Lateralized Approach for Robustness Against Attacks in Emotion Categorization from Images

Harisu Abdullahi Shehu, Abubakar Siddique, Will Browne and Hedwig Eisenbarth

April 6, 2021

Lateralized Approach for Robustness Against Attacks in Emotion Categorization from Images

Harisu Abdullahi Shehu[1][0000−0002−9689−3290], Abubakar Siddique[1][0000−0002−3253−802X], Will N. Browne[1][0000−0001−8979−2224], and Hedwig Eisenbarth[2][0000−0002−0521−2630]

[1] School of Engineering and Computer Science, Victoria University of Wellington, 6012 Wellington, New Zealand
{harisushehu,abubakar.siddique,will.browne}@ecs.vuw.ac.nz

[2] School of Psychology, Victoria University of Wellington, 6012 Wellington, New Zealand
hedwig.eisenbarth@vuw.ac.nz

Abstract. Deep learning has achieved a high classification accuracy on image classification tasks, including emotion categorization. However, deep learning models are highly vulnerable to adversarial attacks. Even a small change, imperceptible to a human (e.g. one-pixel attack), can decrease the classification accuracy of deep models. One reason could be their homogeneous representation of knowledge that considers all pixels in an image to be equally important is easily fooled. Enabling multiple representations of the same object, e.g. at the constituent and holistic viewpoints provides robustness against attacking a single view. This heterogeneity is provided by lateralization in biological systems. Lateral asymmetry of biological intelligence suggests heterogeneous learning of objects. This heterogeneity allows information to be learned at different levels of abstraction, i.e. at the constituent and the holistic level, enabling multiple representations of the same object.

This work aims to create a novel system that can consider heterogeneous features e.g. mouth, eyes, nose, and jaw in a face image for emotion categorization. The experimental results show that the lateralized system successfully considers constituent and holistic features to exhibit robustness to unimportant and irrelevant changes to emotion in an image, demonstrating performance accuracy better than (or similar) to the deep learning system (VGG19). Overall, the novel lateralized method shows a stronger resistance to changes ($10.86 - 47.72\%$ decrease) than the deep model ($25.15 - 83.43\%$ decrease). The advances arise by allowing heterogeneous features, which enable constituent and holistic representations of image components.

Keywords: Adversarial Attacks · CK+ · Emotion categorization · Facial expression · Lateralization · Learning Classifier Systems (LCS) · sUpervised Classifier System (UCS) · VGG19.

1 Introduction

Emotion categorization, based on facial expression, plays an important role in human-computer interaction [1]. Nowadays, there is a growing demand for robots in hotels and retail stores to interact with customers. However, these robots need to understand human emotions in this close-proximity situation. It helps to improve their interaction with the customers to achieve an enhanced customer experience [2]. The term emotion categorization is used here as we contend that humans can superficially express an emotional state that is different from the one that they are experiencing internally.

Deep Learning (DL) based systems have widely been used for image classification [3], including emotion categorization. These systems have demonstrated limited competency by achieving high performance on many state-of-the-art datasets as well as having won many challenges set up by the data science community such as the ImageNet challenge [4]. However, their homogeneous representation of knowledge has made them vulnerable to adversarial attacks, i.e. deliberate changes to the image in an attempt to fool the classifier [5]. For instance, a small modification made to the test or train data might mislead the model to misclassify the input object [6].

On the other hand, biological intelligence supports heterogeneity. It has been hypothesized that lateral asymmetry of the vertebrate brains enables the processing of information at different levels of abstraction, i.e. at a constituent level and holistic level [7]. For instance, the left hemisphere processes sensory input at the constituent (elementary) level, whereas, the right hemisphere processes the same signal at a higher level of abstraction, up to the top holistic level. This heterogeneity concept has recently been shown beneficial at handling noisy data in artificial visual classification systems [8].

The main goal of this work is to create a lateralized system, inspired by the lateralization in biological intelligence, for emotion categorization that will be robust against image changes. As the lateralized approach is considered to be heterogeneous, we anticipate that the novel system will lead to obtaining a much higher accuracy than a homogeneous DL based system when obfuscate changes are made to an image. This is because an emotion, such as happy, may be visible in individual features (e.g. eyes, mouth, jaw) plus their higher-order relationships rather than simply pixel colors (e.g. on a cheek or foreground in an image). Since a constituent or a holistic feature may exhibit robustness against a specific change, these features could be combined, at different levels of abstraction, to obtain overall robustness against a variety of changes. One half of the system will consider the constituent features, whereas, the other half will handle the higher level holistic features. Subsequently, constituent level likelihood and holistic level likelihood will be computed by utilizing constituent and holistic features, respectively. Finally, these likelihoods will be utilized at different levels of abstraction to predict the emotional category of the given image.

The holistic level derived its prediction from a deep model whereas the constituent level derived its prediction from a deep model, as well as a sUpervised learning classifier systems (UCS) to reduce the spread in the average skill of a predictive model, in order to improve the overall accuracy of the system. We will compare the performance of the lateralized system with the performance of a typical DL algorithm before and after changes are made to the images. VGG19 [9] was chosen as the benchmark approach as it is the latest among other VGG models and also because it is a well tested standard model.

