# EmotionNet: Enhancing Facial Emotion Recognition with Spatial Attention

Yi Lin[1]

The Australian National University, Canberra ACT 2601, AU

**Abstract.** Traditional facial emotion recognition technology is severely limited in its ability to generalize in complex real-world environments. Methods that rely on weight analysis with a limited quantity of principal component information frequently fail to underscore the model's emphasis on specific features, according to previous research. This paper introduces EmotionNet, a novel model that combines the pre-trained ResNet-18 architecture with a spatial attention mechanism, with the goal of enhancing the accuracy and efficacy of facial emotion recognition. We conduct extensive experiments using the SFEW dataset to validate our model and demonstrate its superiority over the ResNet-18 model. Our findings validate the efficacy of convolutional neural networks in emotion recognition tasks and emphasize the significance of weight analysis in comprehending, enhancing, and ensuring the accuracy of models.

**Keywords:** Facial emotion recognition · Convolutional neural network · Weight analysis · Attention network

## 1 Introduction

Facial expression analysis, a prominent area of study within the domains of computer vision and artificial intelligence, has significant obstacles when applied to intricate real-world settings [1]. The aforementioned issues mostly stem from variations in ambient illumination conditions, alterations in face posture configurations, and slight discrepancies in facial movements. The primary reliance in traditional face expression analysis research is on datasets that are acquired inside controlled laboratory contexts [2]. While this approach contributes to the advancement of research to a certain degree, it also imposes constraints on the use of the model in practical situations. The statement made is a generalization. The SFEW dataset [3] was introduced as a means to explore face emotion analysis in real-world scenarios. Its efficiency was evaluated using two commonly used feature extraction techniques, namely LPQ [4] and PHOG [5].

Extracting relevant information from the SFEW data set is a complex challenge, despite its near resemblance to real-life circumstances. In the preceding research, we conducted a thorough examination of the dataset in order to diminish the presence of extraneous data and preserve the most meaningful components. Specifically, we only used the first five main components of LPQ and PHOG [3]. Previous studies have shown notable accomplishments in facial feature extraction via the use of LPQ and PHOG techniques. However, it is important to acknowledge that these methods also possess some drawbacks, including elevated computing complexity, challenges in parameter adjustment, and a lack of flexibility in feature extraction [6]. The aforementioned restrictions provide challenges for both approaches in accurately capturing the intricate nuances of face emition, particularly inside intricate real-world environments. To address these constraints and enhance suitability for intricate and dynamic real-world contexts, we choose to use neural networks for the purpose of emotion categorization. Neural networks have emerged as a promising approach for facial emotion identification due to their remarkable ability to automatically extract relevant features and achieve high performance across a range of tasks [7]. Within this particular framework, neural networks has the capability to autonomously extract significant and distinguishing characteristics from data, hence facilitating precise categorization of emotions.

Nevertheless, it is insufficient to merely depend on the outcomes generated by a neural network model in order to comprehensively comprehend the decision-making mechanism of the model, particularly when confronted with intricate real-world data. To provide a comprehensive investigation into the underlying workings of the model, we used the weight analysis technique [8]. Through the process of quantifying the individual contributions of each feature to the output of the model, a deeper comprehension of the decision-making process used by the model can be attained, along with a clearer identification of the characteristics that have the highest significance in the model. The matter has significance. This analysis not only facilitates a comprehensive comprehension of the decision-making process used by the model, but also aids in the identification of possible vulnerabilities within the model and offers guidance for enhancing its resilience. Previous research has mostly used datasets with a restricted number of features, hence leaving scope for enhancing the accuracy of the retrieved features. In particular, when doing weight analysis on the main component data, the observed weight variations are rather minor, posing challenges in accurately discerning the specific regions of heightened model attention.

To gain a more comprehensive comprehension of the model's behavior in handling intricate image data and to validate the efficacy of weight analysis in enhancing model performance, we opted to employ a Convolutional Neural Network (CNN) [9] for the processing of the SFEW (Static Facial Expressions in the Facial emotion recognition task on Wild) image dataset.

The baseline model used for our research was ResNet-18 [10], which had been pre-trained on the ImageNet dataset. To illustrate the focus of the model on samples and emphasize the dispersion of model weights, we employed Grad-CAM [11]. To improve model performance and increase focus on significant feature regions, we used the Spatial Attention module [12] to amplify the influence of model weight settings on outcomes. The inclusion of a spatial attention module in the model facilitates the allocation of more attention towards picture regions that are pertinent to the given task, thereby enhancing the precision of the outcomes.

