Convolutional Neural Network Model Comparisons For Face Emotion Recognition

Burcu Alakuş-Çınar
Cognitive Science Department
Middle East Technical University
Ankara, Türkiye
burcu.alakus@metu.edu.tr

***Abstract*—This study compares various Convolutional Neural Network (CNN) architectures and parameter configurations for face emotion recognition. Multiple CNN models were evaluated using a facial emotion dataset, adjusting key parameters. The results indicate significant variations in accuracies among the models. These findings offer insights into optimising CNNs for emotion recognition, with applications in human-computer interaction and security systems.**

***Keywords*—convolutional neural network (CNN), face emotion recognition, emotion detection, deep learning, parameter tuning**

I. Introduction

Face emotion recognition is a challenging and essential task in the fields of computer vision and artificial intelligence. It involves identifying and classifying human emotions from facial expressions captured in images or videos. Accurate emotion recognition has wide-ranging applications, including human-computer interaction, psychological analysis, security systems, and entertainment.

A. Problem Statement

Despite significant advancements in deep learning, accurately recognizing emotions from facial expressions remains a complex problem due to variations in facial features, lighting conditions, occlusions, and individual differences in expressing emotions. Traditional machine learning approaches often fail to handle these challenges, necessitating more sophisticated techniques such as CNNs.

B. Objectives

This study, conducted as part of a deep learning course project, aims to design, implement, and compare various CNN architectures and parameter configurations to identify the optimal model for face emotion recognition. By systematically adjusting key parameters like learning rate and model architectures, this research seeks to improve the accuracy and efficiency of CNN models in detecting facial emotions. The insights gained from this study can contribute to developing more robust and effective emotion recognition systems.

II. Related Work

A. Studies on Face Emotion Recognition

Face emotion recognition has been an active area of research for several decades. Early approaches relied on handcrafted features and traditional machine-learning algorithms. However, these methods often struggled with variations in lighting, occlusions, and the subtlety of human emotions.

With the advent of deep learning, CNNs have become state-of-the-art in face emotion recognition due to their ability to automatically learn hierarchical feature representations from raw images. Several notable studies have made significant contributions to this field:

Viola and Jones (2001) [1] introduced a robust face detection framework that paved the way for subsequent emotion recognition research. Their work focused on detecting faces in real-time using Haar-like features and a cascade of classifiers. Fasel and Luettin (2003) [2] provided a comprehensive review of early face emotion recognition methods, highlighting the limitations of traditional approaches and the potential of emerging techniques. Kahou et al. (2013) [3] demonstrated the effectiveness of CNNs for emotion recognition by combining deep features with temporal information from videos. Their model achieved high accuracy on benchmark datasets, showcasing the potential of deep learning in this domain. Mollahosseini et al. (2016) [4] developed a deep neural network called 'AffectNet' that achieved state-of-the-art performance on the challenging FER-2013 dataset. Their work highlighted the importance of large-scale annotated datasets and data augmentation techniques for training deep models. Li et al. (2018) [5] proposed an attention mechanism to enhance the performance of CNNs in emotion recognition. Their model selectively focused on salient regions of the face, improving accuracy by reducing the impact of irrelevant background information. Wang et al. (2020) [6] Investigated the impact of different CNN architectures and training strategies on emotion recognition performance. Their comprehensive study provided valuable insights into the trade-offs between model complexity and accuracy.

B. CNN Architectures in Emotion Detection

CNNs have revolutionized the field of emotion detection from facial images due to their ability to automatically learn and extract hierarchical features. Several CNN architectures have been proposed and evaluated for their effectiveness in recognizing facial emotions. This section provides an overview of notable CNN architectures and their contributions to emotion detection:

