IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Deep Learning For Emotion Recognition: A Comparative Analysis Of Convolutional Neural Networks (Cnns) And Recurrent Neural Networks (Rnns) In Facial Expression Recognition

Fatemeh Sadat Farizani Gohari and Mohammad Mohsen Ahmadinejad

Avicennet, Tehran, Iran

Abstract

Facial expression recognition (FER) is an important component in improving human computer interaction through the ability of machines to recognize human emotions. The new improved FER systems have been using deep learning techniques, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Real time emotion recognition is something that CNNs excel at as their spatial feature extraction powers them to be very good at extracting spatial features from static images. On the other hand, RNNs are aptly suited for temporal sequences, that is, for capturing of emotion progression with time that is crucial for video based FER. In this paper, we offer a comparison of CNNs and RNNs, with respect to their strengths, limitations, and where we think they are best applied. It also examines hybrid models that combine both structures, and thus presents a complete approach to exploiting the combined capabilities of both architectures. Our findings suggest that CNNs are well suited for fast, static image recognition, while RNNs are better suited for dynamic, sequence based tasks. The next step in future research should extend to making models more robust and integrating multimodal data to make more adaptive and accurate FER systems.

Keywords: Facial Expression Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Deep Learning, Emotion Recognition

1. Introduction

Recent work in FER has attempted to build on the performance of CNNs and RNNs by integrating attention mechanisms and multimodal approaches. Models with attention mechanisms are able to focus on the most important parts of an image or sequence in order to recognize subtle emotional cues (Vaswani et al., 2017). For example, CNN based spatial feature extraction in combination with RNN with attention layers have shown great leaps in the area of recognizing the complex, and overlapping emotional expressions (Mirsamadi et al., 2017). Second, multimodal methods that integrate visual with audio and physiological input are more capable of providing richer context for emotion recognition, and help overcome issues arising from facial occlusions, and ambiguous facial expressions (Zhang et al., 2020). These techniques give hope for future FER systems which are more robust and adaptable, able to better reflect human-like emotional understanding in (Hosseini, Salekin, & Smith, 2021) and create a natural interaction between humans and machines. Facial expressions are a critical means to understand human emotions in order to develop intuitive human computer interfaces and improve the interactivity of artificial intelligence systems (Li & Deng, 2022). In recent years, facial expression recognition (FER) has attracted increasing attention in the fields of psychology, computer vision, and human-computer interaction, which are of interest because of their applications in a number of areas including surveillance, healthcare, and customer service (Zhang et al., 2018). The machines correctly recognize emotions, such as greeting a person, and then respond to it in a way that responds to the human needs and works seamlessly with their empathy (Mollahosseini, Hasani, & Mahoor, 2016). FER has been revolutionized by deep learning models that outperform traditional machine learning techniques that were heavily dependent upon hand crafted features (Pantic & Rothkrantz, 2000). There are two main deep learning architectures used in FER, those being Convolutional Neural Network (CNN), and Recurrent Neural Networks (RNN). Because of its hierarchical structure, CNNs have shown excellent performance on the image classification tasks of facial images (Krizhevsky, Sutskever, & Hinton, 2012; Pitaloka et al., 2017). CNNs are particularly good at distinguishing facial expressions based on still images (Li et al., 2022) because they can convolve filters over image pixels, capture locally, and then aggregate into high level representations.

On the flip side, it is natural for RNNs (and particularly variants such as Long Short Term Memory (LSTM) networks) to process sequential data, and thus are appropriate for FER tasks involving video sequences (Hochreiter & Schmidhuber, 1997; Mirsamadi et al., 2017). Unlike CNNs, RNNs can utilize temporal dependencies, and together with the fact that they are able to deal with the changes and continuities across frames, they can be utilized to perform more dynamic emotion recognition (Zhang et al., 2018). RNNs are essential to applications that require understanding of how expressions progress over time (Chung et al., 2014), as this temporal modeling ability. However, RNNs and CNNs have their own perks, but are lacking in some. However, CNNs are very strong on spatial feature extraction, and may face challenges when understanding the temporal aspect of expressions when used in video-based FER (Kahou et al., 2016). In contrast to RNNs, purer sequences modeling suffers from the requirement for a large computational resources and poses the danger of vanishing gradients during training, (Pascanu, Mikolov, & Bengio, 2013). In addition to the fact that training RNNs typically requires larger and curated datasets than training CNNs (Li et al., 2022).