The rest of the paper is organized as follows: Section 2 provides the required background knowledge from computer vision and machine learning. It also includes the state-of-the-art relevant techniques that have been investigated for emotion categorization. Section 3 presents the lateralized system, its critical components, and the learning mechanism. The robustness of the developed lateralized approach against attacks is evaluated in Section 4. Section 5 provides a further explanation of the obtained results. It also explains the decision-making process of the novel system. Finally, Section 6 concludes the paper and hints at further studies.

2 Background

The goals of this section are two-fold: first, to review the relevant techniques that have been investigated for emotion categorization in images; and second, to provide the required background knowledge from machine learning and computer vision techniques.

2.1 Computer Vision

This section presents a brief introduction to the attacks that will be applied on the data set to evaluate the robustness of the categorization techniques. It also includes the feature extraction techniques that will be utilized in this work.

Modification Attacks An adversarial attack is any change made to an input image with the intention to mislead a classifier to misclassify the input image. An adversarial attack can be targeted, which aims to mislead the classifier to misclassify an input to a specific/target class, or non-targeted, which aims to fool the classifier to misclassify an input image but does not specify to which class should the input be misclassified. DeepFool is one of the commonly used and well-recognized methods to generate adversarial attacks [17]. It is a simple and accurate perturbation method designed to fool a deep network model. The algorithm repeatedly applies a small change/perturbation to the original image until the newly produced image, which is known as the perturbed image, is predicted incorrectly by the deep model. This work will apply three types of adversarial attacks, i.e. (i) an enhanced version of DeepFool based adversarial attack (named Distractor Attack), (ii) sunglasses based adversarial attack (named Wrapper Attack), and a combination of distractor attack and wrapper attack (named Hybrid Attack) (see Section 4).

Features The histogram oriented gradient (HOG) is one of the commonly used features in computer vision problems [18]. The HOG descriptor utilizes the occurrence of gradient orientation for the detection of complex objects. This utilization of the local gradient makes the HOG features invariant to light conditions, geometric transformation, and color variation. These features assist the lateralized system to accurately classify images based on facial expressions.

2.2 Machine Learning

This section provides an overview of the relevant deep learning and evolutionary machine learning (i.e. learning classifier systems) techniques.

Deep Learning Inspired by the neural connections that exist in the human brain, deep learning (DL) is a methodology of extracting higher-level features from unstructured or raw data [19]. VGG19 is one of the commonly used DL models for classification problems [20] [21]. According to the VGG paper, representation depth is beneficial for classification accuracy. As such, VGG19 is chosen to be used to obtain the constituent and holistic level prediction in the novel lateralized system as it has the latest (19) weight layers among other VGG models such as the VGG16 with 16 weight layers.

Learning Classifier Systems Learning Classifier Systems (LCSs) are a rule-based learning method developed to solve complex problems. They combine a learning component with a genetic algorithm (GA) to perform either supervised, unsupervised, or reinforcement learning. In this research, the sUpervised Classifier System (UCS) [23] is used to predict emotion categories of the constituent level likelihood in the attention phase. UCS is chosen to be used because we know the actual label of the constituent parts and also because the representation of rules in the UCS are straightforward for a human to understand [22].

2.3 Related Work

A large number of techniques have been developed for emotion categorization from images. Convolutional deep networks based techniques have been commonly used for emotion categorization. Recently, a convolutional neural network (CNN) based system is created to classify six basic plus neutral emotions of the facial action coding system [10]. Initially, a face in an image was detected using the Viola-Jones algorithm [11]. Subsequently, the face area was cropped to eliminate the surrounding unimportant data. These cropped images were converted to grayscale and facial features were extracted by using the edge detection technique. Finally, these extracted features were used as input problem instances for the CNN model. This system achieved a performance accuracy of 79.8% on the FER2013 dataset [12]. However, an optimization technique, which might provide an improvement on the accuracy, had not been applied on the CNN. Therefore, the achieved accuracy might be improved with the application of an optimization technique.

Another attempt was made to create a deep CNN based framework for emotion classification in real-time [13]. The proposed network consisted of four separate modules, each of which had multiple layers. The generalizability was achieved by using images from various sources, e.g. a mixture of movie snapshots, emotion datasets such as JAFFE [14], personal photos, and publicly available images from the internet. Not only was the developed application fast, but also the detected emotion per frame in the real-time feed was accurate at almost 96%.

Sokolov et al. [15] proposed a CNN-based system, similar to ResNet [16], to categorize facial expressions by using cross-platform data in real-time. Emotions were estimated in the arousal-valence scale, i.e. how valence or aroused a person is. The developed system achieved a classification accuracy of 63.01%. Considering that the system was developed to categorize emotion based on two different classes (high/low valence or high/low arousal), the achieved accuracy is a little bit better than random guessing. As such, questions remain as to whether it will perform well on the six basic (or the six basic plus neutral) emotional expressions.

