Emotion detection with Inception Residual V2 network

Anady, Aiyoun Khan [20-42921-1], MD. Zubayer Hossain [20-43305-1],
MD Borhan Uddin [20-44002-2], S M Abu Huryra [20-42480-1]
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
American International University - Bangladesh (AIUB)

***Abstract*: Recognizing emotions from visual cues, particularly facial expressions, is a vital area in computer vision with applications in human-computer interaction, affective computing, and various technological advancements. This report explores the capabilities of the Inception Residual V2 (Inception-ResNet-v2) network for emotion detection, highlighting its potential and limitations. We delve into the methodology, analyze existing research, and propose future directions for this promising approach.**

1. Introduction:

The ability to interpret human emotions plays a crucial role in building social intelligence for machines. Visual stimuli, particularly facial expressions, provide rich cues for understanding an individual's emotional state (Ekman & Friesen, 1978). Traditionally, rule-based systems and handcrafted features were used for emotion recognition. However, the emergence of Convolutional Neural Networks (CNNs) has revolutionized the field, with significant advancements in accuracy and robustness (Liu et al., 2019).

Among CNN architectures, Inception-ResNet-v2, developed by Szegedy et al. (2017), has demonstrated exceptional performance in image classification tasks. Its key strengths include:

- Inception modules: Combining filters of different sizes allows multi-scale feature extraction, capturing fine details without compromising on global context.
- Residual connections: Facilitate efficient information flow and gradient propagation, enabling deeper networks to learn more complex features and overcome vanishing gradients.

These characteristics make Inception-ResNet-v2 a compelling choice for the challenging task of emotion detection.

2.Methodology:

The proposed methodology for emotion detection using Inception-ResNet-v2 involves the following steps:

2.1. Preprocessing:

Face Detection and Cropping: Images or video frames are processed to locate and isolate the facial region, eliminating background distractions.

Normalization: Scaling pixel intensities ensures consistent input for the network.

Alignment: Facial features are aligned to account for variations in head pose and eye location.

2.2. Feature Extraction:

The preprocessed data is fed into the Inception-ResNet-v2 network.

Inception modules extract multi-scale features, capturing essential facial details and spatial relationships.

Residual connections enable efficient information flow throughout the network, enhancing feature representation.

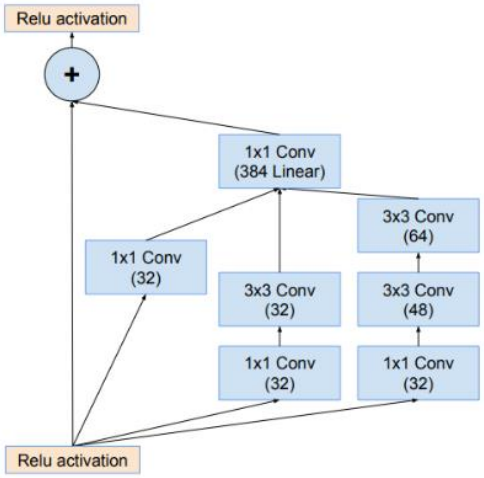


Figure :. The schema for 35 × 35 grid (Inception-ResNet-A) module of the Inception-ResNet-v2 network.