AlexNet, introduced by Krizhevsky et al.(2012) [7], significantly improved image classification performance. While not specifically designed for emotion recognition, its success in general image classification tasks paved the way for its application in various domains, including emotion detection. Simonyan and Zisserman (2014) [8] proposed VGGNet, which emphasized the importance of deeper networks with smaller convolutional filters. VGGNet's architecture has been widely adopted in emotion recognition tasks due to its simplicity and effectiveness in capturing fine-grained features. He et al. (2015) [9] introduced the ResNet architecture, which addressed the degradation problem in deep networks through residual learning. ResNet's ability to train very deep networks without performance degradation has made it a popular choice for emotion recognition, providing high accuracy and robustness. Szegedy et al. (2015) [10] proposed the Inception architecture (also known as GoogLeNet), which introduced the concept of multi-scale processing through Inception modules. This architecture has been effective in capturing diverse features at different scales, making it suitable for emotion detection tasks where facial expressions vary widely. Huang et al. (2017) [11] introduced DenseNet, which connects each layer to every other layer in a feed-forward fashion. This architecture promotes feature reuse and mitigates the vanishing gradient problem, leading to improved performance in emotion recognition by leveraging dense connections. Minaee and Abdolrashidi (2017) [12] specifically designed the EmotionNet architecture for facial emotion recognition. This model combined CNNs with facial landmark localization to enhance emotion detection accuracy. EmotionNet demonstrated the effectiveness of integrating spatial facial features with deep learning models. Howard et al. (2017) [13] developed MobileNet, a lightweight CNN architecture optimized for mobile and embedded applications. MobileNet's efficiency and low computational requirements have made it an attractive choice for real-time emotion recognition on resource-constrained devices. Li et al. (2022) [13] incorporated attention mechanisms into CNNs, allowing the network to focus on salient regions of the face. This approach significantly improved emotion recognition performance by emphasizing important facial features and reducing the influence of irrelevant background information.

III. Methodology

A. Dataset Description

For this study, a well-known facial emotion dataset, FER-2013, that includes a diverse range of facial expressions labelled with corresponding emotions is used. The dataset consists of grayscale images, each of size 48x48 pixels, categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is split into training, validation, and test sets to facilitate model evaluation and comparison.

B. CNN Model Architectures

Several CNN architectures are designed and implemented, varying in depth and complexity, to evaluate their performance in facial emotion recognition. The architectures ranged from simple models with a few convolutional layers to more complex ones incorporating advanced techniques such as residual connections and attention mechanisms. The primary architectures used in this study include:

- **Custom CNN with different learning rates:** A custom CNN model consists of several convolutional layers with ReLU activation and batch normalization, followed by max-pooling and dropout layers, culminating in fully connected dense layers, and outputs through a Softmax layer for classification as shown in Fig. 1 trained with learning rates of 1, 0.01, 0.0001.
- **Custom CNN using data augmentation:** Same architecture as custom CNN but trained with augmented data including re-scaling, random rotations, shifts, shearing, zooming, and horizontal flips, while also splitting 20% of the data for validation, with a learning rate of 0.0001.
- **Simpler Custom CNN:** A simpler CNN model as in Fig. 2 includes three sets of convolutional layers with ReLU activation, batch normalization, max-pooling, and dropout, followed by a flattened layer, a dense layer with 512 neurons, and a Softmax output layer for classification.
- **Deeper Custom CNN:** A model includes multiple sets of convolutional layers with ReLU activation, batch normalization, max-pooling, and dropout, with an additional set of convolutional layers for increased depth, followed by a flattened layer, a dense layer with 2048 neurons, and a Softmax output layer for classification as seen in Fig 3.
- **Custom CNN using Residual Blocks:** This model features initial convolutional layers followed by multiple residual blocks with ReLU activation, batch normalization, max-pooling, and dropout layers, culminating in a flattened layer, a dense layer with 1024 neurons, and a Softmax output layer for classification seen in Fig. 4.
- **Custom CNN using Separable Convolutions:** This CNN model utilises separable convolutional as seen inf Fig. 5 layers with ReLU activation, batch normalization, max-pooling, and dropout layers, followed by a flattened layer, a dense layer with 1024 neurons, and a Softmax output layer for classification.
- **Custom CNN using Global Average Pooling:** This CNN model in Fig. 6 includes convolutional layers with ReLU activation, batch normalization, max-pooling, and dropout, followed by a global average pooling layer, a dense layer with 1024 neurons, and a Softmax output layer for classification.
- **Transfer Learning with VGG16:** This is the model utilising VGG16 base (with pre-trained ImageNet weights and non-trainable layers) for feature extraction, followed by custom fully connected layers with ReLU activation and dropout for regularization, culminating in a Softmax output layer for classification as seen inf Fig. 7
- **Transfer Learning with ResNet50:** This model seen in

