feature propagation, facilitating the reuse of features and enabling the network to propagate information throughout the network more efficiently.

DenseNet comprises several dense blocks, each consisting of multiple convolutional layers with the same number of output feature maps. The output of each dense block is then passed through a transition layer, which performs downsampling of the spatial dimensions and compression of feature maps. The transition layer comprises a convolutional layer followed by a pooling layer and a batch normalization layer. The pooling layer reduces the spatial dimension of the feature maps, while the batch normalization layer normalizes the feature maps across the channel dimension. DenseNet also includes bottleneck

DOI: 10.54254/2755-2721/8/20230248

layers, which reduce the computational cost of the network by using 1x1 convolutions to reduce the number of input channels before the 3x3 convolutions are applied. This reduces the number of parameters in the network, making it more efficient and easier to train.

2.3. Implementation details

All networks are tested using the PyTorch framework, with 25 training iterations and a batch size of 64. CrossEntropyLoss is employed as the loss function, which calculates the cross-entropy loss between the probability distributions of each class. The optimizer sets the learning rate at 0.001 and the momentum at 0.9. The learning rate is adjusted using the StepLR strategy, which reduces the learning rate by one order of magnitude every seven epochs. Data preprocessing is carried out by normalizing each channel of the image, incorporating a random (50% probability) horizontal flip with a random rotation of -10 to 10 degrees for data augmentation, and additional input resizing for certain models that demand supplementary input images.

3. Result and discussion

Table 1. Experimental results of 5 networks based on FER 2013 dataset.

Model name	Test Set Accuracy
VGG16	60.70%
MobileNetV2	60.30%
MobileNetV3	52.24%
ResNet50	65.39%
DenseNet	64.34%

This study trained five distinct neural networks, VGG16, MobileNetV2, MobileNetV3, ResNet50, and DenseNet, on the FER2013 dataset and assessed their accuracy on the validation set. The respective accuracies are 60.7%, 60.3%, 52.24%, 65.39%, and 64.34% shown in Table 1. An analysis and discussion of the experimental results in academic and concise language follows:

VGG16 (60.7%) and MobileNetV2 (60.3%) manifested moderate accuracies. Although these networks exhibit disparate architectural properties, their analogous performance implies that neither may represent the most efficacious solution for the FER2013 dataset. Notwithstanding, they remain plausible alternatives for the facial emotion recognition domain.

MobileNetV3 (52.24%) yielded the lowest accuracy among the assessed networks. This outcome can be ascribed to the network's design principles, which emphasize reduced computational overhead and diminished model size, thereby compromising accuracy. Consequently, the trade-off between computational efficiency and performance appears suboptimal for the FER2013 dataset.

ResNet50 (65.39%) and DenseNet (64.34%) demonstrated the most superior accuracies. This elevated performance can be attributed to their distinct architectural innovations: ResNet50 incorporates residual connections to enable efficient gradient propagation, while DenseNet's dense connectivity pattern fosters effective feature reuse. As a result, these architectures appear particularly well-adapted for the FER2013 dataset.

In summation, ResNet50 and DenseNet emerge as the most efficacious deep neural network architectures for facial emotion recognition tasks employing the FER2013 dataset. Conversely, VGG16 and MobileNetV2 provide moderate performance, and MobileNetV3, notwithstanding its computational efficiency, may not represent the most suitable choice for this dataset due to its diminished accuracy. Future investigations may benefit from a more thorough exploration of model hyperparameters, data augmentation methodologies, and the incorporation of additional architectures to enhance recognition performance.

DOI: 10.54254/2755-2721/8/20230248

4. Conclusion

In summary, this paper provides a comprehensive investigation of various state-of-the-art CNN architectures, including VGG, ResNets, MobileNets, and DenseNet, for facial expression recognition using the FER2013 dataset. Each architecture offers distinct advantages and disadvantages concerning performance, parameter efficiency, and computational complexity. The application of data augmentation techniques, such as translation, rotation, flip, and zoom, helps to alleviate overfitting and enhances model performance. VGG models demonstrate the merits of increased network depth, while ResNets pioneer identity shortcut connections for improved optimization. MobileNets address image classification challenges in resource-constrained settings using depthwise separable convolutions, linear bottlenecks, and inverted residuals. DenseNet, with its densely connected layers, introduces a novel paradigm in CNN architecture design. Selecting an appropriate CNN architecture depends on the specific requirements and constraints of the application. Future research should focus on exploring innovative architectural designs, optimizing training methodologies, and addressing facial expression recognition challenges to improve model performance and expand real-world applicability.

References

[1] Gold J M Tjan B S Shotts M 2009 Integration of facial information is sub-optimal CogSci... Annual Conference of the Cognitive Science Society Cognitive Science Society (US) Conference NIH Public Access 2009: 2897

[2] Savchenko A V 2021 Facial expression and attributes recognition based on multi-task learning of lightweight neural networks 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY) IEEE 119-124

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

[4] Targ S Almeida D Lyman K 2016 Resnet in resnet: Generalizing residual architectures arXiv preprint arXiv:1603.08029

[5] Sandler M Howard A Zhu M et al. 2018 Mobilenetv2: Inverted residuals and linear bottlenecks Proceedings of the IEEE conference on computer vision and pattern recognition 4510-4520

[6] Huang G Liu Z Van Der Maaten L et al. 2017 Densely connected convolutional networks Proceedings of the IEEE conference on computer vision and pattern recognition 4700-4708

[7] Chengeta K Viriri S 2018 Facial expression recognition using local directional pattern variants and deep learning Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence 1-7

[8] Ramalingam S Garzia F 2018 Facial expression recognition using transfer learning 2018 International Carnahan Conference on Security Technology (ICCST) IEEE 1-5

[9] Kaggle 2020 FER 2013 https://www.kaggle.com/datasets/msambare/fer2013

[10] Yu Q Chang C S Yan J L et al. 2019 Semantic segmentation of intracranial hemorrhages in head CT scans 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) IEEE 112-115

[11] Deeba K Amutha B 2020 ResNet-deep neural network architecture for leaf disease classification Microprocessors and Microsystems 103364