Sequence-based Facial Emotion Recognition using EfficientNet and LSTM

Celestine Akpanoko, Alex Esser, Srikanth Narayanan, Chang-Yong Song, Hunter Mast
Vanderbilt University

Abstract

This study advances the field of facial emotion recognition (FER) by enhancing a pre-trained EfficientNet model, originally designed for single-frame image analysis, to accommodate sequence-based inputs using Long Short-Term Memory (LSTM) networks. The integration of LSTM is premised on its ability to capture temporal dynamics, which we hypothesize will significantly improve the classification of the categorization of valence and arousal values across sequences, transforming them into a robust 21-class model ranging from -10 to 10 for each. Our approach modifies the EfficientNet (enet_b0_8_va_mtl.pt) architecture to extract and aggregate spatio-features of every frame in a sequence and then model the temporal features with LSTM, facilitating a deeper understanding of emotional transitions over time. The enhanced model was trained on a AFEW-VA dataset aimed at evaluating our combined EfficientNET and LSTM's efficacy in capturing fluctuating emotional states, with performance compared against other baseline CNN-LSTM models. Results from half of the dataset indicate substantial improvements in accuracy for both training and validation phases, underscoring the methodological benefits of incorporating sequential and temporal analysis in affective computing. This development not only elevates the precision in recognizing complex emotional expressions but also highlights the value of temporal integration in enhancing the capabilities of static image-based FER models.

1 Introduction

Emotions play a crucial part in our everyday communication. It drives fast-growing research in many fields like education, psychology, human-computer interaction, and robotics. The field of Affective Computing has been paired with one or more different fields in recent emotional research. Facial emotion recognition (FER) has been the driving force of research in Affective Computing because our faces contribute more to our emotions than any other part of our body (2). FER is described as a computer vision task with the goal of being able to recognize and categorize expressions into different categories on the human emotional spectrum (1). Providing real-time analysis of emotions allows us to cross facial landmark movements such as eyes, mouth, and eyebrows related to discrete emotions such as joy, sadness, and anger (1). The amount of depth this kind of system provides is endless.

However, despite the importance of facial emotion recognition, there is still much to be desired in terms of creating such systems. Alone, FER can only provide so much recognition and requires other networks to run efficiently. Early on, FER systems were built off methods such as Gabor filters, haar-like features and Local Binary Patterns (LBP). Many issues stem from this as predefined features are incapable of cooperating with data from multiple applications. A deep learning method such as a convolutional neural network (CNN) could be used to integrate feature extraction and determining expressions into a single procedure. CNNs can contribute to an improvement in recognition and accuracy over other methods. With this, translating a video to a sequence of static images may lead to some issues. Facial expressions are dynamic and can change fast, so analyzing images can lead to many issues with this recognition quickly.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Through extensive research, the integration of Long Short-Term Memory (LSTM) networks into different models has made remarkable progress in the realm of facial emotion recognition (FER). As a specialized form of Recurrent Neural Network (RNN), LSTM networks are specifically designed to tackle the vanishing gradient problem that often affects traditional RNNs. Their architecture allows for long-term learning and retention of information, which is essential for effectively handling time-series data.

For our project, the utilization of LSTM networks is especially beneficial as it aligns with our goal of improving emotion prediction models to evaluate emotions over a second interval, rather than in individual frames. Understanding the importance of a temporal approach is crucial as it enables a deeper understanding of emotions and expressions by taking into account their changes over time. Conventional emotion recognition methods tend to disregard the contextual and progressive aspects of human emotions, which can result in a misinterpretation of one's emotional state. With the integration of LSTM networks, our model becomes more adept at capturing the ever-changing and continuous stream of emotional expressions, resulting in a more precise and lifelike evaluation of emotional states.

We hypothesize that leveraging the aggregated spatial features from a pre-trained EfficientNet model, combined with the temporal modeling capabilities of a LSTM network, will enable a more contextual and continuous analysis of emotional expressions. This approach is anticipated to improve the precision of valence-arousal emotion predictions, in contrast to methods that evaluate each frame separately. With the incorporation of LSTM technology, the enhanced model strives to replicate the human approach to emotional assessment by capturing the time-based changes in expressions. This integration holds great promise for enhancing our comprehension of human emotions, potentially paving the way for the creation of more advanced and precise emotion recognition systems. This hypothesis forms the basis of our study, aiming to connect human emotional perception with automated emotional intelligence.