Recently, a lateralized system was created for the classification of cats and dogs [8]. The developed system considered the constituents and holistic features of the given image. The lateralized system outperformed other state-of-the-art deep models by $2.15\% - 25.84\%$. The study was conducted based on an artificial visual recognition system to classify cats and dogs. However, it is unknown if such lateralized systems can accurately work as an emotion categorization system since different facial features might have different contributions to different emotions, e.g. the importance of mouth shape to happy compared with fear.

Deep learning algorithms have been used to categorize emotion from images [10], [13], [15]. However, feeding deep models directly with face images considers the color distribution within pixels by representing all pixels in an image to be equally important. This is anticipated to make these techniques vulnerable to even a small change made to the images. Besides, different people might have a slightly different way of expressing the same emotion depending on their cultural background. This work will create a lateralized system that will be robust against changes to the pixels in an image to categorize emotion.

3 Lateralized System

The overall classification scheme of the novel lateralized system, shown in Fig. 1 is similar to a standard supervised learning system except that the prediction can be generated by two phases, i.e. context phase and attention phase. The context phase is developed by using deep models, whereas, the attention phase is developed by using UCSs. Both the phases identify, extract, and utilize constituent and holistic features to make predictions. These techniques are explained below.

3.1 Context Phase

The context phase consists of six deep models (VGG19). Five deep models are used to obtain the constituent level predictions, i.e. prediction about the face, jaw, eyes, mouth, and nose. One reason for doing this is to enable us to move away from end-to-end learning so as to improve performance by testing important feature groups as we know that certain emotional features are innately recognized [24].

The prediction is the probability that a part belongs to a candidate emotion category (class). For this purpose, a face in the given image is initially detected by utilizing the Haar cascade classifier [25]. Subsequently, the position of each constituent part is obtained by using *dlib* (an open-source c++ library for ML) [26]. These position values are used to segment the respective parts. The

Algorithm 1 Algorithm adopted by the context phase

```
Initilize
$L_{t_i} \leftarrow$ List of images
$Prediction_{list} \leftarrow [\,]$
repeat
    for Image i in $L_{t_i}$ do
        $Image_{copy} \leftarrow Original_{image}$
        Detect Face($Image_{copy}$) % Detect face in the image.
        Locate ROI($Image_{copy}$) % Locate the position of region of interest (ROI) such as face, jaw, eyes, nose, mouth, in the image
        Crop ROI($Image_{copy}$) % Get a cropped copy of each ROI
        $L_{roi} \leftarrow$ List of ROIs
        for each ROI in $L_{roi}$ do
            $P_{roi}$ = getPrediction($ROI$) % Prediction of each constituent part from its associated Deep Model. Returns probability
        end for
        for each CP in all emotion categories do % For each constituent part (CP)
            for each $P_i$ in $P_{roi}$ do
                $CP_{cat}$ +=$P_i$ % Overall prediction that a constituent part belongs to a specific category
            end for
        end for
        $CLL$ = argmax($CP_{cat}$) % The category with highest constituent prediction is considered as a constituent level likelihood (CLL).
        $Prediction_{Holistic}$ = getPrediction($Original_{image}$) % get a holistic level prediction from a DL model
        $HLL$ = argmax($Prediction_{Holistic}$) % The category with highest holistic prediction is considered as a holistic level likelihood (HLL).
        if (CLL and HLL Predict the Same Category) then
            Add $CLL$ and $HLL$
            MakeFinalPrediction ()
            GenerateInhibitSignal() % Generate inhibit signal to stop further processing at the attention phase.
        else
            GenerateExciteSignal() % Generate excite signal to do further processing at the attention phase.
        end if
    end for
until $i$ == len($L_{t_i}$) % all test images are processed
```

