Facial Emotion Recognition using Convolutional Neural Networks

Abstract:
Artificial neural networks, or shortly neural networks, find applications in a very wide spectrum. Neural networks are the essential building parts of deep learning algorithms: they are capable of learning complex patterns, such as hierarchical representations of features (image recognition) and intricate relationships in data (stock market predictions). In this article, we will dive into the image recognition field, specifically facial emotion recognition.

We will first go through a related literature review to get familiarized with the topic. Next, the methodology section comes in, where the applied methods and a CNN structure are explained. Lastly, my experimental results are provided, reflecting on my accuracy levels and reasons for some confusion my neural networks had when predicting emotions.

The applied dataset includes more than 10000 images and involves 3 different types of emotions (fear, anger, happiness). We adopted the VGGNet architecture, as well as the ResNet50 architecture, and experimented with various optimization methods. The purpose of this paper is to observe how well CNNs can recognize facial human emotions.

In the end, while experimenting, we concluded that emotions like fear and anger were confused with happiness images, but oftentimes emotions can be predicted quite correctly. Some wrong predictions could have happened due to the complexity of emotions, and quite big size of the dataset.

Keywords: Artificial Neural Networks, Deep Learning, CNN, Image Classification, VGG Architecture, ResNet50, Facial Emotion Recognition, Optimization Methods.

Introduction:
Since the twentieth century, Ekman et al. [2] defined seven basic emotions: anger, fear, happiness, sadness, contempt [6], disgust, and surprise. Facial expression for emotion detection has always been an easy task for human beings, but for a computer, this is a much more complex endeavor.
Computer vision is a popular field in data science, and CNNs have become very ubiquitous in terms of computer vision techniques. Among the different types of

neural networks (recurrent neural networks (RNN), long short-term memory (LSTM), artificial neural networks (ANN), etc.), CNNs are undoubtedly one of the most popular. They work exceptionally well on computer vision tasks like image classification, object detection, image recognition, etc., and they have been widely used in artificial intelligence modeling.

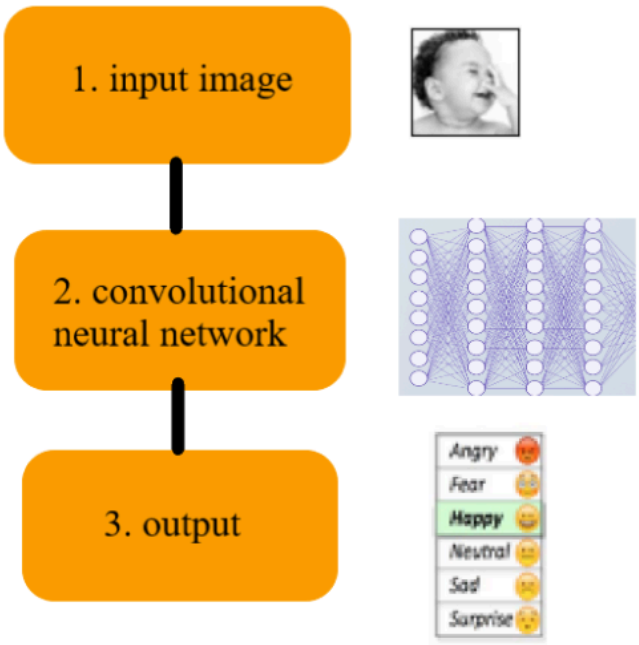
These days, image recognition technology has ways of application in Internet applications such as image search, face recognition, and image detection. In the early image recognition system, feature extraction methods such as Scale Invariant feature transform (SIFT [4]) and histogram of oriented gradients (HOG [5]) were used, and then the extracted Feature input classifier for classification and recognition. These features are essentially a feature of manual design. For different identification problems, the extracted features have a direct impact on the performance of the system, so the researchers need to study the problem areas to be studied to design Adaptability to better features, thereby improving system performance. This period of the image recognition system is generally for a specific identification task, and the size of the data is not big, generalization ability is poor, and it is difficult in the practical application of the problem to achieve an accurate identification effect.

Related work:
Since being introduced in the late 1990s, CNNs have shown a potential in image processing [7]. A typical CNN includes a convolutional layer, a pooling layer, and a fully connected layer. However, at that time, the application of CNNs was limited since there was a lack of training data and computing power. After the 2010s, the growth of computing power and the collection of bigger datasets made CNNs a much more usable tool in image classification [8].