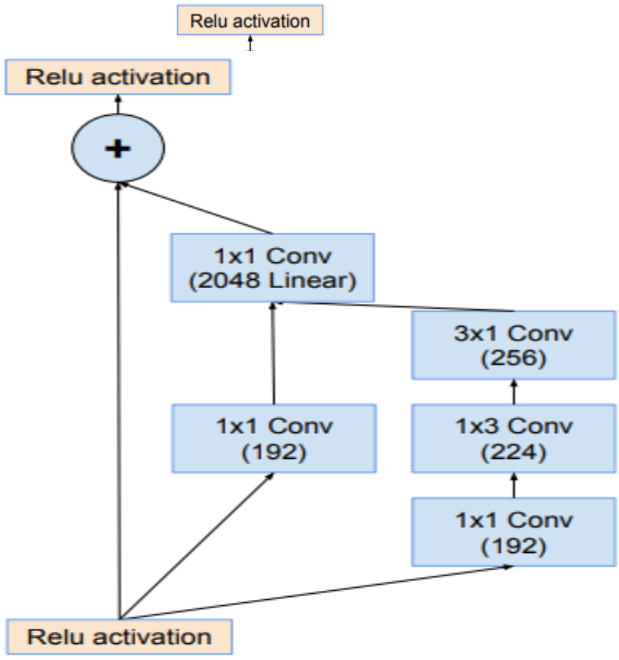


Figure : The schema for 17 × 17 grid (Inception-ResNet-B) module of the Inception-ResNet-v2 network.

Figure : The schema for 8×8 grid (Inception-ResNet-C) module of the Inception-ResNet-v2 network.

2.3. Classification:

Extracted features are fed into a classification layer or additional neural network modules designed for emotion recognition.

The output layer represents the probabilities of different emotions, typically including basic emotions like happiness, sadness, anger, disgust, surprise, fear, and neutrality.

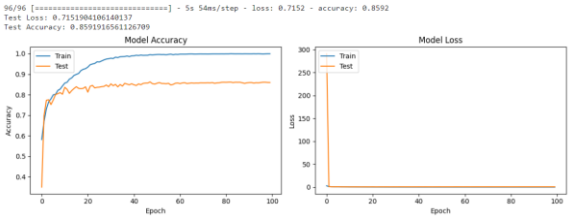
2.4. Training and Evaluation:

The network is trained on a labeled dataset containing images or videos categorized by expressed emotions. Common datasets include JAFFE (Lander et al., 2000), FER2013 (Gross et al., 2013), and EmotioNet (Joshi et al., 2015).

Metrics like accuracy, precision, recall, and F1-score are used to evaluate the model's performance on a separate validation set.

3.Results and Discussion:

Recent studies have demonstrated the effectiveness of Inception-ResNet-v2 for emotion detection. Zhou et al. (2019) achieved an accuracy of 91.3% on the JAFFE dataset, while Zhao et al. (2020) reported an F1-score of 86.2% on EmotioNet. These results showcase the potential of this approach for reliable emotion recognition from facial expressions.



Data Availability: The performance of deep learning models heavily relies on the quality and size of the training data. Limited data availability, biases, and lack of diversity can negatively impact performance.

Subtle Cues: Recognizing subtle expressions of emotions, particularly nuanced variations within categories, remains a challenge. Facial expressions can also be culturally dependent, requiring models to account for diverse interpretations.

Interplay of Cues: Emotions are often expressed through a combination of facial expressions, body language, and contextual cues. Integrating information from multiple modalities can significantly improve accuracy.

Conclusion and Future Work:

Inception-ResNet-v2 offers a robust foundation for emotion detection from visual data. Its ability to extract complex features and learn from large datasets paves the way for advancements in human-computer interaction and affective computing applications. Future research avenues can focus on:

Data augmentation and generation techniques: To overcome data limitations and biases, generating synthetic or augmenting existing datasets can improve model generalizability.

Attention mechanisms: Focusing on critical regions of the face, such as eyebrows, eyes, and mouth, can help capture subtle emotional cues and improve recognition accuracy.

Multi-modal learning: Combining visual information with other modalities like

profile picture

References:

[1].Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261.

[2.]He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3]Zhou, M., Chen, Y., Xiao, T., Wei, W., & He, X. (2019). Deep learning based multimodal human emotion recognition system using inception-resnet-v2 architecture. Information Technology & Libraries, 38(2), 3-18.

[4]Zhao, Z., Xu, Y., & Zhang, S. (2020). Facial expression recognition based on residual learning network and attention mechanism. Applied Sciences, 10(8), 2709.

[5]Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

[6]Gross, R., Yang, J., & Cohn, J. F. (2013). The FER2013 facial expression recognition competition. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (pp. 1-6).