2 Related Work

The field of facial expression recognition (FER) has seen significant progress due to extensive research and collaborative competitions like the Affective Behavior Analysis in-the-wild (ABAW). The 2023 iteration of the ABAW competition showcases the latest advancements, with participants tackling complex challenges using datasets such as Aff-Wild2 and the Hume-Reaction. The competition required participants to tackle four different challenges: valence-arousal estimation, expression classification, action unit detection, and emotional reaction intensity estimation (9). These tasks showcase the various methods used by experts in the field to improve the precision and practicality of FER systems. With the incorporation of these methods into our research, we have established a strong foundation for enhancing machines' comprehension of human emotions.

Hong Guo and Jiayou Chen of Wuhan University of Science and Technology looked into dynamic FER systems based on ResNet residual neural network and LSTM. In this research, the both of them would try to create a system that views emotions in a dynamic way, rather than the usual static image data typically used. The ResNet network would allow for better training on the CNN model, which when used with the LSTM network, would capture dynamic sequence data inside of the CNN model. The facial expression change would be captured in a video sequence for analysis and processing. Overall, it would provide a better application of this data in FER environments. This idea is something we considered using when creating our FER system. We used a sequential-based FER to allow for lots of image data to be collected, even though it is not dynamic. It allows for a simpler computation and deep learning model design (5).

Ye Ming, Hu Qian, and Liu Guanyuan of Southwest University in Chongqing, China wrote research regarding FER systems using CNN-LSTM and a two-player attention mechanism. It tries to attack the issue regarding current algorithms being insufficient in using information and data of the emotions being expressed. To solve this, combine a CNN and LSTM into a CNN-LSTM to allocate and analyze the data given by the FER. A CNN-ALSTM is used to increase accuracy. Along with this, another version is created utilizing a ACNN-ALSTM model to manage two-layers of attention mechanisms for a more accurate analysis of emotions. Results showed that using this ACNN-ALSTM as a hybrid neural network model was superior to other works due to network depth and hidden layer nodes improving accuracy 2% to 4% more. The hidden layer nodes do have an upper limit of around 1024, but still proved to be more efficient than current models (12).

Ryo Miyoshi and Manabu Hashimoto from Chukyo University and Noriko Nagata from Kwansei Gakuin University researched FER from video using a ConvLSTM algorithm. It was proposed that adding skip connections in spatial and temporal directions to most ConvLSTM happen to remove gradient vanishing and older data for creating an enhanced ConvLSTM. Another method proposed in this paper is one that will automatically recognize facial expressions from videos taken. Similar to Hong Guo and Liu Guanyuan's paper, FER would dynamically analyze videos using two enhanced ConvLSTM networks and two ResNet networks. This was tested by comparing both the enhanced method and normal method, showing that the enhanced ConvLSTM achieved 4.44% higher accuracy to the regular method. The results of the FER method for dynamically viewing videos for emotions analysis showed that this method had a 45.29% accuracy, which was higher by 2.31% compared to normal ConvLSTM methods. This lines up with the research by Hong Guo and Liu Guanyuan and how dynamically viewing videos leads to a much better accuracy in facial emotional recognition (7).

Rajesh Sighand Anil Vohra from Kurukshetra University, Sumeet Saurav, Ravi Saini, and Sanjay Singh from CSIR-Central Electronics Engineering Research Institute, and finally Tarun Kumar from Birla-Institute of Technology of Science researched FER in videos using hybrid CNN and ConvLSTM. A 3D-CNN is used with a LSTM network to perform more efficiently than some other video-based facial expression recognition (VFER). Similar techniques are used in a few other papers referenced here. A fully-connected LSTM (FC-LSTM) unrolls an image to a one-dimensional vector. Loss created by the FC-LSTM is avoided due to a ConvLSTM not unrolling the said image. Combining 3D-CNN and ConvLSTM for VFER leads to a hybrid network with spatiotemporal data from these video sequences. The results show that the accuracy of combining multiple differences was around 43.86%, so around what other VFER system's accuracy is (15).