In this set of experiments, we have conducted an investigation to validate the efficacy of convolutional neural networks in the context of face emotion detection tasks. Additionally, we have shown the significance of weight analysis in comprehending and enhancing models. The experimental findings demonstrate that doing a weight analysis on the model enables a more profound comprehension of the decision-making process used by the model. Additionally, this analysis facilitates the identification of crucial aspects that significantly impact the model's performance. Consequently, these insights provide essential direction for future endeavors aimed at optimizing the model.

In conclusion, this paper presents two significant contributions. Firstly, it introduces and validates a facial emotion regression method that utilizes convolutional neural network, demonstrating its effectiveness on real-world datasets. Secondly, the paper delves into the exploration of feature importance through weight analysis and introduces the attention mechanism to demonstrate that weight analysis can greatly enhance the performance and robustness of the model. The accuracy of our model on the test set was **52.17%**. The current state-of-the-art result on the complete SFEW dataset is **56.4%** (RAN Model [13]).

## 2 Related Work

In this section, we will first describe the convolutional neural network with ResNet-18 and then introduce the weight analysis with Grad-CAM and Spatial Attention.

### 2.1 Convolutional Neural Network for Facial Emotion Recognition

Convolutional Neural Network (CNN) is a deep learning model particularly suitable for processing image data [14]. The convolutional layers in the convolutional network automatically learn the features of the spatial hierarchy by sliding across the image to scan the input data [15]. Convolutional layers reduce model complexity through local connections and weight sharing [16] while maintaining awareness of spatial structure.

Facial emotion recognition is a very active research direction in the field of computer vision and artificial intelligence [17]. Its goal is to identify and understand the emotional state expressed by human faces. CNN automatically extracts low-level and high-level features [18] of faces through convolutional and pooling layers and performs effective classification.

**ResNet-18** ResNet-18 [10] is a deep residual network (Residual Network) proposed by Microsoft Research in 2015. The main feature of this network is the introduction of the concept of residual learning (Residual Learning) [10] to solve the problems of gradient disappearance and gradient explosion during the training process of deep neural network, thereby making the network deeper and more accurate. In the face emotion recognition task, ResNet-18 can effectively extract the features of facial expressions and improve the classification accuracy through deep residual learning [19].

On the SFEW data set, ResNet-18 is effective as a baseline model [13].

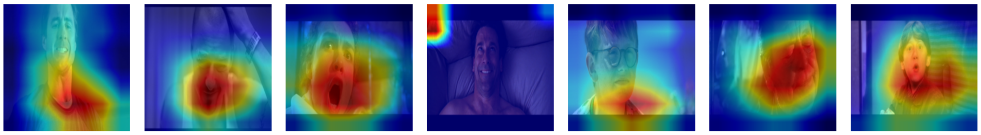
### 2.2 Weight Analysis

In previous work, we used weight calculation formulas as **Equation (1)** in our neural network EmitionNN to view model weights. This is very convenient for datasets with fewer features. However, the image data we have to process now has a large number of features, and the weight distribution cannot be intuitively reflected through the weight data.

$$P_{ij} = \frac{|w_{ij}|}{\sum_j |w_{ij}|}, P_{jk} = \frac{|w_{jk}|}{\sum_k |w_{jk}|}, Q_i = \sum_j P_{ij} \left( \sum_k P_{jk} P_k \right) \tag{1}$$

Here, $P_{ij}$ represents the weight ratio of the connection between the $i^{th}$ input and the $j^{th}$ neuron. $P_{jk}$ is similar. $Q_i$ reflects the total impact of the $i^{th}$ input on the output, considering the contributions of all inter neurons and their connections.

**Grad-CAM** Grad-CAM (Gradient-weighted Class Activation Mapping) [11] is a method for visualizing the learning results of convolutional neural networks (CNN) in different image areas. It can record the weight changes in the model and combine the weighted feature maps to generate a heat map. Finally, the weight distribution of the model is intuitively reflected by overlaying the heat map on the original image.



(a) Angry (b) Disgust (c) Fear (d) Happy (e) Neutral (f) Sad (g) Surprise

Fig. 1: Grad-CAM Heat Maps with Different Emotions

**Spatial Attention** Attention Mechanism is a crucial model enhancement technology that enables the model to concentrate more on certain crucial components when processing input data [20], thereby enhancing the model's performance. For the weight analysis, the attention mechanism is crucial for explaining the model's decision-making process, distinguishing the areas on which the model concentrates, and enhancing model performance [21].