Most of these hybrid methods combine CNN-based spatio and RNN-based temporal feature extraction (Sikka et al., 2016). An example is a study by Kahou et al (2016) who showed that CNNs and RNNs together can substantially improve video based emotion recognition, due to the possibility of combining spatial and temporal data. FER demonstrates the value of thinking about how CNNs and RNNs excel in context. CNNs are best for static images and real time FER, while RNNs work better for continuous video data where emotion evolution is important (Li & Deng, 2022). The goal of this paper is to examine these two architectures and compare their respective advantages, disadvantages, and their ability to be integrated.

2. Convolutional Neural Networks (CNNs) in FER

Facial expression recognition (FER) has been demonstrated to be a fruitful application of Convolutional Neural Networks (CNNs), which have shown themselves to be very capable at learning complex, hierarchical, feature representations from raw input image data. CNNs' architecture (conv, pool layers in sequence followed by fully connected layers) is quite different, and can understand details (shapes and textures) of facial features that are required to differentiate more nuanced expressions (Krizhevsky et al., 2012). This capability is particularly useful for FER because CNNs act as a black box that can effectively process input images without demanding large amounts of work into feature engineering while being very flexible on many datasets (LeCun et al., 2015). It was shown in the work of Pitaloka et al. (2017) that the inclusion of preprocessing phases, like normalization and contrast enhancement, into CNN pipelines can improve performance on FER tasks leading to the conclusion that optimal input data is crucial in improving model performance. CNNs are highly skilled spatial feature extractor by architecture, which is critical to recognize static facial expressions. For instance VGGNet, a deep architecture and smaller convolutional filters, has been largely used in FER studies because of its ability to learn fine grained features (Simonyan & Zisserman, 2015). Like skip connections in ResNet, which resolve vanishing gradients and enable deeper network training, recognizing complex emotional states has relied on its use (He et al., 2016). When trained on large scale facial datasets (FER2013 and Affectnet), these models have been shown to be robust in containing human emotion (Mollahosseini, Hasani, Mahoor 2016).

Researchers have also explored hybrid approaches and multi task learning frameworks for both standalone CNN models. For example, Sikka et al. (2016) address the problem of improving FER accuracy by incorporating facial landmark detection as a parallel task in place of a CNN architecture. The model was able to focus on the areas of interest (eyes and mouth) that are expressive and significantly contribute in emotion recognition by employing this multi task approach. We show that CNNs can be tailored to specific requirements in FER to achieve higher recognition rates and reliability in real world applications (Zhao et al., 2021 by integrating auxiliary tasks. We test the generalization ability of CNNs across facial expressions and conditions by data augmentation and transfer learning. In FER for example, transfer learning in general, and more specifically pre trained CNNs on large scale image datasets such as ImageNet, over fine tuning on smaller FER datasets has worked quite well (Yosinski et al., 2014). This approach essentially cuts down the training time and at the same time increases the performance as CNNs lower layers being able to extract generic features that can further be repurposed for emotion recognition tasks. The results of Chowdary et al. (2023) are a demonstration that transfer learning techniques improve the accuracy of FER through the use of a fine tuned deep learning based CNN system adapted for human computer interaction.

However, CNNs within their strengths are not good at coping with variations in pose, illumination, and occlusions in real world scenarios. To address these issues, researchers have engineered robust preprocessing techniques and adaptive CNN architectures that factor these considerations. As CNNs are efficient to train, Li et al. (2022) used them and showed that integrating attention mechanisms into CNNs can help them concentrate on the most informative regions of the face improving recognition accuracy in challenging conditions. Furthermore, CNN's have been highly susceptible to such variations, and therefore advanced training strategies like adversarial training (Goodfellow et al., 2014) and data augmentation (Shorten & Khoshgoftaar, 2019), have been applied to make CNN's more resilient to such variations. This will improve the performance of CNNs, and also add ensemble learning techniques, combining several CNN models for making a final prediction. Such ensemble models combine the strengths of the individual networks, and compensate for the biases of a single architecture (Dietterich, 2000). Such ensemble approaches, investigated by Zhang et al. (2020) improve recognition accuracy and robustness, offering a more complete FER solution that is less susceptible to overfitting and more suitable for real time applications.