Moshsin Kabir, Tanvir Ahamed Anik, Shahnewaz Abid, and M. F. Mridha from Bangladesh University of Business and Technology and Abdul Hamid from King Abdulasziz University researched a CNN-LSTM approach using FER. This paper discusses how non-posed images contain non-verbal information used to evaluate the mental state of individuals in face-to-face communications. Posed and non-posed facial expression (PNFE) dataset would be built by them to try and evaluate the best performance of these sorts of systems. It introduces the concept of using a convolutional neural network (CNN) with a LSTM to classify expressions such as happiness, anger, disgust, fear, sadness, and surprise. This combination would be used for having CNN learn on the PNFE dataset, then have LSTM bound the relationship between the images and expressions. Overall, this shows the possibilities that LSTM networks can have when combined with different networks (8).

Dandan Liang, Huagang Liang, and Tipu Zhang from Chang'an University and Zhenbo Yu from Nanjing University researched deep convolutional BiLSTM fusion networks for facial expression recognition. With this, they wanted to focus on avoiding what most deep learning methods do with FER and not focus on spatial appearance features for categorization. This includes bidirectional features with the LSTM to use both spatial and temporal information at the same time known as BiLSTM. A framework was created by them to learn spatial features in more of a joint manner. It extracts these spatial features each frame and uses temporal dynamics with a CNN to analyze the data. This will allow for fused features and a framework that is learnable with temporal data for adapting to these features (11).

Jingwei Yan and Wenming Zheng from Southeast University, Zhen Cui from Nanjing University of Science and Technology, and Peng Song from Yantai University researched a joint convolutional bidirectional LSTM framework for FER systems. In this paper, it looks into the relationship between facial regions and spatial dependencies. FER views these different regions to analyze facial muscle movement and categorize different expressions into the emotional spectrum. The joint convolutional bidirectional LSTM (JCBLSTM) framework models facial textures and their spatial relation between different facial regions. Mapping output to a CNN as a sequence and using LSTM to model the dependencies, joint feature representation is used to combine all of these representations to a single representation. In the end, it proves that this sort of LSTM is able to achieve similar results to normal LSTM networks while lowering data usage and the effectiveness of understanding spatial relation data (17).

As we can see with FER systems and LSTM networks today, we can notice a lot of similarities and common features most of them involve. CNNs and LSTMs are vital to creating such systems due to their sequential methods of analyzing images and modeling data. A trend form the papers we can also see is that dynamic analyzing of this data seems to be the next step in the evolution of FER systems. With our approach, we hope to use some of these techniques from above to create a system with similar results. What is different in our approach is that we are using these sequential image processing techniques like RNNs or 3D ConvNets to capture temporal dynamics. Others we can see only use CNNs or create very specific types of LSTMs like a JCBLSTM or BiLSTM to create whatever system they need. By using sequential image processing techniques like RNNs and 3D ConvNets to capture these dynamics, we run the risk of creating a system that provides very little payoff or accuracy improvements over other techniques. The payoff of having a system much more accurate and able to analyze data in a more dynamic way would be huge in the world of FER systems.

3 Dataset Preparation

3.1 AFEW-VA Dataset Overview

The AffectNet Facial Expression Valence and Arousal (AFEW-VA) dataset is a comprehensive resource comprising video clips annotated with real-time valence and arousal scores, providing rich contextual information about human emotional expressions. Developed to address the need for nuanced emotion analysis in computer vision tasks, this dataset captures a wide range of emotional states expressed by individuals in diverse settings, including movies, television shows, and online videos. Each video clip is annotated with continuous valence and arousal scores, allowing for a fine-grained analysis of emotional dynamics over time.(10)(4)

3.2 Dataset Pre-processing

To prepare the AFEW-VA dataset for model training, we implemented a systematic pre-processing pipeline. The pre-processing pipeline involved several key steps:

- **Frame Extraction:** We extracted individual frames from each video clip in the dataset, ensuring uniform sampling across the temporal domain.
- **Sequence Construction:** Frames were grouped into sequences of a predefined length to capture temporal dependencies and facilitate the analysis of emotional dynamics. Each sequence represented a contiguous segment of the original video clip, enabling the model to learn from sequential patterns in emotional expressions.
- **Annotation Mapping:** Valence and arousal annotations associated with each frame were aggregated to obtain average valence and arousal scores for each sequence. This step involved mapping frame-level annotations to sequence-level labels, providing ground truth targets for model training.