Fig. 8 utilises the ResNet50V2 base (with pre-trained ImageNet weights and the last 50 layers trainable) for feature extraction, followed by custom layers including dropout, batch normalization, a dense layer with ReLU activation, and a Softmax output layer for classification.

C. Parameter Settings and Configurations

D. Training and Validation Process

E. Experimental Setup

F. Hardware and Software Environment

G. Implementation Details

H. Evaluation Metrics

IV. Results and Discussion

A. Model Performance Comparison

B. Impact of Parameter Variations

C. Analysis of Findings

V. Conclusion

A. Future Work

B. Figures and Tables

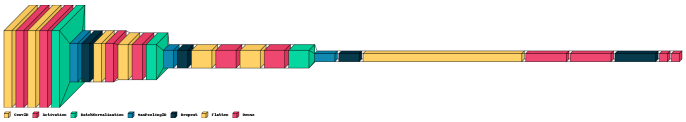

Fig. 1. Custom CNN architecture

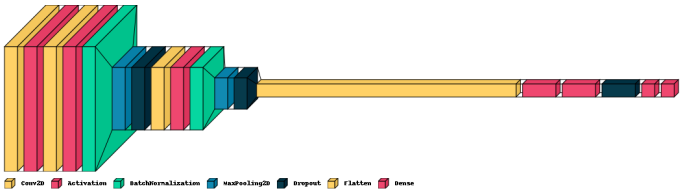

Fig. 2. Simpler custom CNN architecture

Fig. 3. Deeper custom CNN architecture

Fig. 4. Custom CNN using Residual Blocks

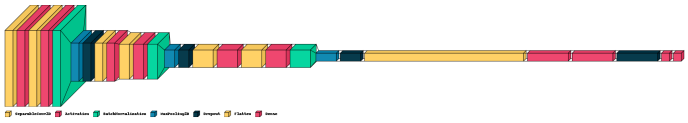

Fig. 5. Custom CNN using Separable Convolutions

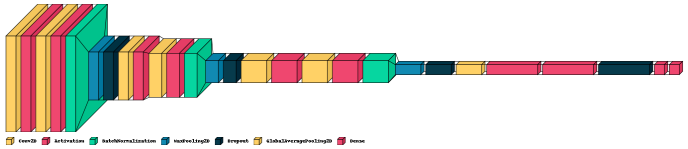

Fig. 6. Custom CNN using Global Average Pooling

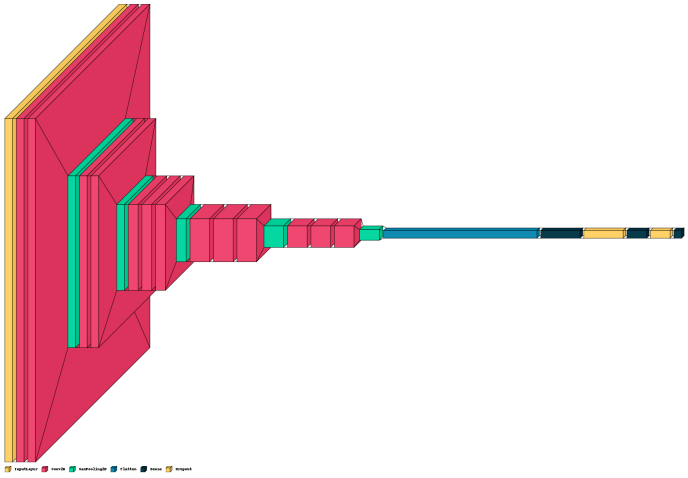

Fig. 7. Transfer Learning with VGG16

Fig. 8. Transfer Learning with ResNet50