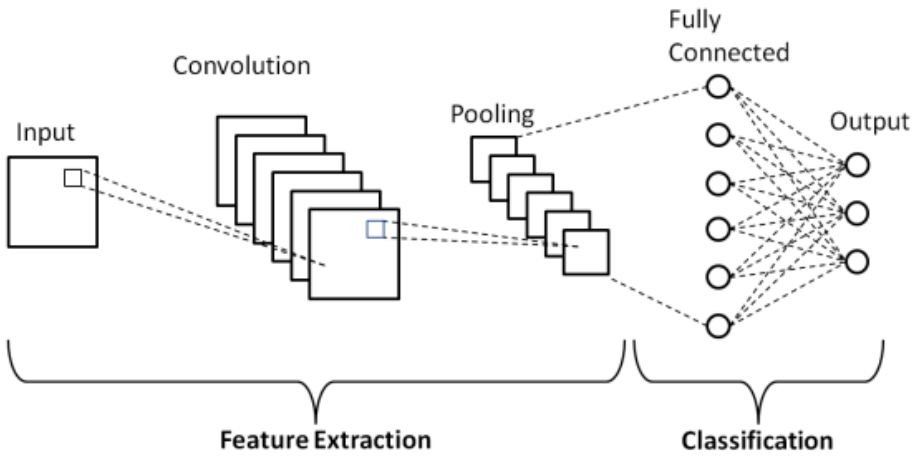
Many researchers are interested in constantly improving the learning environment with Face Emotion Recognition (FER). For example, Tang et al. [9] developed a system that can analyze students' facial expressions to estimate classroom teaching environments. The system is composed of five phases: data acquisition, face detection, face recognition, facial expression recognition, and post-processing. The approach uses K-nearest neighbor (KNN) for classification and Uniform Local Gabor Binary Pattern Histogram Sequence (ULGBPHS) for pattern analysis.
Savva et al. [10] proposed a web application that performs an analysis of students' emotions while participating in active face-to-face classroom instruction. The application uses webcams that are installed in classrooms to collect live recordings, then they apply different machine learning algorithms.

Various techniques have been suggested to even further improve performance. For instance, the Sigmoid activation function has been used instead of Rectified Linear Unit (ReLU) activation to eliminate gradient dispersion problems and speed up training [14].
A great deal of research has also been done in creating different optimization algorithms used in training. Though there are no systematic rules on choosing an optimizer, empirical results show that a suitable optimization algorithm can significantly enhance a model's performance [19].
Among many others, one significant factor that could impact performance is the learning rate. A large learning rate could lead to oscillations around the minima or divergence in the loss. A small learning rate would slow down the model's convergence significantly and could trap the model at a non-optimal local minimum. A commonly used technique is to employ a learning rate scheduler that changes the learning rate during training [12].

Methodology:
In this section, we discuss the proposed system to analyze facial expressions using a Convolutional Neural Network (CNN) architecture. First, the system detects a face from the input image. Then, these images are applied as an input to CNN. Finally, the output is the facial emotion recognition results (anger, happiness, sadness, disgust, surprise, or neutral). The figure below presents the structure of our proposed approach.



Convolutional neural network is a class of deep learning methods that has become pervasive in various computer vision tasks and is attracting interest across a variety of domains [13]. They can identify visual patterns from input images with minimal preprocessing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered [14]. The term convolution refers to the use of a filter or kernel on the input image to produce a feature map.
CNN model contains 3 types of layers as represented in the figure below:



source.
https://www.researchgate.net/figure/Schematic-diagram-of-a-basic-convolutional-neural-network-CNN-architecture-26_fig1_336805909