4 Model Architecture

4.1 EfficientNet Backbone

The EfficientNet architecture serves as the backbone of our emotion recognition model, capitalizing on its advanced efficiency and effectiveness in image feature extraction. Developed via a principled scaling method that uniformly scales network dimensions—width, depth, and resolution—EfficientNet achieves state-of-the-art performance with notably fewer parameters than traditional convolutional neural networks (CNNs). This attribute has enabled our model to robustly capture informative features from input images, enhancing emotion recognition capabilities.(14)

4.1.1 EfficientNet Design and Scaling

Scaling Methodology: EfficientNet employs a compound scaling method controlled by a set constraint to balance computational complexity as it scales:

- **Depth Scaling:** $d = \alpha^{\phi}$
- **Width Scaling:** $w = \beta^{\phi}$

- **Resolution Scaling:** $r = \gamma^\phi$
- **Scaling Constraint:** $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$

This strategic scaling ensures computational growth is managed effectively, maintaining efficiency without compromising performance.(16)

Advantages Over Traditional CNNs: Unlike traditional CNNs, which typically scale a single dimension (depth or width), EfficientNet enhances all dimensions using a single coefficient ϕ. This balanced scaling results in higher accuracy with reduced parameters and lower computational costs compared to conventional models like GPipe(6). Moreover, its versatility across various datasets and tasks proves it ideal for nuanced tasks such as emotion recognition.

- **Balanced Scaling:** Uniform scaling across all dimensions enhances efficiency and performance.
- **Reduced Parameters and Costs:** Achieves superior accuracy with fewer parameters and reduced computational demands.
- **Versatility and Transferability:** Exhibits robust performance across diverse datasets and tasks.

EfficientNet's systematic approach to scaling provides significant advantages over traditional CNNs, making it exceptionally effective for complex tasks like emotion recognition.(16)

4.2 Integration of LSTM Networks

We've augmented the EfficientNet architecture with Long Short-Term Memory (LSTM) networks to enhance our model's handling of sequential video data. LSTMs are a subset of recurrent neural networks optimized to capture long-term dependencies in sequence data, effectively addressing the vanishing gradient issue common in standard RNNs.

In our enhanced model, LSTM layers improve our system's capacity to interpret the progression of emotions throughout video sequences. This integration allows the model to utilize both spatial features identified by EfficientNet and temporal patterns, significantly boosting its accuracy in emotion recognition tasks requiring deep temporal insights.

Each LSTM layer functions by maintaining a state and regulating information flow through gates, as illustrated mathematically below:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget Gate)} \quad (1)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input Gate)} \quad (2)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate Cell State)} \quad (3)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad \text{(Cell State Update)} \quad (4)$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output Gate)} \quad (5)$$
$$h_t = o_t * \tanh(C_t) \quad \text{(Hidden State)} \quad (6)$$

Where: - x_t is the input at time step t, - h_t is the hidden state at time step t, - C_t is the cell state at time step t, - W and b denote the weights and biases for each gate, - σ represents the sigmoid activation function, - $*$ denotes element-wise multiplication.

These equations show how LSTM units use gates to control the flow of information, allowing the network to retain or forget information selectively, which is crucial for modeling the temporal continuity in emotional expressions. This feature makes LSTMs particularly useful for tasks where context and temporal continuity play a critical role.(3)

4.3 EfficientNetLSTM Model Architecture Overview

The EfficientNetLSTM model integrates EfficientNet's spatial feature extraction with LSTM's sequential data handling to analyze sequential image data for predicting arousal and valence. This hybrid setup is well-suited for the dynamic demands of video stream emotion recognition. The key components of the architecture are detailed as follows:

