Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230135

Multiple SOTA convolutional neural networks for facial expression recognition

Yuming Huang

Department of Computer Science, University of Wisconsin – Madison, 53703, WI, US

huang463@wisc.edu

Abstract. Facial Expression Recognition (FER) has been a popular topic in the field of computer vision. Various and plentiful facial expression datasets emerged every year for people to train their models and compete. ImageNet, as a massive database for image classification, became a standard benchmark for new computer vision models. Many excellent models such as VGG, ResNet, and EfficientNet managed to excel and were regarded as state-of-the-art models (SOTAs). This study aims to investigate whether SOTA models trained on ImageNet can perform exceptionally well in FER tasks. The models are categorized into three groups based on different weight initialization strategies and then trained and evaluated on the FER-2013 dataset. The results indicate that models with weights trained on ImageNet can be fine-tuned and perform well in FER-2013, particularly when compared to other groups. Finally, simpler models with less computational costs are promoted considering the need for real-time application of facial expression recognition.

Keywords: facial expression recognition, convolutional neural network, deep learning.

1. Introduction

Facial expression is a crucial indicator of a person's emotions. It is also an effective way to strengthen the tone of voice and interact with others in daily life. Although the use of facial expressions might slightly vary due to cultural differences in different regions, by observing people's facial expressions, people's primary emotions can reasonably be inferred. Ekman and Friesen proposed that different cultural regions with the least literary interactions also had a similar way of expressing emotion [1]. This further proves the universality of facial expression, and thus it is feasible to systematically analyze people's emotions.

With the development of machine learning, and especially the emergence of convolutional neural networks [2], Facial Expression Recognition (FER) became a popular area of study that lets computers do expression classification tasks and apply them to real-world problems. According to Li and Deng [3], FER systems are divided into two types: static image FER and dynamic sequence FER based on their feature representation. Static-based methods encode feature representations using only spatial information from the current single image, whereas dynamic-based methods consider the temporal relationship between contiguous frames in the input facial expression sequence. The application of FER is promising. For example, expression recognition can be applied to polygraph tests where computer cameras can capture test subjects' every micro expression that is hard to be collected by human eyes. In addition, FER can serve as a useful tool in Human-Computer Interaction (HCI). By allowing the camera to capture faces from the operators, computers can directly obtain facial data from them and analyze their emotions to generate correct responses that improve efficiency.

In the early stage of FER, traditional computer vision techniques were employed. One popular way of FER was the feature-based approach. Namely, important features of the face such as eyes, nose, mouth, and eyebrows are extracted and sent to the traditional machine learning classifier such as the decision tree or SVM (Support Vector Machines). Shang et al. suggest a study of facial expression based on the Local Binary Pattern (LBP) feature [4]. Multi-layer Perceptron was also used to solve the FER [5]. CNN has been applied to FER for a long time since its emergence. With the development of the semiconductor industry, stronger computing units enabled CNN models to become more complicated and accurate. It surpassed traditional machine learning methods and became one of FER's mainstream choices.

To find out whether popular CNN models trained from the ImageNet have the ability to classify facial features from the FER-2013 dataset, this paper evaluates the accuracy of pre-trained SoTAs (In the research, MobileNet, VGG16, EfficientNet, and ResNet50 were used), that are originally trained from the ImageNet dataset, on the FER-2013 dataset. In particular, three different groups of SoTAs are tested. The first group is labelled as the “pre-trained group”. Their weights from ImageNet were frozen, and only the fully connected part can be trained. The second group, which is called the “uninitialized group”, all have uninitialized weights, but they are free to adjust any weight. Finally, the third group, which is fine-tuned group, employed the idea of transfer learning, so they have weights from ImageNet and are allowed to be fully trained during the process. After conducting the experiment, the research shows that the fine-tuned group gives the best performance in accuracy, recall, and precision which shows that the weights from ImageNet can improve models’ performance on facial expression tasks.

2. Methodology

The following part introduces the dataset used in the research and the approach to get the results.