The Spatial Attention Module [12] is a mechanism used to enhance the performance of convolutional neural networks. It strengthens the model's focus on key features by assigning different weights to different areas in the image[22]. In complex practical application scenarios, there may be a lot of background noise in the image. The spatial attention mechanism can help the model suppress this irrelevant information and ensure that the weight analysis focuses on the features that are really important to the task [23].

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ M_s(F) &= \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{2}$$

where $M_s(F) \in R^{H\times W}$ is the spatial attention map with input image size $H \times W$, $\sigma$ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with the filter size of $7 \times 7$, and $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$ are the average and max pooling features across the channel.

## 3 Methodology

In this section, we will first describe the structure of our convolutional neural network EmotionNet and then show the impact of the area of the spatial attention module on the model.

### 3.1 Model Structure

The attention mechanism helps identify the areas that are most important to the model's decisions by assigning different weights to different parts of the input data. In order to further prove the role of weight analysis technology on model performance, we proposed EmotionNet, a convolutional neural network integrated with a spatial attention module.

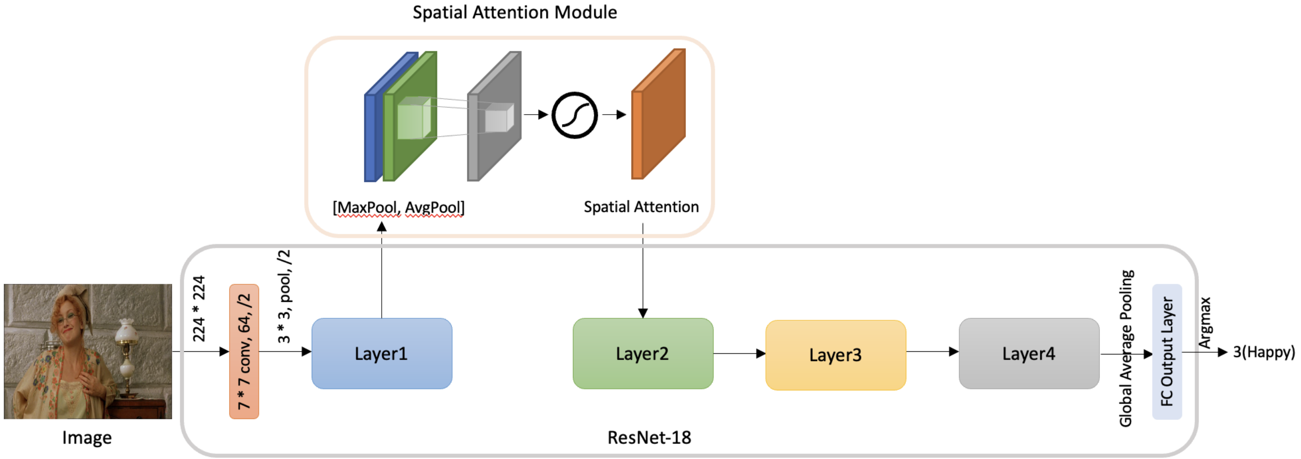


Fig. 2: The framework of our EmotionNet

EmotionNet aims to improve the accuracy of emotion recognition by improving the model's sensitivity to key facial regions through a spatial attention mechanism. The architecture of EmotionNet is shown in **Fig. 2**, which mainly includes three modules: feature extraction module, spatial attention module and emotion classification module. First, we use pre-trained ResNet-18 as the base model to extract features of facial images. These features are then passed to the spatial attention module, which assigns attention weights by learning the importance of each region in the image. We use convolutional layers and Sigmoid activation functions to implement this process, ensuring that the weight assignment of each region is adaptive and ranges between 0 and 1.

The output of the spatial attention module is a weighted feature map that emphasizes the most important regions for emotion recognition while suppressing the influence of irrelevant or interfering regions. This weighted feature map is then passed to the emotion classification module, which consists of fully connected layers and uses a softmax activation function to predict the emotions of the people in the image.

In this way, EmotionNet is not only able to leverage the powerful feature extraction capabilities of deep learning, but also adaptively focus on the most relevant facial regions through a spatial attention mechanism. In addition, EmotionNet modifies the weight distribution through weight analysis, thereby improving the accuracy and robustness of emotion recognition. We conduct a comprehensive evaluation of EmotionNet in the experimental section, and the results show that it outperforms traditional methods in various situations and is able to effectively handle subtle changes in facial expressions.