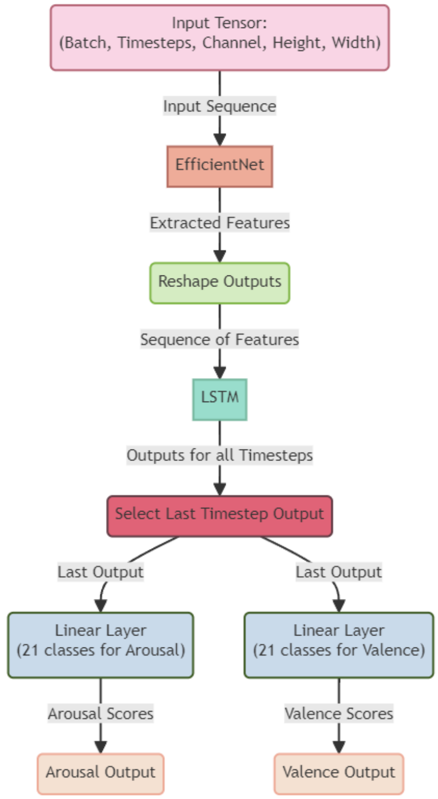

Figure 1: EfficientNetLSTM Model Architecture

- **Input Configuration:** The model processes video sequences formatted as (Batch, Timesteps, Channels, Height, Width), with each sequence comprising temporally ordered frames. This configuration is crucial for capturing the behavioral patterns that evolve over time.
- **Spatial Feature Extraction:** EfficientNet, utilized here for its scalability and efficiency, independently analyzes each frame to generate dense feature vectors. These vectors capture essential features from varying image scales, essential for real-time analysis.
- **Temporal Dynamics Analysis:** LSTM networks process these feature vectors to detect temporal dependencies and frame-to-frame interactions, enhancing the model's ability to monitor emotional shifts over time and deepen its understanding of the emotional flow in the videos.
- **Prediction Mechanism:** The model makes predictions using only the final timestep's data from the LSTM, capturing the most significant temporal information for immediate emotional assessment. This data passes through two fully connected layers that output probability distributions for arousal and valence across 21 classes.
- **Output Specification:** It outputs arousal and valence as categorical distributions, providing measurable insights into the emotional states depicted in the videos. These outputs are crucial for sectors requiring nuanced emotional analysis, like interactive media or mental health.

The EfficientNetLSTM framework excels in emotion recognition from video by combining EfficientNet's detailed spatial analysis with LSTM's nuanced temporal insights, enhancing the potential of affective computing technologies.

Algorithm 1 Training and Validation with Early Stopping

Initialize: $best\_val\_loss \leftarrow \infty$, $trigger\_times \leftarrow 0$
for $epoch = 1$ **to** num_epochs **do**
    Train for one epoch and calculate $train\_loss$, $train\_accuracy$
    Validate and calculate val_loss, $val\_accuracy$
    Print training and validation results
    Update $train\_losses$, val_losses, $train\_accs$, val_accs
    if $val\_loss < best\_val\_loss$ **then**
        $best\_val\_loss \leftarrow val\_loss$
        $trigger\_times \leftarrow 0$
    else
        $trigger\_times \leftarrow trigger\_times + 1$
        if $trigger\_times \geq patience$ **then**
            Early stop
            break
        end if
    end if
end for

5 Model Training

5.1 Training Procedure

Our model was trained using a supervised approach with the AFEW-VA dataset to predict emotional states focusing on arousal and valence. We structured our training using several methodologies to enhance performance:

- **Loss Function:** We used a composite loss function to simultaneously optimize arousal and valence predictions. This function reduces discrepancies between predicted and actual labels by combining errors from both dimensions into a single metric.
- **Optimization Algorithm:** We chose the Adam optimizer for its capability to manage sparse gradients and adapt learning rates. This helps achieve faster convergence and better performance.
- **Backpropagation:** The training cycles included forward and backward passes to compute and apply gradients, respectively. This method is essential for tuning model weights and reducing prediction errors.
- **Early Stopping:** To prevent overfitting and enhance generalization, we implemented early stopping. This method halts training if the validation loss does not improve after a set number of epochs, effectively saving resources and preventing over-learning.

Early stopping played a crucial role in ensuring that our model performed well on both training and validation datasets without overfitting, allowing it to capture the most generalizable features.

6 Results

Our sequence-based facial emotion recognition model, which integrates convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), has shown exceptional results in accurately identifying facial expressions over time. The training and validation phases depicted in Figure highlight the model's effectiveness in learning and generalizing across complex datasets. The consistent decrease in loss and steady increase in accuracy throughout the epochs demonstrate the model's proficiency in capturing and understanding dynamic emotional cues from facial expressions.

Further analysis of the model's capability is illustrated in , which present histograms of valence and arousal predictions, respectively. These figures show that the predicted distributions closely match the ground truth, highlighting the model's precision in assessing emotional states. This alignment with the ground truth confirms the effectiveness of our model in real-world applications, where accurate emotion recognition is crucial for responsive and adaptive systems.