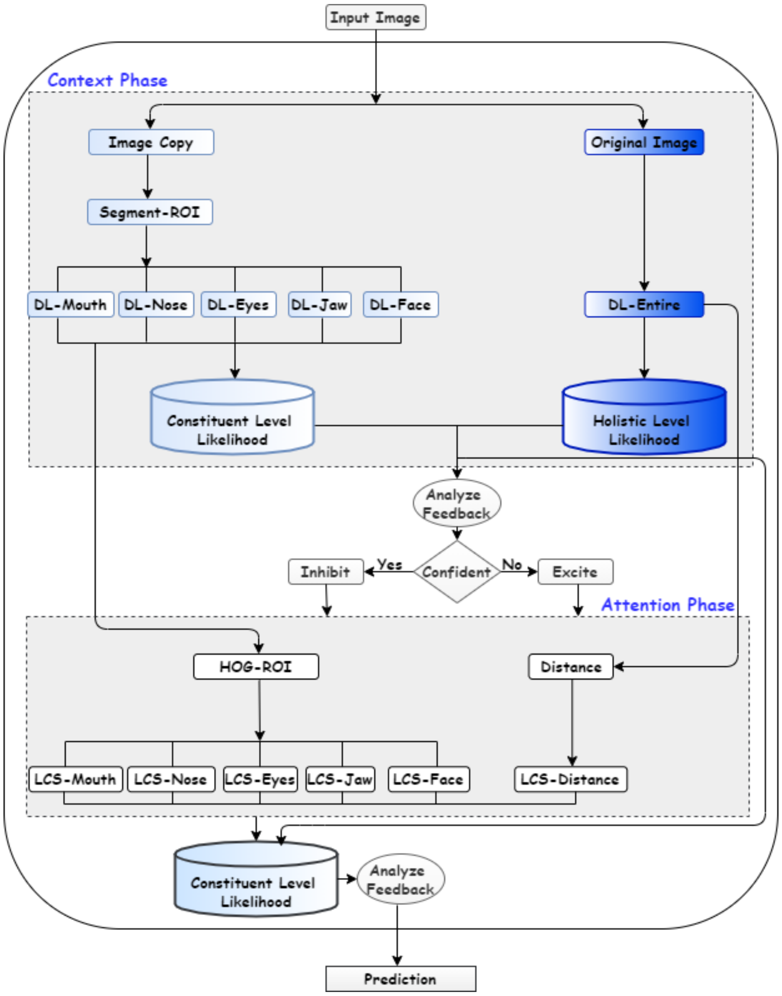


Fig. 1. Flow chart of the laterized system

segmented images are given to the respective deep models and predictions are computed for each emotion category. These prediction values of each category are summed to obtain the vote for that category, e.g. the prediction values of face, jaw, eyes, mouth, and nose for category anger are added to obtain the anger vote.

$$CP = \sum_{i=1}^{n} P_i \tag{1}$$

where CP is the overall prediction that a constituent part belongs to a specific category, P_i represents the probability of each constituent part for that category, and n is the number of total constituent parts. Finally, the CP of each category are compared and the category with highest vote is considered as a constituent level likelihood (CLL), as given below.

$$CLL = \max_{x \in [1,\ldots,m]} CP(x) \tag{2}$$

where m is the number of emotion categories.

Moreover, a deep model is used to obtain the holistic level prediction, i.e. the prediction of the whole image. The resultant highest prediction value for an emotion category is considered as a holistic level likelihood ($\mathcal{HLL}$), as given below.

$$HLL = \max_{x \in [1,\ldots,m]} P(x) \tag{3}$$

where P is the prediction value.

The system analyses the feedback received from the context phase. If the CLL and HLL predict the same category, the system makes the final prediction with confidence and generates an inhibit signal to the attention phase to stop processing. However, if the CLL and HLL predict different categories, the system generates an excite signal to the attention phase to do further analysis. The pseudo-code of the technique developed for the context phase is presented in Algorithm 1.

3.2 Attention Phase

The attention phase consists of six UCSs. Five of the UCSs are used to obtain constituent level predictions about the parts, i.e. face, jaw, nose, eyes, and mouth. This phase utilizes the segmented images generated for each part during the context phase. The HOG features are computed for the segmented images. The resultant features are used as input instances for the respective UCS to obtain the constituent level prediction for each part. Here, the prediction is the probability that each constituent part belongs to a specific category. It is computed by dividing the votes that favor a specific category by the total votes in the UCS prediction array. Subsequently, the respective constituent level prediction values for each category are added to obtain the overall prediction probability for that category (see equation 1). Moreover, we identify the facial landmark[3] using *dlib*[2] [26]. Subsequently, we compute the distance of each (x, y) landmark coordinate from the center of the face, assuming the tip of the nose to be the center. These distances are the holistic level features that represent the relationship between constituents (parts). The sixth UCS is used to obtain the holistic level prediction by using these distances as an input instance. Subsequently, the computed UCS-based constituent level and holistic level prediction values are normalized. These values are added to the corresponding CLL and HLL values from the context phase to obtain the overall prediction probability for each category. Finally, the category with the maximum probability value is predicted.

[3] The facial landmark is a set of coordinates that cover the whole face.

Algorithm 2 Algorithm adopted by the attention phase

```
Check Inhibit Signal() % Stops if receive inhibit signal
$L_{roi}$ ← List of ROIs % List of ROI from the Context Phase
for each ROI in $L_{roi}$ do
    $HOG_{ROI}$ = Compute HOG() % Compute HOG Feature of each constituent part
    $P_{roiUCS}$ = getPredictionFromUCS($HOG_{ROI}$) % Prediction of each constituent part from its associated UCS model
end for
for each CP in all emotion categories do % For each constituent part (CP)
    for each $P_i$ in $P_{roiUCS}$ do
        $CP_{cat}$ +=$P_i$ % Overall prediction that a constituent belongs to a specific category
    end for
end for
$CLL_{UCS}$ = argmax$CP_{cat}$ % The category with the highest constituent prediction is considered as the CLL from UCS.
Face = DetectFace($Image_{copy}$)
Detect FacialLandmark(Face) % Detect (x, y) landmark coordinates from the face
$Dist$ = ComputeDistance() % Get the distance of each landmark coordinate from the center(tip of the nose)
$Prediction_{Holistic}$ = getPredictionFromUCS($Dist$) % Get holistic level prediction from UCS
$HLL_{UCS}$ = argmax($Prediction_{Holistic}$) % The category with highest holistic prediction is considered as a holistic level likelihood (HLL).
Normalize $CLL_{UCS}$ and $HLL_{UCS}$
Add ($CLL_{UCS}$, $HLL_{UCS}$, $CLL$, and $HLL$) % Add all the perceptions from the context phase and the attention phase.
MakeFinalPrediction ()
```

The pseudo-code of the technique developed for the attention phase is presented in Algorithm 2.