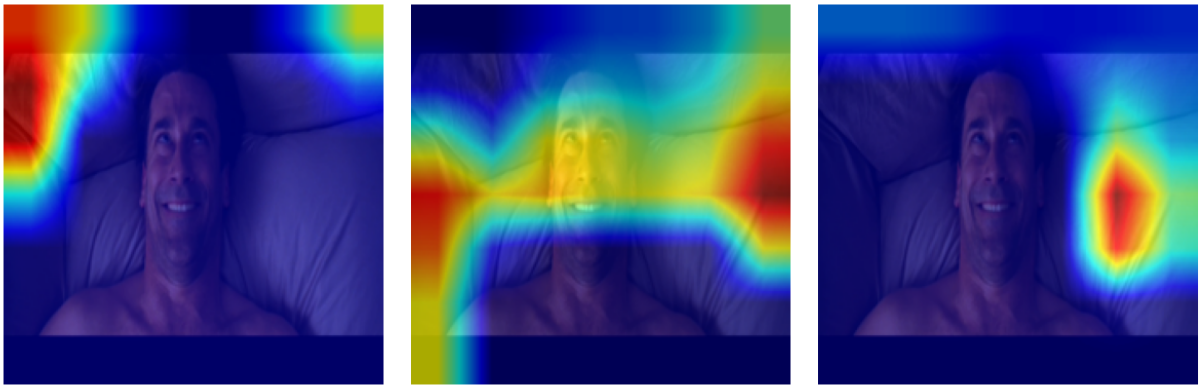
### 3.2 Area Selection of Spatial Attention

During the process of building the EmotionNet model, we explored the impact of placing the spatial attention (SA) module in different locations on the model performance. We chose pre-trained ResNet-18 as the base network and inserted SA modules after different layers to observe its impact on model performance.

**Input Part of the Network** When the SA module is placed in the input part of the network, it mainly focuses on low-level features such as edges and texture. Although the SA module can help the model better capture the subtle structure of the face at this time, it may ignore higher-level, emotion-related features.

**Middle Layer of the Network** Placing it in the middle layer of the network can help the model better focus on key areas of the face, such as the eyes, mouth, and eyebrows, which often contain rich emotional information. In this position, the SA module is able to catch low-level and high-level features to provide more comprehensive emotion recognition.

**Deep Layer of the Network** In deep layer, it mainly processes high-level semantic features. In this position, the SA module helps the model better understand the overall expression, but may sacrifice some detail accuracy.



(a) Input Part of the Network (b) Middle Layer of the Network (c) Deep layer of the Network

Fig. 3: Grad-CAM Heat Maps with Different Attention Area

Through multiple experimental comparisons, we chose to place the spatial attention (SA) module after layer 1 of ResNet-18 (The Structure is shown in **Fig. 2**). This can be thought of as placing the spatial attention module in the middle layer of the network. This location helps the model better capture and understand the emotional information in facial expressions, locating the main areas through weight analysis.

## 4 Experiments

In this part, we provide a comprehensive summary of the dataset used in our research. Subsequently, the dataset was partitioned to serve as a data source for the purpose of assessing the efficacy of the EmotionNet model. Based on this premise, a thorough comparison was carried out between EmotionNet and several ResNet models. By conducting a thorough evaluation and comparison of EmotionNet, we have substantiated its effectiveness and emphasized its potential as a dependable tool in the domain of facial emotion analysis. This assessment include the investigation of model performance in relation to evaluation metrics and weight distributions.

### 4.1 Data Preprocessing

The dataset used in this paper is part of the data extracted from the SFEW dataset. The dataset contains seven different emotions and the samples are evenly distributed. In order to balance the learning degree of the model for all categories, we randomly selected 80% from each category as training data, and the remaining 20% was evenly divided into the validation set and the test set. Furthermore, to normalize the input data, we resize all images to size 224 and encode all categories with index values as **Table 1**.

Table 1: Emotion Category Index Comparison Table

| **Emotion** | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| **Index** | 0 | 1 | 2 | 3 | 4 | 5 | 6 |

### 4.2 Model Settings

In order to fully learn the model, we train all models with high epoch and high batch size in a high computing environment. It was finally determined that under the condition of 50 epochs and batch size of 32, the model effect was excellent and there was no obvious over-fitting phenomenon. These are the best options after experiments (**Table 2**):

Table 2: The Best Options in Experiments

| | |
|---|---|
| Epoch | 50 |
| Batch size | 32 |
| Loss function | Cross Entropy Loss |
| Optimizer | Adam (weight decay=1e-5) |
| Learning rate | 1e-3 |
| Learning rate Scheduler | StepLR (step size=10, gamma=0.9) |

**Loss Function** This experiment chooses **Cross Entropy Loss** [24] as the loss function of the model. It is a commonly used loss function in classification problems and is directly related to classification performance by maximizing the predicted probability of the correct class [25]. This loss function is suitable for classification problems with two or more categories, and the task is to predict the probability of each category. Experiments show that Cross Entropy Loss works well in facial emotion recognition tasks. The formula is as follows:

$$Loss_{CrossEntropy}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{3}$$

$N$ is the sample size, $C$ represent all categories, $y_{ij}$ represents the true label and $\hat{y}_{ij}$ is predicted probability respectively.

