Image-based Facial Expression Recognition using CNN

Temirlan Nurmakhan
dept.of Computer Science
Nazarbayev University
Astana, Kazakhstan
temirlan.nurmakhan@nu.edu.kz

Zhanna Mukhametsharip
dept.of Computer Science
Nazarbayev University
Astana, Kazakhstan
zhanna.mukhametsharip@nu.edu.kz

Ainur Khamitova
dept.of Computer Science
Nazarbayev University
Astana, Kazakhstan
ainur.khamitova@nu.edu.kz

I. PROJECT SUMMARY

Facial Expression Recognition (FER) stands at the forefront of computer vision, playing a pivotal role in various domains, including human-computer interaction, emotion analysis, and beyond. As technology continues to weave itself into our daily lives, the ability to understand and interpret human emotions through facial expressions becomes increasingly critical.

The main goal of our project is to improve the accuracy and robustness of an existing Convolutional Neural Network (CNN) model tailored for FER. For this purpose, we have chosen the FER-2013 dataset, a repository of diverse facial expressions that serve as our testing ground. However, some challenges need to be addressed.

The first challenge is accuracy. Our model must excel in distinguishing subtle nuances in expressions, ensuring that it interprets happiness, sadness, anger, surprise, fear, disgust, and neutrality, and minimizes biases and overlaps between these expressions.

Secondly, robustness is of great concern. In real-world scenarios, lighting conditions, facial angles, and occlusions are variables that can wreak havoc on FER models. Our goal is to equip our model with the resilience to perform consistently under various environmental conditions.

The significance of Facial Expression Recognition extends far beyond its academic purpose. Its practical applications are extensive and profound. In the realm of virtual environments, improved FER can lead to more immersive and responsive virtual reality experiences. User experience testing can provide valuable insights into product and service design by analyzing user emotions and reactions accurately. Moreover, in the context of human-computer interactions, where emotions often convey unspoken cues, FER becomes the key to creating machines that understand us better, fostering more natural and efficient exchanges between humans and technology. Whether it's in healthcare, entertainment, marketing, or countless other domains, optimized FER models can transform the way we engage with the digital world.

In summary, our mission is to enhance the performance of a baseline CNN model for Facial Expression Recognition using the FER-2013 dataset. It holds the promise of enabling technology to understand and respond to human emotions more accurately, thereby enhancing our daily experiences and interactions. This project represents a significant step towards advancing the state-of-the-art in FER to aid research in computer vision

II. DATA DESCRIPTION

For the project, our team is planning to use the FER-2013 dataset. The environment of the dataset is Wild. The dataset can be obtained from the Platform Kaggle [1]. The dataset contains approximately 35,887 facial RGB images of different expressions with a size restricted to 48×48. The training set consists of 28,709 examples and the public test set consists of 3,589 examples. The dataset includes 4953 Anger, 547 Disgust, 5121 Fear, 8989 Happiness, 6077 Sadness, 4002 Surprise, and 6198 Neutral images. Examples of some images can be seen in Fig. 1.

Fig. 1. Example from the Dataset

The main labels of it can be divided into 7 types: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. Each image in the dataset represents one of the emotions with the label.

The dataset was used for different research papers in recent years. The research that used FER-2013 was ranked #9 by using the VGG model with an accuracy of approximately 72.7%. Another research that used the model Res-net acquired an accuracy of 72.4% and was ranked 10th position. The Inception model-based project was placed in 13th place with an accuracy of 71.6%.

Overall, the dataset has a wide range of implementations in different projects and got relatively good results for a variety of models.

III. METHODOLOGY

a. Baseline Machine Learning Model

A Convolutional Neural Network (CNN) is a frequently used baseline model for the FER-2013 dataset. CNN variations have proven to be effective in face expression recognition by achieving classification accuracies ranging from 65% to 72.7% [2]. The basic CNN model typically consists of several convolutional layers, pooling layers for feature extraction, and one or more fully connected layers for classification.