4 Experimental Work

4.1 Data Set

This work is designed to evaluate the robustness of the lateralized approach in emotion categorization. This is achieved by conducting experiments on one of the commonly used data set, i.e. CK+ [27]. The data set contains facial expressions of 201 adult participants. Each participant's posed emotions are recorded in the form of a video that has a varied number of image frames, i.e. 10 to 60 frames. This work uses 3368 images of six basic expressions [28] plus neutral expression. These images are extracted from the last-half frames of the videos by using the technique developed by Shehu et. al. [29]. The sample expression images are shown in Fig. 2.

4.2 Experimental Setup

The learning methodology of the context phase is developed by using state-of-the-art VGG19 deep models. These models are trained for 200 epochs. To avoid

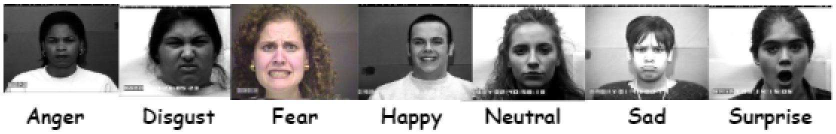


Fig. 2. Sample of six basic plus neutral expressions extracted from the CK+ database, the majority of the images are grayscale.

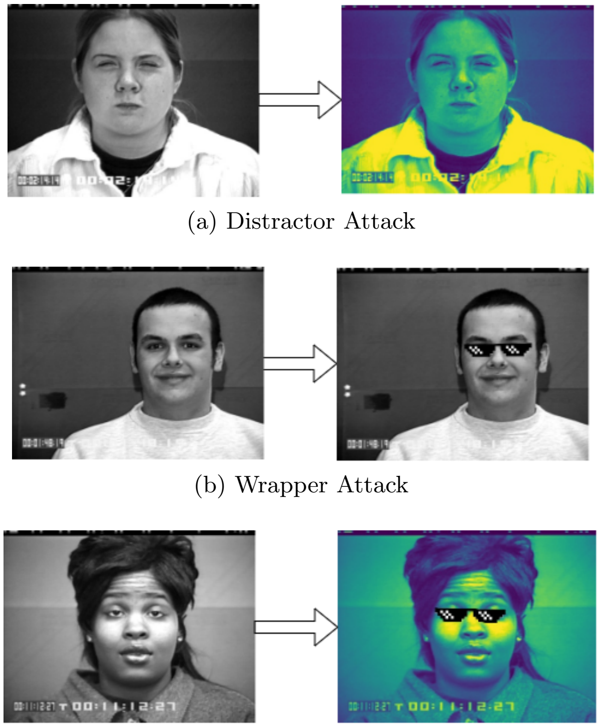
overfitting, the learning rate is reduced by 10% after 80, 100, 120 and 160 epochs, and 5% after 180 epochs. The learning methodology of the attention phase is developed by using UCSs. The configuration settings of the UCS are the same as used by the majority of the researchers [30][31],except for the population size which is set to 10000 as this has shown to give a high performance result. The UCS is coded with upper and lower bound representation and configured as follows: Genetic Algorithm's (GA) threshold θ_{ga} = 20; Crossover probability of χ = 0.8; Crossover type = "two point"; Probability of mutating an allele $\mu = 0.04$; Deletion threshold θ_{del} = 20; Subsumption threshold θ_{sub} = 20, Subsumption accuracy $\epsilon_0 = 0.99$, Initial fitness $f_i = 0.01$; Fitness reduction $\alpha = 0.1$; GA parent selected strategy (s) = tournament; Fraction included in tournament τ = 0.4; Learning rate β = 0.2; finally, the UCSs is set to run over 500000 iterations to ensure convergence. The HOG features are computed with the following parameters: Window size = (64, 64), block size = (16, 16), cell size = (16, 16), window sigma = 4, normalization type = 0, L2-normalization threshold = $2.1x10^{-15}$, number of levels = 64, window stride = (8, 8), and location = (10, 20).

4.3 Experiments

For all the experiments, the expression images are randomly divided into 80% and 20% train and test images, respectively. The novel lateralized system is trained by using the original train images only. The adversarial attacks are applied only to the test images. The performance accuracy of the novel system is evaluated by using the test images, whereas the robustness of the novel lateralized approach is evaluated by using the adversarial images.

Three types of adversarial attacks are applied to the test images, i.e. (i) Distractor Attack, (ii) Wrapper Attack, and (iii) Hybrid Attack as shown in Fig. 3. The distractor attack is generated by using the DeepFool with the following settings: overshoot = 0.02, CenterCrop = 224, mean = $[0.516, 506, 0.496]$, std = $[0.375, 0.365, 0.355]$, $max\_iteration = 1$. The wrapper attack is applied by first detecting the landmark coordinates of the left and right eyes and then adding sunglasses on top of the eyes. However, the width of the sunglasses is reduced to 90% so as not to cover the entire face. The Hybrid attack is the combination of both the distractor attack and the wrapper attack.