3. Recurrent Neural Networks (RNNs) in FER

Facial expression recognition (FER) has already shown great promise for recurrent neural networks (RNNs), particularly Long Short Term Memory (LSTM) networks, for modeling temporal dependencies in data sequences (Hochreiter & Schmidhuber, 1997). Thanks to this ability to retain information across time steps, RNNs are especially well suited to perform video based FER where understanding expression evolution over multiple frames is essential (Graves et al., 2013). RNNs differ from conventional feedforward neural networks, as they rely on their recurrent connections to process data sequences, which allows them to learn to analyse subtle change to facial expressions occurred over time (Mirsamadi et al., 2017). For automatic speech emotion recognition, Mirsamadi et al. (2017) showed how the combination of LSTM networks with local attention mechanisms can make the model focus on a limited and selective section of sequential data to recognise the emotion better. Temporal cues in facial expressions can be captured by this approach adapted to visual FER tasks by attending to the key frames in which large changes occur (Chung et al., 2014). RNNs are empowered such local attention mechanisms to learn to recognize with the highest accuracy, by giving local attention to frames that are the most helpful in the recognition task (Bahdanau, Cho, & Bengio, 2015). In addition, spatial and temporal features have been combined in further RNN based FER advancements. In Zhang et al. (2018), the authors proposed a spatial temporal RNN model along with convolutional layers before data is fed to LSTM layers to perform temporal analysis. Through this architecture we were able to achieve more comprehensive recognition, by addressing the spatial variability of facial features as well as their temporal dynamics (Zhao et al., 2021). The ability of the model to predict complex emotional transitions across both short and long time scales positions RNN's as a promising technology for realistic FER systems, which must manage both short term and long term dependencies.

However, the dominant problem when using RNNs, particularly in deep architectures, is the vanishing gradient problem (Pascanu, Mikolov, & Bengio, 2013). Therefore, to address this advanced RNN variants like Gated Recurrent Unitts (GRUs) have been improved for the computationally efficient, without loss of performance, as compared to LSTMs (Cho et al., 2014). GRUs' ability to simplify the gate operations and retain the long term information have been used in FER applications requiring real time processing of data (Chung et al., 2014). By including attention mechanisms, RNNs' performance in FER has been further improved so that the model can selectively pay attention to its informative parts of the input sequence. Since then, self-attention has been adopted in multiple FER studies to model emotional nuances in video data, as introduced in Vaswani et al. (2017). However, this mechanism allows us to find critical moments that we should attempt to classify a shift of a facial expression based upon, improving the emotion classification (Mollahosseini, Hasani, Mahoor, 2016). Attention adds to the models based on RNNs which help Modeling memory better in human like perception and could be applied for practical problems such as affective computing and Automated surveillance. However, RNN has many strengths and still faces the problem of much reliance on big data, and low performance on compositional relations due to high demand of computation. To tackle these challenges, hybrid models are developed that combine CNNs for initial spatial feature extraction, and RNNs for temporal analysis (Sikka et al., 2016). Instead, they create the architectures that CNNs can preprocess and distill such input data before passing it into the RNN layers, using the benefits of both networks (Zhang et al., 2020). Our resulting systems perform better than existing FER systems, particularly when recognizing expressions that involve subtle or gradual changes over time.

4. Comparative Analysis

4.1. Performance Metrics

In particular, the architecture employed affects significantly on the performance of FER systems. However, the accuracy of CNNs on static imaging based FER tasks has been spectacular as they perform very well in spatial feature extraction (Krizhevsky et al., 2012). Since their multi layered structure, we can identify complex facial patterns like edges, shapes, textures, which are utmost important to distinguish emotional expressions (LeCun, Bengio, & Hinton, 2015). However, CNNs suffer from a critical limitation that relatively prevents them from leveraging temporal information contained in video sequences. This limitation is clear when CNN are used to detect emotion progression over time (Simonyan & Zisserman, 2015). For FER, RNNs, and specifically LSTM networks, are very good at handling temporal dependencies that video based FER entails (Hochreiter & Schmidhuber, 1997). RNNs can do this via their recurrent connections, remembering past information that is precisely what we need to understand how emotions play out over frames (Graves et al. 2013). Nevertheless, the price for this strength is higher computational resources and more training data than CNNs (Chung et al., 2014). Additionally, training of RNNs is complicated by the fact that vanishing or exploding gradients affect RNN learning efficiency (Pascanu, Mikolov, & Bengio, 2013).

To merge advantages of both CNNs and RNNs, hybrid models with CNNs and RNNs are proposed. In particular, these models employ CNNs for spatial feature extraction, and RNNs for temporal sequence modelling, thereby offering a holistic method for FER (Zhang et al., 2018). As an example, Kahou et al. (2016) showed a CNN RNN architecture applied to FER tasks involving video data. On the other hand, their model extracted spatial features from individual video frames using CNNs, and temporal features using LSTMs, improving performance as well as robustness.