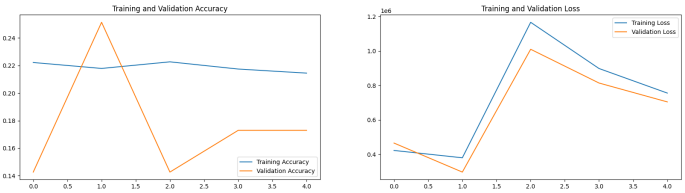

Fig. 9. Train/Validation Accuracy & Loss Graphs for Custom CNN Trained with Learning Rate of 1

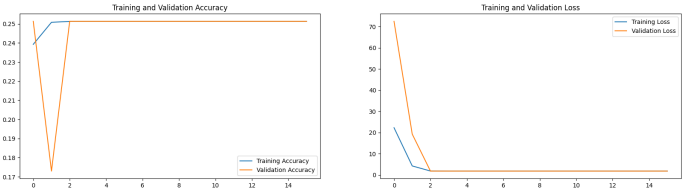

Fig. 10. Train/Validation Accuracy & Loss Graphs for Custom CNN Trained with Learning Rate of 0.01

TABLE I
PERFORMANCE OF DIFFERENT CNN ARCHITECTURES

Model	Learning rate	Train Accuracy (%)	Validation Accuracy (%)	Loss
CNN from scratch	1	25.13	24.71	296462.0625
	0.01	25.13	24.71	1.8133
	0.0001	60.92	54.18	1.9963
CNN from scratch + Image Augmentation	0.0001	67.18	64.92	1.0603
Simpler CNN from scratch	0.0001	63.63	51.32	1.4897
Deeper CNN from scratch	0.0001	61.15	57.01	2.2941
CNN with Residual Blocks	0.0001	54.61	51.53	3.5216
CNN with Separable Convolutions	0.0001	71.80	55.54	1.2115
CNN with Global Average Pooling	0.0001	64.59	58.74	1.4685
Transfer Learning VVG16	0.0001	59.23	57.86	1.1077
Transfer Learning Resnet50	0.0001	70.11	65.78	0.9542

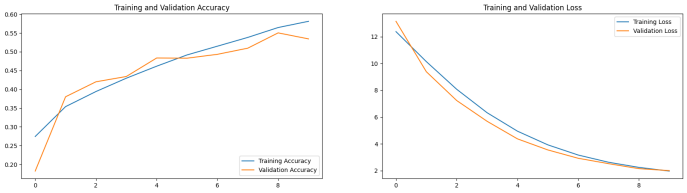

Fig. 11. Train/Validation Accuracy & Loss Graphs for Custom CNN Trained with Learning Rate of 0.0001

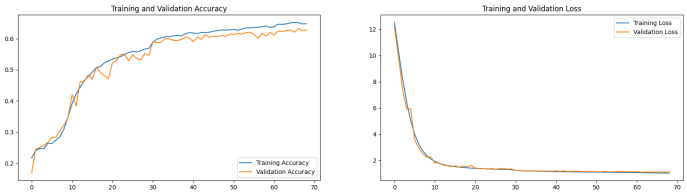

Fig. 12. Train/Validation Accuracy & Loss Graphs for Custom CNN Trained with Augmented Data