Three variants of the lateralized approach are evaluated, i.e. (i) Lateral All, (ii) Lateral Anecdotal, and (iii) Lateral Sense. *Lateral All* (LatAll) utilizes the constituent level and holistic level predictions obtained from all the parts, i.e. eyes, mouth, nose, jaw, and face. *Lateral Anecdotal* (LatAne) utilizes the predic-



(a) Distractor Attack

(b) Wrapper Attack

(c) Hybrid Attack

Fig. 3. Sample original images and the resultant adversarial images after applying distractor, wrapper, and hybrid attacks.

tions obtained from the top three parts, i.e. mouth, jaw, and face. These three parts are selected as they anecdotally contribute to emotion. *Lateral Sense* (LatSen) utilizes the predictions obtained from three sensing parts, i.e. eyes, nose, and mouth. These parts are selected as they are believed to be used by humans for the expression of their emotions [32]. In reporting the statistical test, the letters *a, b, c, and d* are used to indicate if the result is significantly different compared to the VGG19 model. The same letter infers that there is no significant difference whereas different letters show that there is a significant difference.

The experimental results show that all the variants of the lateralized approach obtained a performance accuracy better than or equal to the conventional DL model (VGG19), see Table 1. For original test images (none attack), LatSen obtained a classification accuracy of 99.14%, whereas the VGG19 model obtained an accuracy of 98.86%. The lateralized systems outperformed the conventional DL model as they consider the image at different levels of abstraction. For distractor adversarial images, two of the lateralized systems outperformed the VGG19 model. The LatAll system exhibited strong robustness against the distractor attack and achieved an accuracy of 88%, whereas VGG19 obtained an

Table 1. Classification Accuracy(Highest Accuracy is in bold).

Atttack	**VGG19**	**LatAll**	**LatAne**	**LatSen**	**ANOVA**	
					F	p
None	$98.86 \pm 0.2^{*a}$	$98.86 \pm 1.9^{*a}$	$98.86 \pm 0.1^{*a}$	**99.14**$\pm 2.3^{*a}$	30.03	$< .001$
Distractor	$73.71 \pm 0.2^{*a}$	**88.0**$\pm 6.6^{*b}$	$86.86 \pm 3.8^{*c}$	$73.71 \pm 6.7^{*a}$	2.04 x 10^{29}	$< .001$
Wrapper	$36.57 \pm 0.2^{*a}$	$75.43 \pm 13.0^{*b}$	**87.43**$\pm 11.3^{*c}$	$59.71 \pm 16.1^{*d}$	2.39 x 10^{30}	$< .001$
Hybrid	$15.43 \pm 0.2^{*a}$	$49.28 \pm 8.6^{*b}$	**51.14**$\pm 9.8^{*c}$	$47.14 \pm 10.2^{*d}$	5.20 x 10^{30}	$< .001$

accuracy of 73.71%. Similarly, LatAne exhibited robustness and achieved a classification accuracy of 86.86%. For wrapper adversarial images, all the lateralized system shows better robustness to the attack than the VGG19 model. The lateralized systems have achieved an accuracy of 87.43%, 75.43%, and 59.71% for LatAne, LatAll, and LatSen respectively, compared to the VGG19 model that achieved an accuracy of 36.43%. As the hybrid attack is the strongest attack, the lateralized systems could only achieve an accuracy of 51.14% (LatAne), 49.14% (LatAll), and 34.29% (LatSen). Yet, the achieved accuracy is higher when compared to the VGG19 model that achieved an accuracy of only 15.43%.

The statistical significance of the novel lateralized system is determined by applying one-way ANOVA and post hoc comparison tests (see Table 1). Initially, a one-way ANOVA was conducted to determine the significance of interaction between the groups. Here, we have three groups, i.e. (i) VGG19-LatAll, (ii) VGG19-LatAne, (iii) VGG19-LatSen); and four scenarios, i.e. None (original images), distractor, wrapper, and hybrid. A significant interaction was found between these groups (all $p < .001$).

Similarly, a post hoc comparison of a two-sample t-test with *Bonferroni correction* was performed on the obtained experimental results and no significant difference found for the original images. However, the experimental results for all the lateralized systems after the attack (except the LatSen system for distractor images) were found to be significantly higher than for VGG19 at the $\alpha = .017$.

4.4 Interpretation of Decisions

The decision-making process of the novel lateralized system is interpretable as we can read the rules generated by the LCS. The analysis of predictions obtained for original and adversarial images reveals the reasons behind the robustness against adversarial attacks. The constituent and holistic models at the context phase may generate wrong predictions for adversarial images. In the majority of such cases, the constituent and holistic models predict different classes. Consequently, the system considers that it is not confident to predict the class of the given image and generates an excite signal to the attention phase to do further analysis. After receiving the feedback from the attention phase, the system combines all the predictions and confidently predicts the class of the given adversarial image.