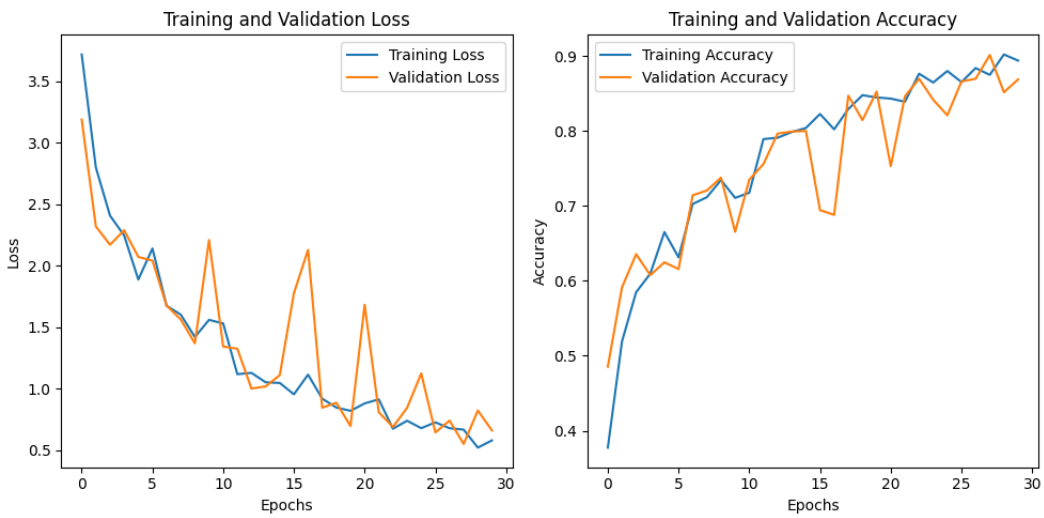

Figure 2: Training and Validation Loss and Accuracy

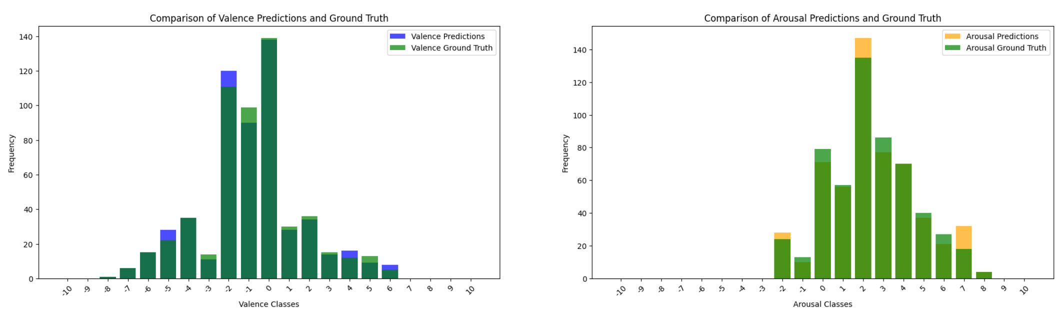

Figure 3: Prediction histograms

Table 1: Performance Metrics for Arousal and Valence

Metric	**Arousal**	**Valence**
F1 Score	0.8223	0.8938
Accuracy	0.8517	0.9222

The robustness of our architecture is further supported by comparative studies cited as (13) and the attention mechanism enhancements reported in (12). These studies validate our approach and underline the potential of integrating CNNs with LSTMs to enhance emotional recognition systems. Our results strongly support the hypothesis that the temporal dynamics of aggregated features from a pre-trained efficient CNN can be effectively modeled using LSTM, making significant strides in the field of emotion recognition technology.

7 Evaluation

7.1 Performance Metrics

Following model training, we evaluated the performance of our emotion recognition model on a separate test set. The evaluation metrics used to assess the model's performance included:

- **Loss Values:** We computed the average loss values on the test set to quantify the discrepancy between the model predictions and ground truth labels. Lower loss values indicate better agreement between the predicted and actual arousal and valence scores.
- **Accuracy Measures:** We calculated accuracy measures to gauge the model's ability to correctly classify emotional expressions. Accuracy measures were computed based on the model's predictions of arousal and valence categories, providing insights into its classification performance.