**Optimizer** This experiment uses the **Adam optimizer** [26]. It is an adaptive learning rate optimization algorithm that combines the advantages of Momentum and RMSprop [27]. It can dynamically adjust the learning rate of each parameter based on the first-order moment estimate and second-order moment estimate of the parameter, making the model converge faster and less sensitive to the initial learning rate. And the optimizer provides specified parameters for L2 regularization to prevent model overfitting [28].

**Learning Rate Scheduler** This experiment uses **StepLR** as the learning rate scheduling strategy. It will multiply the learning rate by gamma after every step size epochs. In this way, a larger learning rate can be used to decrease quickly in the early stages of training, and then gradually reduce the learning rate to stabilize the model in the later stages of training. This helps improve model performance and speeds up convergence.

**Evaluation Index** In addition to comparing the accuracy of the validation set and the test set, this experiment used three evaluation indicators in the SFEW paper [3]: **Precision, Recall and Specificity**.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{4}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{5}$$

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}} \tag{6}$$

And in these three formulas:

True Positives (TP): The number of samples that are actually positive and predicted by the model to be positive.

False Positives (FP): The number of samples that are actually negative but predicted by the model to be positive.

True Negatives (TN): The number of samples that are actually negative and predicted by the model to be negative.

False Negatives (FN): The number of samples that are actually positive but predicted by the model to be negative.

### 4.3 Results and Comparison

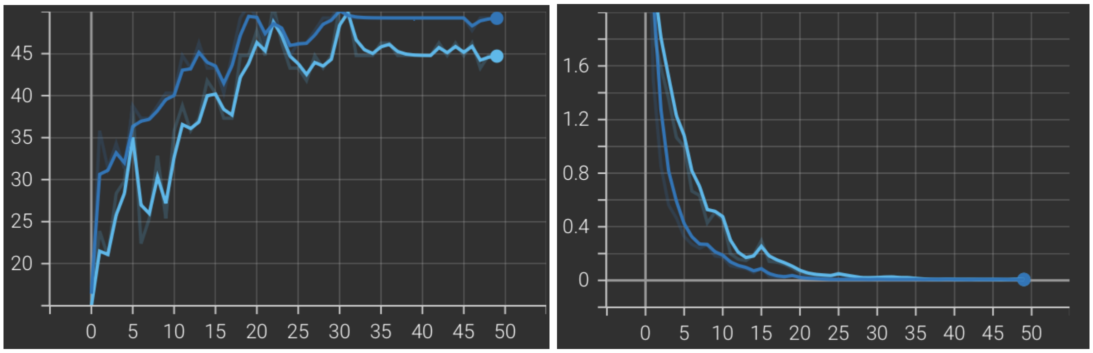
**Analysis of the Network Components** We assess the effectiveness of our proposed EmotionNet model, which integrates a pretrained ResNet-18 architecture with self-attention (SA) modules. Then we conduct a comparative

Table 3: Model Accuracy on Test Set

| **Model** | ResNet-18 | ResNet-50 | ResNet-50+sa | EmotionNet (ResNet-18+sa) |
|---|---|---|---|---|
| **Accuracy** | 42.03% | 40.58% | 36.23% | **52.17%** |

analysis of the test accuracy in several models, like pretrained ResNet-18 and ResNet-50, as well as ResNet-50 integrated with a spatial attention (SA) module.

The EmotionNet model we propose integrates pretrained ResNet-18 and spatial attention modules, with the objective of enhancing performance via the identification and emphasis of emotion-related areas inside pictures. **Table 3** indicate that EmotionNet outperforms other models in terms of accuracy on the test set (52.17%).