There are four widely recognized architectures: AlexNET, VGG, Inception, and ResNet. In our project, we will consider VGG for the following reasons: Firstly, pre-trained Visual Geometry Group (VGG) models on large datasets like ImageNet are available for use. Secondly, results from fine-tuning a VGG model on a smaller dataset can be fairly accurate. It enables quick evaluation of the facial expression recognition system's performance. Thirdly, it has a consistent structure with repeated convolutional and pooling layers which eases capturing a wide range of features at different scales, such as facial expressions. VGG models, especially smaller variants like VGG16 and VGG19, are computationally less costly when compared to newer architectures like ResNet or Inception [3].

We will employ the VGG16 classification model for emotion recognition. VGG16 is made up of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. These convolutional layers extract features from the input image, and the fully connected layers classify the image based on these extracted characteristics. The network's first layer processes the input image's raw pixel values, while successive layers gradually extract more complex characteristics; the max-pooling layers help to minimize the features' spatial dimensions. To achieve more accurate results in the model's performance on the chosen dataset, we will adjust the architecture and fine-tune weights and hyperparameters. Presumably, ReLU will be used in the convolutional layers, while softmax will be used in the output layer for multi-class classification. We will apply the categorical cross-entropy loss function as our loss function. The model is typically trained using the Adam optimizer or Stochastic Gradient Descent (SGD).

The baseline CNN offers a strong basis for face expression detection tasks on the FER dataset, but more complex models, architectures, and approaches can be investigated to enhance performance. In the provided code for the paper at [4] "Facial Emotion Recognition: State of the Art Performance on FER-2013" the proposed model achieved single-network accuracy of 73.28 % on FER-2013 without using extra training data (Fig 2). The model is made up of 4 convolutional stages and 3 fully connected layers. Each convolutional stage contains two convolutional blocks and a max-pooling layer. The convolution block has a convolutional layer, a ReLU activation layer, and a batch normalization layer. In this paper, Khaireddin et al.[2] did a grid search across all parameters using six different optimizers and five learning schedulers for tuning. The deep neural networks were visualized using a saliency map. The classification of "happiness" and "surprise" by this model was the most accurate, whereas the classification of "disgust" and "anger" was less accurate. The small sample size in the training set contributed to the low classification accuracy in the categories of "disgust" and "fear". The misclassification of the categories of "fear" and "sadness" was attributed to inter-class similarities.

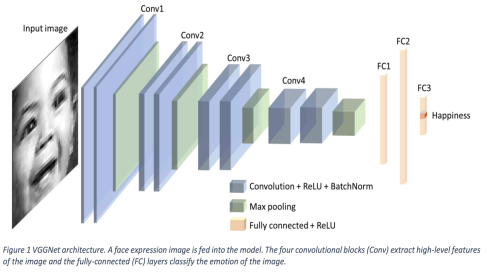

Figure 1 VGGNet architecture. A face expression image is fed into the model. The four convolutional blocks (Conv) extract high-level features of the image and the fully-connected (FC) layers classify the emotion of the image.

Fig 2. VGGNet Architecture[4]

IV. Proposed Idea

Multiple types of research were done to improve Facial Expression Recognition (FER) accuracy using FER-2013. Most of them apply different CNNs due to automatic feature extraction and computational efficiency. Our main proposed idea is to take the baseline model and enhance its accuracy for FER 2013. In this project, we intend to experiment with data augmentation techniques and optimize computational approaches.

Firstly, we want to try random transformations on the dataset, such as rotations, translations, zooming, shearing, and horizontal flips to increase the training data variance. In addition, data augmentation increases the size of the training set and reduces overfitting, thus allowing the model to have a better generalization [5].