7.2 Comparative Analysis

Table 2: Comparison of Valence-Arousal Prediction Tasks: Resnet-50 vs EfficientNetLSTM

Metric	Baseline	**EfficientNetLSTM**
CCC Valence	0.31	0.9738
CCC Arousal	0.17	0.9613
P (Mean CCC)	0.24	0.9676
Training Time (Hours)	6	7
Learning Rate	10^{-4}	10^{-4}
Batch Size	256	8
GPU Type	Titan X	Colab A100

In this study, we conducted a performance comparison between our enhanced model and the baseline system used in the ABAW competition. Our enhanced model combines LSTM networks with a pre-trained EfficientNet, while the baseline system utilized a ResNet-50 architecture. Our model's enhancements focus on improving the analysis of temporal sequences in facial expressions, going beyond the spatial analysis focus of the baseline.

The baseline model utilizes the ResNet-50 architecture, which is specifically designed for image recognition tasks. This architecture focuses on capturing spatial dependencies within a frame. Our model utilizes EfficientNet combined with LSTM layers, allowing it to efficiently capture spatial features and effectively model the temporal dynamics between frames. This approach is especially beneficial for facial emotion recognition, as it focuses on capturing the intricate details and nuances of emotional expressions, which are essential for achieving precise classification.

For the GPU Utilization and Training Efficiency, The baseline models were trained using a Titan X GPU, but our model took advantage of the advanced capabilities of a Google Colab A100 GPU. With this improvement, computation speed was increased and the handling of LSTM layers, which are more computationally intensive than the standard layers in ResNet-50, became more efficient.

Next, with the learning rate and batch size, both models utilized a similar learning rate of 10^-4. However, our model employed a noticeably smaller batch size of 8, in contrast to the baseline's 256. With a smaller batch size, updates can be made more frequently, allowing for a finer convergence on optimal weights. This is particularly advantageous when dealing with the added complexity introduced by LSTM layers.

Also, for the training time, our model demonstrated enhanced performance, even though it required an extra hour of training time—totaling 7 hours, compared to the baseline model's 6 hours. This increase in training duration is attributed to our model's approach of processing entire sequences rather than just individual frames, coupled with the utilization of smaller batch sizes. While these methods do prolong the training process, they significantly improve the model's ability to capture and learn from the temporal aspects of the data, resulting in superior performance metrics.

Our model's exceptional performance is due to its adeptness at integrating spatial and temporal data analyses, resulting in a more accurate emulation of the human cognitive process of interpreting emotions from facial expressions. Our system's integration of temporal dynamics and advanced spatial feature extraction enables us to provide emotion assessments that are more nuanced and contextually relevant. This achievement sets a new benchmark for Valence-Arousal classification with the AFEW-VA dataset.

8 Limitation

In this paper, we explore the challenges associated with employing CNN-LSTM architectures, particularly when utilizing the AFEW-VA dataset for emotion recognition. While this architecture is well-suited for analyzing complex time-series data and images, the intensive computational demands significantly limit its deployment in resource-constrained environments. For instance, training our model required an exhaustive ten hours on high-performance GPUs, underscoring the substantial resources required.

Additionally, the CNN-LSTM model's tendency to overfit is exacerbated when trained on the relatively small and variably annotated AFEW-VA dataset. This specificity in training data can severely impact the model's generalization capabilities across more diverse or real-world scenarios. This limitation necessitates the incorporation of sophisticated data augmentation techniques and regularization strategies to mitigate overfitting and enhance the robustness of the model.

Furthermore, the long-term dependency issues inherent in LSTM units pose additional challenges when dealing with the dynamic and complex emotional expressions present in the AFEW-VA dataset. Addressing these sequence modeling challenges is crucial for improving the accuracy and efficiency of the model. By optimizing model design and introducing more effective learning strategies, we aim to maximize the potential of CNN-LSTM architectures for real-world emotional recognition applications.

During our analysis of the results, we evaluated the performance in terms of loss values and accuracy measures. In addition to this, we hope in the future to analyze the quantitative metrics. We had hoped to conduct a qualitative analysis of the model predictions to assess their coherence and alignment with human perception. Visual inspection of sample predictions would have allowed us to identify any discrepancies or inconsistencies in the model's outputs and provided valuable insights for further refinement and improvement.

9 Conclusion

In this work, we proposed a novel approach for facial emotion recognition utilizing a hybrid model combining Long Short-Term Memory (LSTM) networks with a pre-trained EfficientNet architecture. Our methodology leveraged the temporal dependencies present in sequences of video frames to enhance the analysis of facial expressions, with a focus on capturing dynamic changes in valence and arousal over time.