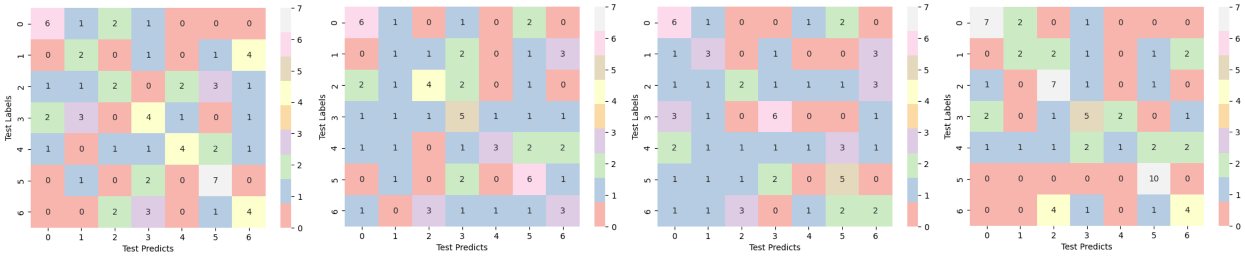
(a) Validation Accuracy

(b) Validation Loss

Fig. 4: Validation Set Results (Blue: ResNet-18, Cycan: ResNet-50)

This is noteworthy considering that ResNet-50 possesses a deeper network architecture and increased feature extraction capabilities. **Fig.4** illustrates the impact of the two models on the validation set. The ResNet-18 model has superior convergence speed and accuracy compared to the ResNet-50 model, which demonstrates slower convergence speed and more variability in the accuracy curve. Furthermore, the incorporation of the spatial attention module did not provide any significant improvement in the performance of ResNet-50+SA. This suggests that the settings presently configured are not appropriate for the ResNet-50 model. The relationship between model complexity and performance is not necessarily positively correlated.

To visualize how well the model recognizes different categories, we use a confusion matrix to show the relationship between model predictions and actual categories.



(a) ResNet-18

(b) ResNet-50

(c) ResNet-50+sa

(d) EmotionNet

Fig. 5: Confusion Matrix on Test Set

For a confusion matrix, the numbers on the diagonal represent the number of correct classifications for each category. The larger the value on the diagonal, the better the model's performance on this category. Numbers off the diagonal represent instances of model confusion, i.e. the number of instances that were actually in one class but predicted in another. The smaller these values are, the better because they represent the number of classification errors.

Based on the above conclusion, **Fig.5** shows that the prediction effect of our model is ideal, and the prediction ability for different categories is different. Among them, our model has weak prediction ability for the category with index 3. Additionally, we can use confusion matrix to get Precision, Recall, and Specificity. EmotionNet demonstrates its superiority in performance metrics in **Table 4**.

Table 4: Precision, Recall and Specifity results on Test Set

| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Precision | 0.5 | 0.14 | 0.38 | 0.30 | 0.50 | 0.58 | 0.27 |
| Recall | 0.60 | 0.12 | 0.50 | 0.27 | 0.20 | 0.70 | 0.30 |
| Specificity | 0.90 | 0.90 | 0.86 | 0.88 | **0.97** | **0.92** | 0.86 |

(a) ResNet-18

| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Precision | 0.55 | 0.17 | 0.44 | 0.36 | **0.60** | 0.43 | 0.30 |
| Recall | 0.60 | 0.12 | 0.40 | 0.45 | **0.30** | 0.60 | 0.30 |
| Specificity | 0.92 | 0.92 | **0.92** | 0.84 | **0.97** | 0.86 | 0.88 |

(b) ResNet-50

| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Precision | 0.40 | 0.33 | 0.29 | **0.55** | 0.25 | 0.38 | 0.20 |
| Recall | 0.60 | **0.38** | 0.20 | **0.55** | 0.10 | 0.50 | 0.20 |
| Specificity | 0.85 | 0.90 | **0.92** | **0.91** | 0.95 | 0.86 | 0.86 |

(c) ResNet-50+sa

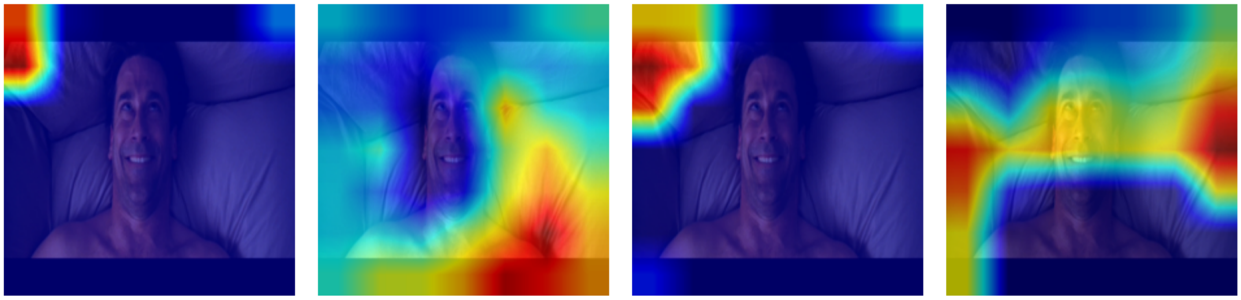
| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Precision | **0.64** | **0.40** | **0.47** | 0.45 | 0.33 | **0.67** | **0.44** |
| Recall | **0.70** | 0.25 | **0.70** | 0.45 | 0.10 | **1.0** | **0.40** |
| Specificity | **0.93** | **0.95** | 0.86 | 0.90 | **0.97** | **0.92** | **0.92** |