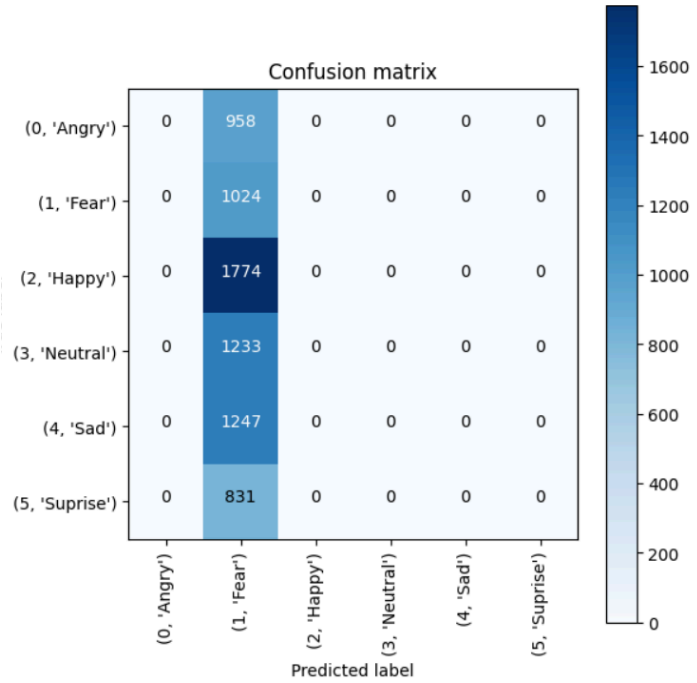
1. Input: We would load the input, usually in the form of a multidimensional vector, to the input layer which will spread it to the hidden layers [16].
2. Convolutional Layer is the first layer to extract features from an input image. The main goal of Convolution in the case of a ConvNet is to extract features from the input image. [17]
3. A Pooling Layer reduces the dimensionality of each feature map but retains the most important information [15].
4. A fully connected layer is a traditional Multi Layer Perceptron that applies an activation function in the output layer. The term “Fully Connected” implies that every neuron in the previous layer is connected to every neuron in the next layer. The goal of the Fully Connected layer is to use the output of the two previous layers for classifying the input image into various classes based on the training dataset. [17]
5. Output: Artificial Neural Networks are mainly comprised of a high number of interconnected computational nodes (neurons), which work entwined in a

distributed fashion to collectively learn from the input to optimize its final output [16].

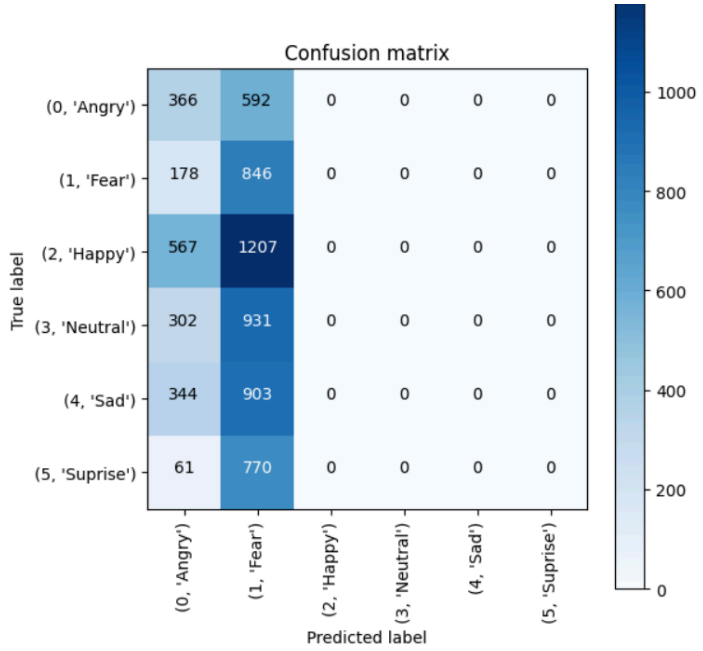
Experimental Results:

We trained our neural networks using the Facial Recognition Dataset which includes six emotions (happiness, anger, sadness, neutral, fear, and surprise). In our first experiment, we only compared images of fear and determined how accurately the neural network can recognize the fear emotion and not confuse it with any other. The detected face images were resized to 48×48 pixels, and converted to grayscale images then were used for inputs to the CNN model.

Our model predicts fear greatly but often confuses it with a happy emotion, as well as other emotions.