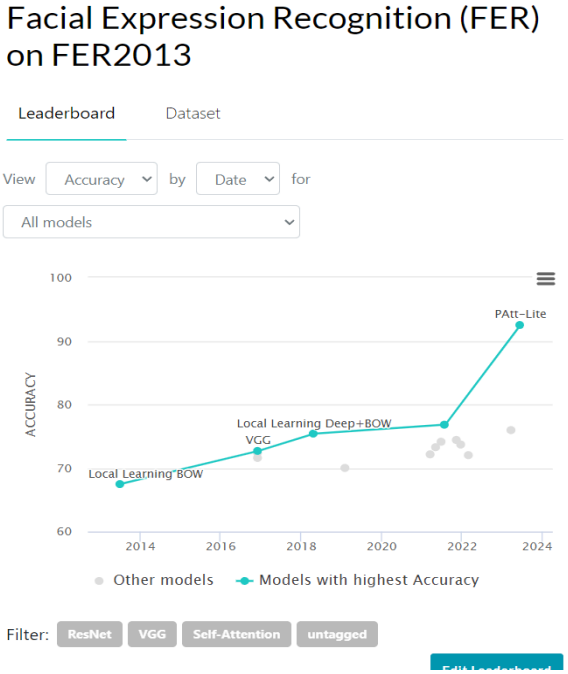

Fig. 3. Accuracy of different models on FER2013 [9].

Secondly, we can apply transfer learning based on other CNN architectures that might improve the performance of the model. There are multiple types of research without published source code that report high accuracy on FER2013. “Papers with Code” is a database of Machine Learning papers, code, datasets, methods, and evaluation tables [6] . According to information in the database presented for the FER2013 dataset shows that there are multiple models with different levels of accuracy(Fig. 3). The most recent research on FER uses a light patch and attention network based on MobileNetV1, also known as PAtt-Lite, and they have achieved the best state-of-the-art results on FER2013 of 92.5% [7]. Another model achieved a state-of-the-art single-network accuracy of 73.70 % on FER-2013 without using extra training data [8]. We will experiment with such CNN models and try at least to repeat their success.

Thirdly, we will try to integrate various model architectural modifications by changing the width, depth, and dropout layers. Furthermore, we will look into hyperparameter optimization methods that can affect the model's performance.

We will compare our solution’s performance with baseline models by using the same performance metrics, namely accuracy and F1-score or confusion matrix. As a result of these methodologies, we expect to exceed the performance of the chosen baseline model.

References

[1] M. Sambare, “Fer-2013,” Kaggle, https://www.kaggle.com/datasets/msambare/fer2013 (accessed Sep. 15, 2023).

[2] Y. Khaireddin and Z. Chen, “Facial Emotion Recognition: State of the Art Performance on FER2013,” arXiv, May 08, 2021.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ‘Densely Connected Convolutional Networks’, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269 (accessed Sep. 15, 2023).

[4] Y. Khaireddin., "fer," GitHub. [Online]. Available: https://github.com/usef-kh/fer (accessed Sep. 15, 2023).

[5] R. Romano, “How data augmentation can improve ML model accuracy,” Qwack, May 19, 2022. [Online]. Available: https://www.qwak.com/post/how-data-augmentation-can-improve-ml-model-accuracy. (accessed Sep. 15, 2023).

[6] "Papers with Code," About. [Online]. Available: https://paperswithcode.com/about. (accessed Sep. 15, 2023).

[7] J. Ngwe, K. Lim, C.-P. Lee, and T. Ong, "PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition," June 2023, https://arxiv.org/pdf/2306.09626v1.pdf (accessed Sep. 15, 2023).

[8] LetheSec, "Fer2013-Facial-Emotion-Recognition-Pytorch," GitHub, [Online]. Available: https://github.com/LetheSec/Fer2013-Facial-Emotion-Recognition-Pytorch/blob/main/README.md (accessed Sep. 15, 2023).

[9] "Papers with Code," State-of-the-Art (SOTA) for Facial Expression Recognition on FER2013, [Online]. Available: https://paperswithcode.com/sota/facial-expression-recognition-on-fer2013 (accessed Sep. 15, 2023).