(d) EmotionNet (ResNet-18+sa)

Based on the conducted tests, it was observed that for face emotion detection tasks, the use of relatively uncomplicated models, such as ResNet-18, may be more suitable, particularly when the dataset size is limited. Furthermore, the incorporation of the spatial attention module has the potential to enhance the model's focus on areas associated with emotions, leading to improved performance. However, it is crucial to exercise caution and make precise adjustments based on the individual tasks and datasets at hand. The EmotionNet model shows significant promise in the domain of face emotion identification tasks via its integration of both components.

**Weight Analysis** The weight visualization approach is used to investigate the significance of individual features in the job of emotion detection. This study aims to assess the impact of the attention mechanism on the overall performance of the model. To do this, we use the pre-trained ResNet-18 architecture, which is augmented with the spatial attention (SA) module.

We first used the Grad-CAM method to visualize the weight distribution results. Generate heat maps to visually demonstrate the areas that the model focuses on when making predictions. The areas highlighted by Grad-CAM

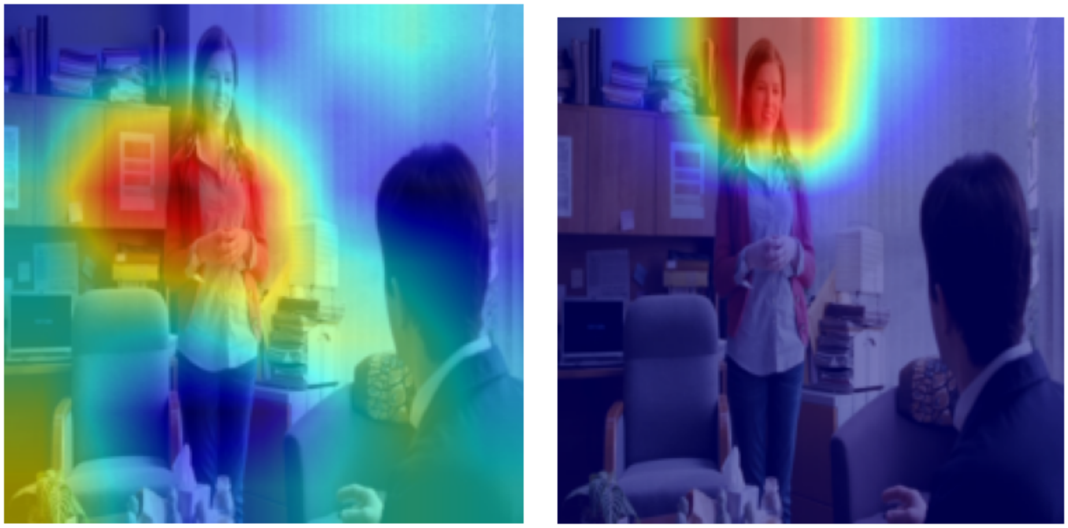


(a) ResNet-18 (b) ResNet-50 (c) ResNet-50+sa (d) EmotionNet

Fig. 6: Grad-CAM Heat Maps on a Happy Sample

directly correspond to the highest weighted areas within the model, indicating that these areas are critical to the model's decision-making process. By comparing the heat maps in **Fig.6** which generated by different models, we can see that they focus on different areas, and our EmotionNet's results are closer to the face area. The errors of the two models, ResNet-18 and ResNet-50+sa, are particularly obvious. These two models' focus areas appear at the edges of the image.

In addition to single-object images, our model also achieves desirable results on multi-object samples.



(a) ResNet-18

(b) ResNet-18+sa

Fig. 7: Grad-CAM Heat Maps on Multi-object Sample

The findings shown in **Fig.7** demonstrate that our model exhibits exclusive attention towards the accurate facial region and places more emphasis on the distribution of weights within this specific area. The ResNet-18 model pays attention to the whole area of the person. This finding reflects that after introducing the spatial attention mechanism, our model can efficiently retrieve important regions of samples and enhance the weight distribution of the regions.