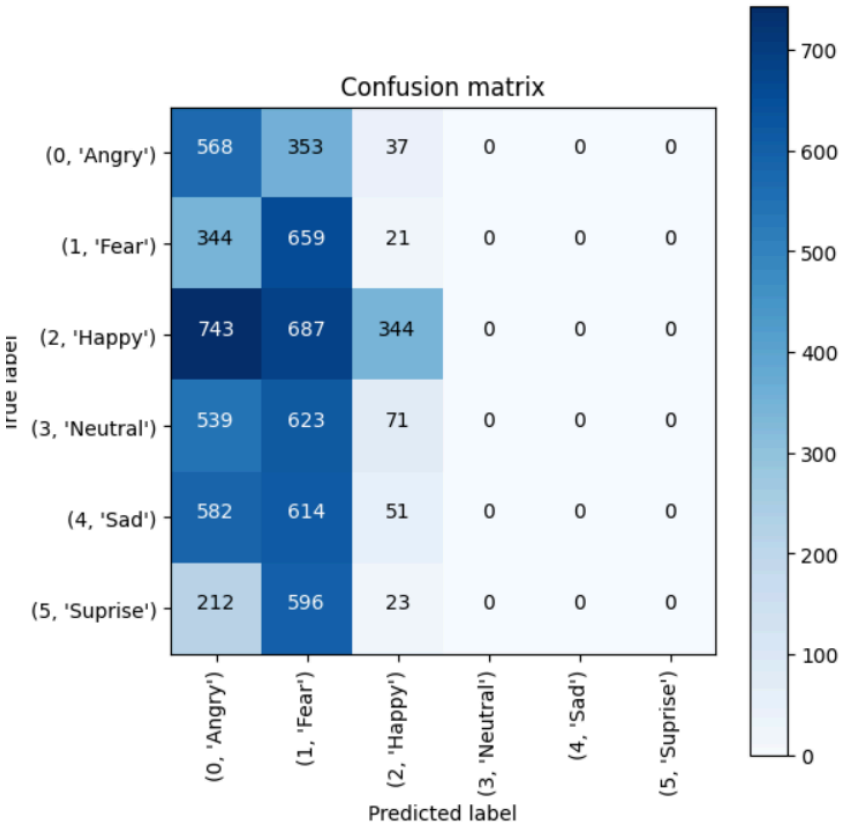


For a comparison, we tested a simpler model at 25 epochs.



Here, we added the next emotion (anger). The confusion matrix clearly changed and the angry emotion became more recognizable, but the neural network still confuses it with the happy emotion. Fear is quite often confused with the happy emotion, and less recognized precisely.

Lastly, we run the ResNet50 model (a 50-layer convolutional neural network).



We finally added a third emotion (happy). Test accuracy came out a bit higher (0.22). The ‘angry’ emotion is still being confused with the happy emotion; However, the ‘happy’ emotion is mostly recognized correctly.

Conclusion:

In this article, we showed several types of Convolutional Neural Networks and analyzed their performance in terms of facial emotion recognition. We revised architectures like VGG and ResNet50, using 3 main emotions (anger, fear, happiness) and comparing with a wider range of emotions (anger, fear, happiness, neutrality, sadness, surprise), and determined their performance. We concluded that even though our models do confuse some emotions with others, generally they recognize emotions correctly.

Annotated bibliography:

[1] Koushik, J. (2016). Understanding Convolutional Neural Networks
[2] Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124
[3] Sajid M, Iqbal Ratyal N, Ali N, Zafar B, Dar SH, Mahmood MT, Joo YB (2019) The impact of asymmetric left and asymmetric right face images on accurate age estimation. Math Problem Eng 2019:1–10
[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 2004
[5] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. San Diego, USA
[6] Matsumoto D (1992) More evidence for the universality of a contempt expression.
[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient Based learning applied to document recognition,"
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks,"
[9] C. Tang, P. Xu, Z. Luo, G. Zhao, and T. Zou, "Automatic Facial Expression Analysis of Students in Teaching Environments," in Biometric Recognition,
[10] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence
[11] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout,"
[12] W. S. Chin, Y. Zhuang, Y. C. Juan, and C. J. Lin, "A learning-rate schedule for stochastic gradient methods to matrix factorization,"
[13] Convolutional neural networks: an overview and application in radiology, Rikiya Yamashita, 2018
[14] aionlinecourse.com/tutorial/machine-learning/convolution-neural-network. Accessed 20 June 2019
[15] ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/.
[16] An Introduction to Convolutional Neural Networks, Keiron O'Shea
[17] Facial Emotion Recognition of Students using Convolutional Neural Network, Imane Lasri