For example, during the classification process of an original image '*Img-org*' (see Fig. 4a)), the holistic level deep model predicted its class as 100% happy.

The constituents level deep models predicted the mouth as 99.99% happy and 0.01% anger, the nose as 99.99% happy and 0.01% disgust, and the eyes as 99.61% happy and 0.39% surprise. Since all the holistic and constituents level models were predicting the same class (see Eq. 2 and 3), the lateralized system predicted the class of the given original image as happy and generated an inhibit signal to the attention phase to stop further processing of the image.

The image '*Img-adv*' was generated by applying a hybrid adversarial attack to the *Img-org*, see Fig. 4b). The holistic level deep model predicted its class as 99.90% disgust and 0.10 sad. However, the constituents level deep models predicted the mouth of the image as 99.68% happy and 0.32% anger, the nose as 99.99% disgust and 0.01% fear, and the eyes as 49.91% sad, 44.68% happy, and 5.41% disgust. In this case, the holistic and constituents level deep models were at odd with each-others. Consequently, the system generated an excite signal to the attention phase for further analysis. Subsequently, the CLL (144.36 happy) was computed by using the equations 2 and the HLL value (99.90% disgust) was computed by using equation 3.

At the attention phase, the holistic level UCS models predicted the class of the given image as 72.73% disgust and 27.27% anger. Similarly, the constituent level UCS model predicted it as a 100% happy mouth. All the other constituent level UCS models were not able to predict their respective parts (could not find a matching rule). These prediction values were normalized and the winner class prediction probability was shared with the system. Subsequently, the returned value was added with the CLL and HLL values computed at the context phase. Finally, the lateralized system predicted the given image class as a happy class with a likelihood of 169.36.

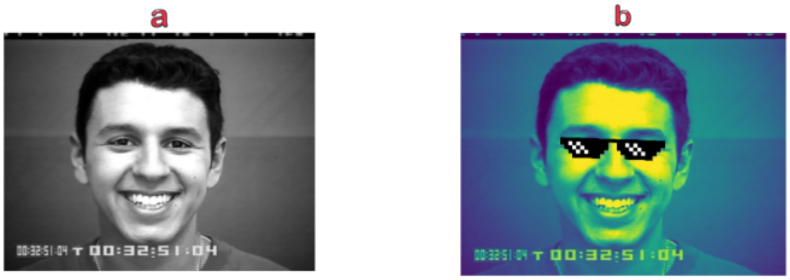


Fig. 4. a) Happy expression, original image(*Img-org*). b) Happy expression, adversarial image after hybrid attack ('*Img-adv*)

5 Discussion

This work is designed to provide robust solutions for emotion categorization against adversarial images. The novel lateralized system considers the given image instance at the constituents level and the holistic level simultaneously. This empowers the novel system to effectively counter the disruptive patterns generated by the adversarial attacks. An adversarial attack needs to successfully challenge all the constituents and holistic patterns to fool the novel lateralized system.

The classification accuracy achieved by all the variants of the lateralized systems is better than or equal to the state-of-the-art VGG19 model. The experimental results demonstrated that the novel system successfully exhibited robustness against the majority of the adversarial attacks. In worse case, the classification accuracy of the novel system (LatAne) was 51.14% against the hybrid attack. It is understandable because hybrid is such a strong adversarial attack that the VGG19 model could not resist it and obtained a very low classification accuracy, i.e. 15.43% (close to random guess). Moreover, the statistical tests show that the improvement in the performance accuracy of the lateralized system is statistically significant.

The decision-making process of the novel lateralized system is interpretable. During the analysis of the results, it is revealed that the lateralized system may wrongly predict some of the constituents or holistic parts but the overall prediction made by the novel system is correct. Moreover, the utilization of inhibit and excite signal assists the novel system to achieve performance efficiency and makes it a more lateralized system rather than an ensemble system. All this suggests that it is worthy to create lateralized classification systems to achieve robustness against noisy and irrelevant real-world data.

In-spite of that, it is also important to keep in mind that these improvements have the negative consequences of increasing the computational costs. While it took an average of 2 hours/run for the VGG19 model in the holistic level to train on an 8GB Graphical Processing Unit (GPU) device GeForce RTX 2080ti with CUDA version 10.2, an average of 2 hours is required to train each of the five deep models in the context phase, i.e. an approximate 10 hours on a single machine (GPU slot). The UCS at the attention phase was run on grid computing, so there is no accurate estimate of time. However, on average, it took about 4-5 hours for each UCS model to run completely. It is noted that this work did not optimize for time.

6 Conclusion

The novel system successfully exhibited robustness against adversarial attacks by applying lateralization. The ability to simultaneously consider the parts of the face (constituents level) and the whole face (holistic level) empowers the lateralized system to correctly classify emotions. The utilization of inhibit and excite signals enable the novel system to efficiently classify original images and pay more attention to the noisy and corrupt images. Consequently, the novel system made correct decisions for badly corrupt images and exhibited robustness against strong adversarial attacks. The novel lateralized system outperformed the state-of-the-art VGG19 model by $15 - 36\%$ point.