2.1. Dataset

The dataset of facial expressions used in this study is from Kaggle FER-2013 [6]. It contains over twenty thousand grey-scale facial images with a size of 48×48 pixels. All the images were obtained from the Google Images search engine and labelled by using crowd-sourcing techniques. After the release, this dataset soon became a popular benchmark in the field of facial expression recognition. Many scholars used FER models to evaluate the models’ performance. The whole dataset can be classified into seven categories and is labelled from 0 to 6 (i.e. 0=angry, 1=disgust, 2=fear, 3=happy, 4=neutral, 5=sad, 6=surprise) as shown in the Figure 1.

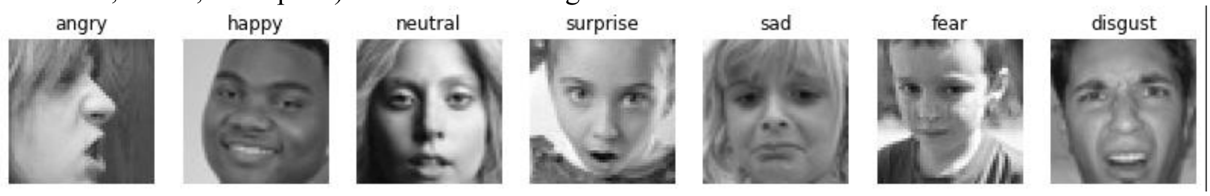

Figure 1. Seven categories of facial expressions from the dataset.

In addition to FER-2013, all pre-trained models used in this paper are based on the dataset called ImageNet which is a massive collection of image data used for training computer vision models. The dataset contains over 14 million labelled images with over 20000 categories [7]. Because this dataset is so instrumental to the development of deep learning models, a famous annual competition named ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was held each year. This competition used a partial set of the whole dataset and participants tried their best to build the model with the highest accuracy. Several excellent models which stood out from the crowd were used in this research.

2.2. Data pre-processing

Because FER-2013 is a fairly small dataset, the training set fed in the model is limited and might cause overfitting. As a result of that, the model will learn to match the noise and peculiarities of the training data rather than the underlying patterns and correlations that generalize to new data. There is a higher chance of overfitting with small datasets. Because of this, the model could perform poorly on hypothetical data and struggle to generalize to new unseen data. On the other hand, it can be more challenging to obtain accurate estimations of the model's performance when the training data size is small since the model's performance on the training data might not be indicative of its performance on real-world external data. Hence, it is reasonable to augment the dataset to reduce problems like overfitting and underfitting. In this research, each image was geometrically augmented through vertical and vertical flips. The image's brightness was also augmented by setting the original brightness to 80% brightness, and 50% brightness.

In addition to data augmentation, regularizers were introduced to further avoid overfitting. Specifically in this research, Lasso regularization(L1) was employed. It is a machine learning and statistics strategy used to reduce overfitting and increase model generalization. It works by introducing a penalty term to a model's cost function, which pushes the model to select fewer characteristics for its predictions. The penalty term is proportional to the total of the absolute values of the model coefficients. This means that Lasso regularization can reduce some of the coefficients to zero, thereby deleting the corresponding features from the model [8].

2.3. CNN and state-of-the-art

The Convolutional Neural Network (CNN) is a type of neural network where convolutional layers are introduced to extract important features of images. Generally, convolutional layers, pooling layers, and fully connected layers are the three main types of heterogeneous layers in a CNN [2]. The convolutional layer employs a set of learnable filters or kernels to perform a series of convolutions on the input picture. Each filter moves through the input image, computing a dot product between its weights and the local pixel values at each place to generate a feature map that highlights various patterns and characteristics in the image. The pooling layer works by splitting the feature map into a series of non-overlapping rectangular sections and then replacing each region with a single value that summarizes the information in that region. After repetitions of this feature learning process, the data are sent to fully connected layers. In this set of layers, every neuron in this layer is linked to every neuron in the previous layer. The layer's output is a vector of values, with each value representing the activity of a specific neuron in the layer. At the layer, the activation function is commonly a rectified linear unit (ReLU) [9], or a sigmoid function, which introduces nonlinearity into the network.