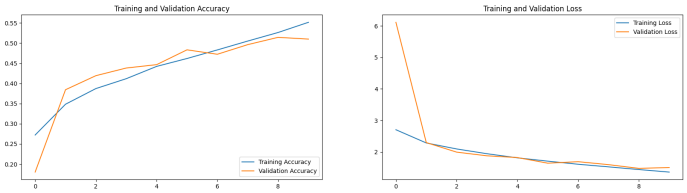

Fig. 13. Train/Validation Accuracy & Loss Graphs for Simpler Custom CNN

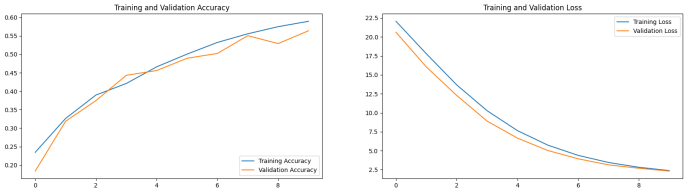

Fig. 14. Train/Validation Accuracy & Loss Graphs for Deeper Custom CNN

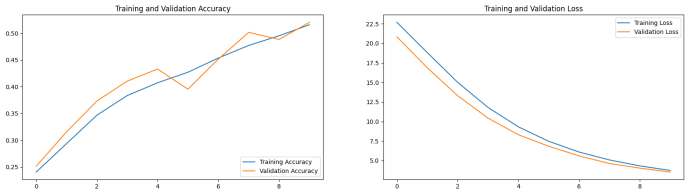

Fig. 15. Train/Validation Accuracy & Loss Graphs for Custom CNN with Residual Blocks

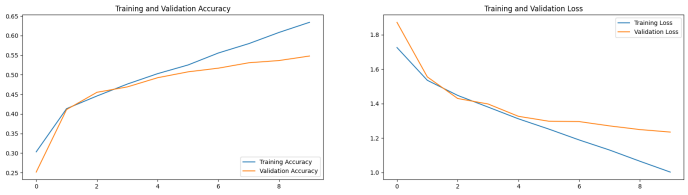

Fig. 16. Train/Validation Accuracy & Loss Graphs for Custom CNN with Separable Convolutions

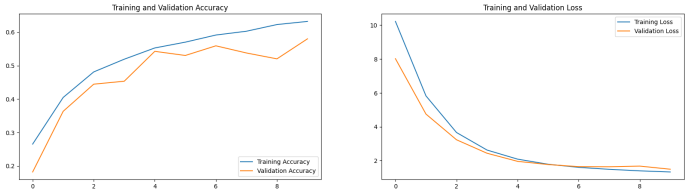

Fig. 17. Train/Validation Accuracy & Loss Graphs for Custom CNN with Global Average Pooling

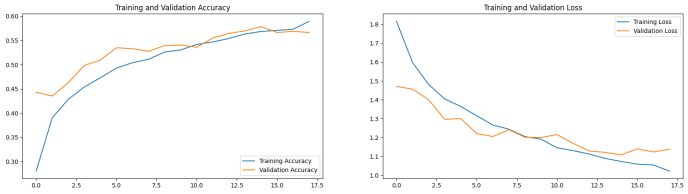

Fig. 18. Train/Validation Accuracy & Loss Graphs for Transfer Learning with VGG16

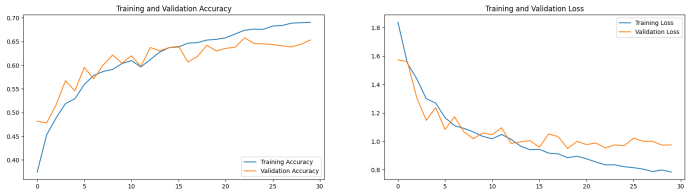

Fig. 19. Train/Validation Accuracy & Loss Graphs for Transfer Learning with ResNet50

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2001, vol. 1, pp. I-511–I-518.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," Pattern Recognit., vol. 36, no. 1, pp. 259–275, Jan. 2003.

[3] S. R. J. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio, "Combining modality specific deep neural networks for emotion recognition in video," in Proc. ACM Int. Conf. Multimodal Interact., 2013, pp. 543–550.

[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.-Mar. 2019.

[5] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2852–2861.

[6] S. Wang, Y. Guo, Y. Yang, L. Yu, and C. Zhang, "Emotion recognition by a fully end-to-end deep neural network with attention mechanism," in Proc. Int. Joint Conf. Neural Netw., 2020, pp. 1–7.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent., 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708.

[12] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," arXiv preprint arXiv:1902.01019, 2019.

[13] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[14] S. Li, W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affect. Comput., vol. 13, no. 1, pp. 119-134, Jan.-Mar. 2022.