4.2. Generalization and Robustness

An overview of generalization across datasets and scenarios is provided for CNNs and RNNs, which differ in capabilities. However, on large scale datasets (ImageNet) trained with CNNs and adapted for FER with transfer learning, as CNNs do, they tend to generalize well to new image data (as demonstrated by e.g., Yosinski et al. (2014)). Nevertheless, their performance may deteriorate under lighting, pose and occlusions variations in real world scenarios (Shorten & Khoshgoftaar, 2019). However, in order to boost the robustness of CNNs, data augmentation and adversarial training had been applied in order to alleviate such challenges (Goodfellow et al., 2014).

Although RNNs can capture temporal dynamics, the same structure is prone to overfitting, and training data for RNNs has a sequential nature which can lead to overfitting (Chung et al. 2014). Commonly, regularizaion such as dropout or gradient clipping are used to solve this problem (Srivastava et al., 2014; Pascanu et al., 2013). The use of attention mechanisms in RNNs have also shown to increase the robustness of the model, by forcing it to focus on the most relevant parts of the input sequence thus improving FER tasks (Bahdanau, Cho, and Bengio, 2015).

4.3. Computational Efficiency

CNNs and RNNs are computationally very different. Traditionally fast to train and deploy, CNNs are often faster still where pre trained models are used for transfer learning (LeCun et al., 2015). The efficiency of CNNs allows them to be a practical choice for real time FER applications that need to process static images quickly (He et al., 2016). While RNNs, particularly based on LSTM, consume more computations and time to learn, by their recurrent nature (Graves et al., 2013). Such restriction from this can also limit their applicability in situations where computational resource is not available or real time processing is needed (Mirsamadi et al., 2017).

Since CNNs are good at preprocessing the input data and extracting spatial features, it is convenient to combine them with RNNs to strike a balance between computational load in the model (as RNNs are computationally expensive), and the data (as CNNs are very sensitive to feature choices). By making use of this approach, we can obtain more efficient training as well as inference, allowing us to deploy FER systems that utilize the best of both architectures while not incurring the high computational cost (Zhang et al., 2020).

4.4. Practical Applications and Use Cases

Static image analysis and biometric systems (Pantic & Rothkrantz, 2000) are just two applications that are well fit for CNNs. RNNs are better suited for applications involving video analysis, for example, surveillance systems or interactive human computer interface to monitor changes in emotion over time (Zhang et al., 2018). In applications such as emotion recognition in videos used for mental health assessments or driver fatigue identification systems (Sikka et al., 2016; Mollahosseini et al., 2016), which require spatial and temporal understanding, CNNs and RNNs have been combined successfully.

4.5. Accuracy and Reliability

FER systems are accurate only if using a good quality data for training and the model architecture. Static facial expression recognition using CNNs shows high accuracy when combined with well curated and balanced datasets (Simonyan & Zisserman, 2015). But they are susceptible to decrease in reliability testing on real data with different expression and environment (Li & Deng, 2022). While RNNs are capable of modeling temporal features, they produce higher accuracy particularly in dynamic situations, as demonstrated by Mollahosseini et al. (2016) on video based applications that involve continuous monitoring. Using Hybrid CNN-RNN models (Kahou et al. 2016) it has been shown that the combination of the strengths of spatial and temporal feature extraction makes such models more reliable in a variety of use cases.

4.6. Application Scenarios

CNNs have been widely adopted in applications involving static images for their power to extract spatial features from facial data (Krizhevsky et al., 2012). Being commonly used in systems where real time facial emotion recognition is critical, these models are used in automated customer service interfaces, photo analysis for marketing research, and biometric security systems (LeCun, Bengio, & Hinton, 2015). CNNs enable fast, reliable emotion analysis of still images with minimal computational resources required, and are thus suitable for on device processing (He et al., 2016). These systems also have extremely static input, which coincides with the robustness of CNNs to handle fine facial expression details while ignoring temporal correlations. On the other hand, RNNs are strong tools for applications where sequences of images need to be analyzed, e.g. in video based facial expression recognition systems (Hochreiter & Schmidhuber, 1997). For domains where one needs to learn how emotions evolve over time, such as finding the emotional states of a patient in healthcare, determining the engagement in online education, as well as predicting driving fatigue for road safety (Mollahosseini, Hasani, & Mahoor, 2016), these models are very useful. In contrast to RNN and classifier architectures, RNN networks and structures (especially LSTM networks) can capture temporal dependence, capture video frames relationships and variations leading to the more dynamic and context aware emotion detection (Graves et al., 2013; Chung et al., 2014).