## 5 Conclusion and Future Work

Through extensive experiments and evaluations, our findings highlight that weight analysis provides valuable insights into the model's decision-making process, enhancing the model's understanding of emotion expression regions. We analyze the impact of different network components and attention modules, demonstrating the importance of each component in the overall performance of the emotion recognition task. In particular, the attention mechanism plays a crucial role in guiding the model to focus on relevant areas of the face, resulting in more accurate and reliable emotion predictions.

Our proposed EmotionNet model is rich in attention mechanisms and has demonstrated its ability to focus on key facial regions, ensuring that the most informative features are exploited for emotion recognition. This approach not only improves the accuracy of the model, but also improves its robustness, making it more resilient to changes and complexities in facial emotion regression.

Our future research is to extend the existing work of EmotionNet, which successfully integrated Convolutional Neural Network (CNN) and weight analysis, in order to advance the field of face emotion identification. Our research aims to investigate the integration of more sophisticated attention processes in order to improve the model's capacity to prioritize the most crucial variables for the categorization of emotions. We also considered using more sophisticated data augmentation methods into the dataset.

Efforts will also be focused on refining the network architecture and fine-tuning the hyperparameters in order to get enhanced performance and efficiency. Within the domain of weight analysis, our objective is to enhance the sophistication of methodologies used, aiming to provide more profound insights into the decision-making process of the model. This pursuit is crucial as it aids in the identification of possible biases, so maintaining the fairness and transparency of our system.

## References

1. Canedo, D., Neves, A. J. (2019). Facial expression recognition using computer vision: A systematic review. Applied Sciences, 9(21), 4678.
2. Adjabi, I., Ouahabi, A., Benzaoui, A., Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. Electronics, 9(8), 1188.
3. Dhall, A., Goecke, R., Lucey, S., Gedeon, T. (2011, November). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In 2011 IEEE international conference on computer vision workshops (ICCV workshops) (pp. 2106-2112). IEEE.
4. M. U. Khan, M. A. Abbasi, Z. Saeed, M. Asif, A. Raza and U. Urooj, ”Deep learning based Intelligent Emotion Recognition and Classification System,” 2021 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2021, pp. 25-30, doi: 10.1109/FIT53504.2021.00015.
5. M. T. B. Iqbal, B. Ryu, A. R. Rivera, F. Makhmudkhujaev, O. Chae and S. -H. Bae, ”Facial Expression Recognition with Active Local Shape Pattern and Learned-Size Block Representations,” in IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1322-1336, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.2995432.
6. Li, S., Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3), 1195-1215.
7. Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C. H., Xiang, Y., He, J. (2019). A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors, 19(8), 1863.
8. Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. International journal of neural systems, 8(02), 209-218.
9. Albawi, S., Mohammed, T. A., Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET) (pp. 1-6). Ieee.
10. 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).
11. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
12. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).
13. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29, 4057-4069.
14. Traore, B. B., Kamsu-Foguem, B., Tangara, F. (2018). Deep convolution neural network for image recognition. Ecological informatics, 48, 257-268.
15. Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., Chen, Y. (2015). Convolutional recurrent neural networks: Learning spatial dependencies for image representation. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 18-26).
16. Khan, A., Sohail, A., Zahoora, U., Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 53, 5455-5516.
17. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. sensors, 18(2), 401.
18. Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9), 1036.
19. Zhou, Y., Ren, F., Nishide, S., Kang, X. (2019, November). Facial sentiment classification based on resnet-18 model. In 2019 International Conference on electronic engineering and informatics (EEI) (pp. 463-466). IEEE.
20. Chen, L. C., Yang, Y., Wang, J., Xu, W., Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3640-3649).
21. Niu, Z., Zhong, G., Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.
22. Itti, L., Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research, 40(10-12), 1489-1506.
23. Mozer, M. C., Sitton, M. (2016). Computational modeling of spatial attention. In Attention (pp. 341-393). Psychology Press.
24. De Boer, P. T., Kroese, D. P., Mannor, S., Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. Annals of operations research, 134, 19-67.
25. Zhang, Z., Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems, 31.
26. Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
27. Yedida, R., Saha, S. (2019). A novel adaptive learning rate scheduler for deep neural networks. arXiv preprint arXiv:1902.07399.
28. Van Laarhoven, T. (2017). L2 regularization versus batch and weight normalization. arXiv preprint arXiv:1706.05350.