Even though the novel lateralized system achieved a significantly better classification accuracy as compared to VGG19, it could not resist the strong adversarial attack (classification accuracy 51.14%). The future work will improve the lateralized method to exhibit robustness against such strong adversarial attacks.

References

1. Brave, S., and Nass, C. (2009). Emotion in human-computer interaction. Human-computer interaction fundamentals, 20094635, 53-68.
2. Shehu H. A., Browne W. N., Eisenbarth H. (2020). An Adversarial Attacks Resistance-based Approach to Emotion Recognition from Images using Facial Landmarks. 2020 IEEE International Conference on Robot and Human Interactive Communication.
3. B. Zoph, V. Vasudevan, J. Shlens and Q. V. Le, ”Learning Transferable Architectures for Scalable Image Recognition,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 8697-8710, doi: 10.1109/CVPR.2018.00907.
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.
5. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
6. Heaven, D. (2019). Why deep-learning AIs are so easy to fool. Nature, 574(7777), 163-166.
7. Grimshaw, G. M., & Carmel, D. (2014). An asymmetric inhibition model of hemispheric differences in emotional processing. Frontiers in Psychology, 5, 489.
8. Siddique, A., Browne, W. N., & Grimshaw, G. M. (2020, June). Lateralized learning for robustness against adversarial attacks in a visual classification system. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference (pp. 395-403).
9. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
10. Babajee, P., Suddul, G., Armoogum, S., & Foogooa, R. (2020). Identifying Human Emotions from Facial Expressions with Deep Learning. 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 2020, pp. 36-39, doi: 10.1109/ZINC50678.2020.9161445.
11. Happy, S.L., Member, S., & Routray, A., (2015). Automatic facial expression recognition using features of salient facial patches. IEEE Transactions on Affective Compututing 6, 1–12.
12. Goodfellow, I., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Lee, D., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L. Xu, B., Chuang, Z., & Y. Bengio. (2013). Challenges in Representation Learning: A Report on Three Machine Learning Contests. In: Lee M., Hirose A., Hou ZG., Kil R.M. (eds) Neural Information Processing. (ICONIP 2013). Lecture Notes in Computer Science, vol 8228. Springer, Berlin, Heidelberg
13. Pathak, K. M., Yadav, S., Jain, P., Tanwar, P. & Kumar, B. (2020). A Facial Expression Recognition System To Predict Emotions. 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2020, pp. 414-419, doi: 10.1109/ICIEM48762.2020.9160229.
14. Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998, April). The Japanese female facial expression (JAFFE) database. In Proceedings of third international conference on automatic face and gesture recognition (pp. 14-16).

15. Sokolov, D. & Patkin, M. (2018). Real-Time Emotion Recognition on Mobile Devices. IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 787-787, doi: 10.1109/FG.2018.00124.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 770–778. doi: 10.1109/CVPR.2016.90.
17. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2574-2582).
18. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.
19. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
20. Mateen, M., Wen, J., Song, S., & Huang, Z. (2019). Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry, 11(1), 1.
21. Oloko-Oba, M. & Viriri, S. (2020). Pre-trained Convolutional Neural Network for the Diagnosis of Tuberculosis: 2020 International Symposium on Vision Computing (ISVC).
22. H. H. Dam, H. A. Abbass, C. Lokan & X. Yao, "Neural-Based Learning Classifier Systems," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 1, pp. 26-39, Jan. 2008, doi: 10.1109/TKDE.2007.190671.
23. Bernadó-Mansilla, E. & Garrell-Guiu, J. M. 2003. Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. Evolutionary computation 11, 3 (2003), 209–238.
24. Addabbo, M., Longhi, E., Marchis, I. C., Tagliabue, P., & Turati, C. (2018). Dynamic facial expressions of emotions are discriminated at birth. PloS one, 13(3), e0193868.
25. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1. https://doi.org/10.1109/cvpr.2001.990517
26. Dlib Python API Tutorials [Electronic resource] – Access mode: http://dlib.net/python/index.html
27. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, 94–101.
28. Ekman, P. & Friesen, W. V., (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, vol. 17, no. 2, p. 124.
29. Shehu, H. A., Browne, W., & Eisenbarth, H. (2020, October). Emotion Categorization from Video-Frame Images Using a Novel Sequential Voting Technique. In International Symposium on Visual Computing (pp. 618-632). Springer, Cham.
30. Siddique, A., Iqbal, M., & Browne, W. N. (2016). A comprehensive strategy for mammogram image classification using learning classifier systems. In Evolutionary Computation (CEC), IEEE Congress on. IEEE, 2201–2208.
31. T. B. Nguyen, W. N. Browne & M. Zhang, (2019). Online Feature-Generation of Code Fragments for XCS to Guide Feature Construction, IEEE Congress on Evolutionary Computation (CEC), pp. 3308-3315.
32. Moore, K. L., Dalley, A. F., & Agur, A. M. R. (2010). Moore's clinical anatomy. United States of America: Lippincott Williams & Wilkins. pp. 843–980. ISBN 978-1-60547-652-0.