2.3.1. Model selection. For the list of the selected models, the most representative models shown in Table 1 were chosen in this research. ResNet is distinguished by the employment of residual connections, which allow the network to learn residual functions rather than entire mappings [10]. A residual connection is a skip connection that adds a layer's input to the layer's output, allowing the network to learn the difference between the input and output rather than the full function. This method has been found to increase the training stability and performance of very deep neural networks. The EfficientNet architecture is based on a compound scaling method that systematically scales up the depth, width, and resolution of the network [11]. Lastly, VGG model was used because it has generally good performance on a variety of image recognition tasks [12].

Table 1. Choice of models.

	The number of parameters/in million	depth	Accuracy for ImageNet
ResNet50	25.6	107	74.9%
VGG16	138.4	16	71.3%

Table 1. (continued).

EfficientNetB0	5.3	132	77.1%

2.4. CNN and state-of-the-art

All the models mentioned above were introduced by using TensorFlow Keras. Since FER-2013 only has seven categories, the output layer was changed into seven neurons with softmax activation. Additionally, the Fully Connected Nodes in each model were replaced with three customized FCN layers. Each layer has 32 neurons and gets passed by a batch normalization layer [13], and a dropout layer [14]. For the pre-trained group, the weights in the convolutional layers were frozen whereas the untrained group was allowed to adjust weights freely. For the last fine-tuned group, models were initialized with the weights from ImageNet and trained additionally from the FER-2013 training set. All models were compiled with Adam optimizers with a learning rate of 0.001. By making decreasing rate extremely small, the weights can gradually converge to the optimal value instead of overshooting [15]. Every model has trained 50 epochs, and its accuracy and loss were further summarized. To evaluate all models' performance, metrics such as accuracy, recall, precision, and f1 score were used. A confusion matrix can better help to understand those metrics.

Table 2. Metrics for evaluating the performance of the model.

	Actual positive	Actual negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

As shown in the Table 2. Accuracy is the sum of TP and TN divided by all components. The percentage of correctly categorized photos is how accuracy is calculated. Although it is a widely used statistic for classification tasks, it might not be the ideal one in cases of imbalanced datasets when the accuracy of the majority class predominates. Recall examines the percentage of genuine positive forecasts among all positive predictions, whereas precision evaluates the percentage of true positive predictions across all occurrences of positive outcomes [16]. These measurements are helpful for applications like object recognition and segmentation, where the objective is to locate every occurrence of a certain item in an image. When the dataset is unbalanced, the F1 score—which is the harmonic mean of accuracy and recall—is a helpful indicator. It is frequently used in tasks involving object recognition and segmentation since it incorporates accuracy and recall into a single metric.

3. Results and discussion

The results of all groups of models are shown below.

3.1. The performance of various models

Table 3. A slightly more complex table with a narrow caption.

	Models	Accuracy	Precision	Recall	F1-score	Loss
Pre-trained	VGG16	0.8583	0.5719	0.0322	0.0597	1.6923
Pre-trained	ResNet50	0.8571	0.4333	0.0017	0.0033	1.674
Pre-trained	EfficientNet	0.8571	≈0	≈0	≈0	1.8101
Uninitialized	VGG16	0.8905	0.8020	0.3102	0.4440	1.1808
Uninitialized	ResNet50	0.8878	0.7854	0.2952	0.4259	1.2634

Table 3. (continued).

Uninitialized	EfficientNet	0.8873	0.7736	0.2980	0.4276	1.2795
Fine-tuned	VGG16	0.8950	0.7914	0.3598	0.4922	1.3215
Fine-tuned	ResNet50	0.8892	0.8075	0.2980	0.4311	1.2204
Fine-tuned	EfficientNet	0.9002	0.7699	0.4296	0.5479	1.0794

The obtained results shown in Table 3 revealed that the VGG16 model displayed exceptional accuracy in all experimental groups, exhibiting the highest accuracy in general. Nonetheless, the recall rate was significantly low in the pre-trained VGG16 model. Conversely, the pre-trained ResNet model exhibited suboptimal performance during the experimental process, with fluctuating metrics indicative of overfitting, despite ultimately achieving satisfactory accuracy.