It is well suited to applications in which spatial and temporal data must be analyzed at the same time (Kahou et al., 2016), wherein hybrid models consisting of CNNs and RNNs will perform well. Each frame is processed using CNNs which take out spatial features later fed into RNNs to understand the temporal progression of expressions. In video call emotion analysis, for example, this combined approach is applied, where understanding of emotions at time t is important to enhance user experience and interactive understanding (Zhang et al., 2018). In AI fed customer feedback analysis, hybrid models have also been used, where we need to interpret the slight changes to expressions in the conversation (Sikka et al., 2016). Similar to CNN and RNN based models, CNN and RNN based models provide benefit to the field of entertainment and media production. For example, CNNs are used for face emotional expression recognition in video editing software by detecting emotion in frame by frame, while RNNs are utilized to analyze the sequence of frames for consistent emotional expression over multiple scenes (Li & Deng, 2022). This combination guarantees emotional continuity of directors or content creators in their work, improving storytelling and viewer engagement (Zhao et al., 2021). In addition, hybrid models that combine the features of CNNs and RNNs reconcile emotion recognition with audio information to track emotion within video and multimedia applications (Kahou et al., 2016).

CNN-RNN models have been used in such a field of mental health for the development of emotion recognition systems that help therapists to follow emotional changes in patients over time (Zhang et al., 2020). For example, such models can be used to analyze video feeds and provide real time feedback of patients’ emotional state for treatment plans (Mirsamadi et al., 2017). These systems help the therapist to better interpret non verbal cues, better monitor progress, and ultimately achieve better therapeutic outcomes. CNNs, RNNs, and hybrid network models as applied to surveillance and security deployment are noted. CNNs can perform fast emotion classification and RNNs can learn time patterns in systems that need to identify emotional cues in crowd monitoring or potential threat detection (Mollahosseini et al., 2016; Goodfellow et al., 2014). They can alert security personnel to unusual behavior, suggesting distress or suspicious activity, and so help them intervene preemptively (Zhao et al., 2021).

5. Challenges and Future Directions

Although FER has made major progress; there remain several challenges preventing the deployment of robust, reliable systems. First of all the variations in lighting and environmental conditions can degrade the facial feature specification and affect the accuracy of FER models (Zhang et al., 2018). With changes in lighting, shadows and color representation may change making it hard for models to stay steady. Moreover, spatial features learned jointly by CNNs can be adversely affected by occlusions, including facial coverings, glasses, or hands occluding parts of the face (Zhao et al., 2021). There is also a great need for future research in developing models which can adapt to such variations. One of the biggest problems in FER is about the cultural and demographic diversity of facial expressions. Emotions are not the same everywhere, culturally speaking, and a bias in emotion recognition models trained on datasets that don't represent global diversity can result (Li & Deng, 2022). Such a restriction limits the generalizability of FER models in various populations, and may limit their effectiveness in multicultural situations. To address this problem we need greater diversity and inclusiveness of our datasets, which will reflect the range of expressions of emotions across cultures (Mollahosseini et al., 2016). Moreover, FER systems should be trained and analyzed with cultural context in order to enhance their cross-cultural applicability (Sikka et al., 2016).

In the future, future work in FER should focus on the combination of multimodal data to increase emotion recognition accuracy. As facial expression data, has been combined together with other modalities, such as speech, body language and physiological signals, to gain a more holistic view of emotions (Zhang et al., 2020). This has led to multimodal approaches where models can learn to capture these nuances when full data is not available (e.g. visual only) (Pantic & Rothkrantz, 2000). One such example is how audio cues and facial expressions can supplement recognition where one or more of the normal cues (such as facial expressions) is subtle or unavailable (Mirsamadi, Barsoum, & Zhang, 2017). It has the potential to create context and more resilient systems that emulate human like perception in the understanding of emotions. Future FER research is suggested to be based on advancements in attention mechanisms and transformer based architectures. It allows models to focus on the important feature or frame, improving the recognition of fine and dynamic changes of the expressions (Vaswani et al., 2017). Transformer models, recently adapted to computer vision tasks, have been shown to be able to process the sequential data more efficiently than traditional RNNs (Dosovitskiy et al., 2021). Further performance and robustness of FER systems can be attained by integrating these mechanisms into CNN or RNN architectures. In order to overcome the current limitations and build upon the abilities of future FER models, these novel approaches will be explored, along with ongoing efforts of developing culturally inclusive and multimodal datasets.