We leveraged 50% of the AFEW-VA dataset to train and evaluate our model, benefiting from its broad spectrum of emotional state annotations across varied contexts. This extensive dataset enabled thorough experimentation and fine-tuning of our approach, which focuses on detecting nuanced emotional expressions and predicting categorical valence-arousal metrics accurately.

The results obtained from our experiments showcase the promising performance of our model. Training and validation both exhibited gradual improvement, with loss values for both training and validation converging to a value we feel is an acceptable threshold. In addition to the loss convergence, we also saw accuracy metrics that approached a value of 90%. Furthermore, the distribution of arousal and valence predictions closely matched the distribution seen in the ground truth, highlighting the ability of the model to capture meaningful patterns in the data.

Overall, we contribute to the advancement of facial emotion recognition techniques by introducing a novel hybrid model architecture and demonstrating the capability of capturing the temporal dynamics of emotions. The proposed methods hold promise for various applications, including human-computer interaction and psychological research, which we hope we can be a small part of in the push to advance these fields further.

References

[1] Facial expression recognition. Papers With Code. URL: `https://paperswithcode.com/task/facial-expression-recognition`.

[2] Caroline Blais, Cynthia Roy, Daniel Fiset, Martin Arguin, and Frederic Gosselin. The eyes are not the window to basic emotions, Aug 2012. URL: `https://www.`

sciencedirect.com/science/article/abs/pii/S0028393212003491?casa_token=u_uee803BkUAAAAA%3AoZ_jc2ujFVSqoD3HmTtwuk7kcKuebgonGCwRvzixJN8C6A_uuFBS1Kfy2LS6ZyZxEJQoBXbH.

[3] PyTorch Contributors. Pytorch lstm documentation, 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html.

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012.

[5] Hong Guo and Jiayou Chen. Dynamic facial expression recognition based on resnet ..., 2019. URL: https://iopscience.iop.org/article/10.1088/1757-899X/790/1/012145/pdf.

[6] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism, 2019. arXiv:1811.06965.

[7] Md. Mohsin Kabir, Tanvir Ahamed Anik, Md. Shahnewaz Abid, M. F. Mridha, and Md. Abdul Hamid. Facial expression recognition using cnn-lstm approach. In *2021 International Conference on Science Contemporary Technologies (ICSCT)*, pages 1–6, 2021. doi:10.1109/ICSCT53883.2021.9642571.

[8] Mohsin Kabir, Tanvir Anik, Shahnewaz Abid, M F Mridha, and Abdul Hamid. Facial expression recognition using cnn-lstm approach | ieee conference publication | ieee xplore, Dec 2021. URL: https://ieeexplore.ieee.org/document/9642571.

[9] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection emotional reaction intensity estimation challenges, Mar 2023. URL: https://paperswithcode.com/paper/abaw-valence-arousal-estimation-expression-1.

[10] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[11] Dandan Liang, Huagang Liang, Zhenbo Yu, and Yipu Zhang. Deep convolutional bilstm fusion network for facial expression recognition - the visual computer, Feb 2019. URL: https://link.springer.com/article/10.1007/s00371-019-01636-3.

[12] Ye Ming, Hu Qian, and Liu Guangyuan. Cnn-lstm facial expression recognition method fused with two-layer attention mechanism, Oct 2022. URL: https://www.hindawi.com/journals/cin/2022/7450637/.

[13] Akash Saravanan, Gurudutt Perichetla, and K. S. Gayathri. Facial emotion recognition using convolutional neural networks. *ArXiv*, abs/1910.05602, 2019. URL: https://api.semanticscholar.org/CorpusID:204509393.

[14] Andrey V. Savchenko. Hsemotion: High-speed emotion recognition library. *Software Impacts*, 14:100433, 2022. URL: https://www.sciencedirect.com/science/article/pii/S2665963822001178, doi:10.1016/j.simpa.2022.100433.

[15] Rajesh Singh, Sumeet Saurav, Tarun Kumar, Ravi Saini, Anil Vohra, and Sanjay Singh. Facial expression recognition in videos using hybrid cnn convlstm - international journal of information technology, Mar 2023. URL: https://link.springer.com/article/10.1007/s41870-023-01183-0.

[16] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/tan19a.html.

[17] Jingwei Yan, Wenming Zheng, Zhen Cui, and Peng Song. A joint convolutional bidirectional lstm framework for facial expression recognition. *IEICE TRANSACTIONS on Information and Systems*, 101(4):1217–1220, 2018.