It is clear that all models in the pre-trained group have extremely low recall values, which means that it is good at correctly identifying positive instances but may be missing many true positive instances. In other words, the model is conservative in making positive predictions and is likely to miss some positive instances. One reason for this phenomenon is that all models are trained based on the images of ImageNet. The weights in models are adjusted for detecting features of different objects instead of facial organs. This leads to difficulty in feature extraction, and the convolutional layers fail to extract crucial features that FCN can use to make the correct classification. Thus, directly applying models trained from the ImageNet to FER may not be a good option.

Upon comparing the uninitialized group with the fine-tuned group, it was observed that the latter displayed marginally superior metrics across various evaluation measures. This observation implies that the incorporation of convolutional layer weights trained on the ImageNet dataset can assist in optimizing the performance of novel FER models, enabling effective loss reduction.

4. Conclusion

Generally, all state-of-the-art models in this research displayed a good performance in generalizing data, and classifying most facial expressions correctly. However, it is not recommended to freeze weights in convolutional layers of all ImageNet-trained models because the original weights are not designed for FER tasks. However, instead of training from the start, the weights from the ImageNet dataset can be further used to improve the loss and accuracy of the new FER models. Fine-tuning can perform better in the training process than the other two groups. Although these models were initially designed for object classification, their structure can be transferred to other image-related classification problems and remain a high accuracy.

Since FER may often be applied to real-time capture and detection. The ultimate goal of this application is to reduce the time of classification within the frame rate so that the model can give the response in time. That also means complex models such as VGG16 and ResNet50 might not be capable of handling this task because it usually takes a long time given so many layers and parameters. Here, the EfficientNet could be the best choice among these three models because it has the smallest number of parameters and layers, and still maintains a satisfying level of accuracy and loss (it is even higher than that of VGG and ResNet in fine-tuned groups). In addition, many real-time applications are usually installed on portable devices, and their computing units are confined by the power of the battery. Thus, the cost of computation will be prioritized. Other lightweight models could also be considered if they have an acceptable range of performance drop.

References

[1] Ekman P and Friesen W V 1971 Constants across cultures in the face and emotion (Journal of Personality and Social Psychology vol 17) pp 124-129

[2] LeCun Y and Bengio Y 1995 Convolutional networks for images, speech, and time series (The handbook of brain theory and neural networks vol 3361) p 1995

[3] Li S and Deng W 2020 Deep Facial Expression Recognition: A Survey (IEEE transactions on affective computing vol 13 ed 2) pp 1195-1215

[4] Shan C et al 2009 Facial expression recognition based on Local Binary Patterns: A comprehensive study (Image and Vision Computing vol 27) pp 803-816

[5] Zhang Z 1998 Feature-based facial expression recognition: Experiments with a multi-layer perceptron INRIA

[6] Zahara L et al 2020 The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi (Fifth International Conference on Informatics and Computing (ICIC)) pp 1-9

[7] Deng et al 2009 ImageNet: A large-scale hierarchical image database (2009 IEEE Conference on Computer Vision and Pattern Recognition) pp 248-255

[8] Robert T 1996 Regression Shrinkage and Selection via the Lasso (Journal of the Royal Statistical Society Series B (Methodological) vol 58) pp267-288

[9] Nair V and Hinton G E 2010 Rectified linear units improve restricted boltzmann machines (Haifa) pp 807-814

[10] He K et al 2016 Deep Residual Learning for Image Recognition (In Proceedings of the IEEE conference on computer vision and pattern recognition) pp 770-778

[11] Tan M et al 2019 EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (PMLR)

[12] Simonyan K and Zisserman 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition (3rd International Conference on Learning Representations (ICLR 2015) pp 1-14

[13] Ioffe S and Szegedy C 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift (32nd International Conference on International Conference on Machine Learning vol 37) pp 448-456

[14] Srivastava N et al Dropout: A Simple Way to Prevent Neural Networks from Overfitting (The journal of machine learning research vol 15) pp 1929-1958

[15] Wilson D R and Martinez T R 2001 The need for small learning rates on large problems (International Joint Conference on Neural Networks) pp 115-119

[16] Powers and David M W 2020 Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation arXiv preprint arXiv:2010.16061