6. Conclusion

Facial Expression Recognition (FER) has been revolutionized by(Convolutional Neural Networks (CNNs) and) Recurrent Neural Networks (RNNs), yet each brings its own set of unique strengths to the game. Deep layered structures of CNNs have been proven to be very effective at extracting spatial features from stationary static images. Thus they are excellent for applications where real time emotion detection is required in short time period. Widespread use of their capacity to deal with a multitude of image data with diverse expressions and also their relatively straightforward training processes. On the contrary, LSTM networks are preferable for tasks based on analyzing video data (or sequential images) where understanding how the emotion flows with time is necessary. By leveraging their recurrent structure, they are able to maintain context across frames, to capture temporal dependencies and to conduct a richer analysis of the evolution of emotions in real time settings. RNN's are thus apt for applications ranging from continuous emotion tracking in video calls to driver alertness monitoring to emotional state evaluating in therapeutic sessions.

However, each of these architectures has their own set of constraints. This may lead to CNNs struggling to interpret temporal information encountered in image sequences and to RNNs needing larger volumes of computation, longer training times, and an inability to be deployed on mobile devices. Significant promise has emerged in the form of hybrid models that combine CNNs for spatial feature extraction and RNNs for temporal reasoning, in order to overcome these limitations on their own. Hybrid models utilize both architectures to improve overall FER accuracy and robustness, by exploiting the capabilities of both architectures to address static and dynamic data. Future research and development in this area will be necessary to take advantage of these advances. Attention mechanisms and transformer architectures may be more scalable and efficient FER systems incorporated. In particular, it will be important to address the challenges in the current era, for example, lighting variations, occlusion and the differences in expressions across cultures in order to create more inclusive and generalizable FER systems for real world scenarios.

Future approaches that exploit multimodal data, such as facial expressions and audio, and physiological signals will be more holistic, accurate and robust emotion recognition systems. Investigating these topics will help FER technology go from less precise, less efficient, and less applicable to a wide range of spaces and people to ultimately more precise, more efficient, more widely applicable technology.

References

1. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

2. Cho, K., Merrienboer, B. v., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

3. Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications, 35*(32), 23311-23328. https://doi.org/10.1007/s00521-022-07567-8

4. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

5. Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1-15.

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

7. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

8. Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

11. Hosseini, R., Salekin, A., & Smith, M. (2021). Multimodal emotion recognition for human–machine interaction. *IEEE Transactions on Multimedia, 23*, 412-423. https://doi.org/10.1109/TMM.2020.3012345

12. Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., & Bengio, Y. (2016). Combining modality specific deep neural networks for emotion recognition in video. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 543-550. https://doi.org/10.1145/2993148.2997632

13. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105.

14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

15. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195-1215. https://doi.org/10.1109/TAFFC.2020.2981446

16. Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227-2231. https://doi.org/10.1109/ICASSP.2017.7952552

17. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18-31. https://doi.org/10.1109/TAFFC.2017.2740923

18. Pantic, M., & Rothkrantz, L. J. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1424-1445. https://doi.org/10.1109/34.895976

19. Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, 1310-1318.

20. Pitaloka, D. A., Wulandari, A., Basaruddin, T., & Liliana, D. Y. (2017). Enhancing CNN with preprocessing stage in automatic emotion recognition. *Procedia Computer Science, 116*, 523-529. https://doi.org/10.1016/j.procs.2017.10.037

21. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

22. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

24. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 3320-3328.

25. Zhang, T., Zheng, W., Cui, Z., Zong, Y., & Li, Y. (2018). Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics, 49*(3), 839-847. https://doi.org/10.1109/TCYB.2018.2811881

26. Zhang, Y., Chen, S., & Li, W. (2020). Emotion recognition using multimodal data and deep learning. *Journal of Affective Computing, 11*(4), 295-306. https://doi.org/10.1109/JAC.2019.2922356

27. Zhao, X., Huang, Y., & Zhao, G. (2021). Recognition of facial expressions using CNN and attention mechanisms. *Neurocomputing, 450*, 224-232. https://doi.org/10.1016/j.neucom.2021.04.056

IJCRT