Best Fit Activation Functions for Attention Mechanism: Comparison and Enhancement

1st Maan Alhazmi
School of Computing
University of Leeds
Leeds, UK.
scmaalh@leeds.ac.uk

2nd Abdulrahman Altahhan
School of Computing
University of Leeds
Leeds, UK.
a.altahhan@leeds.ac.uk

Abstract—**Activation functions are one of the critical elements of neural networks that allow them to produce non-linear, fine and complex decision boundaries. Yet, their effects are not very well understood in the context of attention mechanisms. In this paper, we investigate the role of two widely used family of activation functions in conjunction with three attention mechanisms on two widely used image classification models; ResNet50 and MobileNetV2. We modified the structures of these classification models by infusing them with three attention mechanisms, CBAM, BAM, and Triplet Attention. In addition, we equipped them with different activation functions, including ReLU, ELU, and a newly proposed activation function that we call ELU+. The resultant models' performances were examined in the domain of facial expression recognition using three datasets; two lab-controlled, CK+ and JAFFE, and one real-world, FER2013. Compared with the baseline models, our results show a significant increase of up to +30% of models' performance when using the newly proposed AF.**

I. Introduction

Since activation functions are used to introduce non-linearity in neural networks, their role is significant, and we may argue that one of the main contributions introduced in AlexNet is changing the activation function from sigmoid to ReLU [1]. However, researchers were more attracted to the more prominent aspect related to the network topology, which showed that stacking more layers has the potential to achieve better results and went in that direction. Therefore, following AlexNet, we started to see deeper and deeper networks such as GoogleNet [2], VGG [3], and ResNet [4]. Activation functions nevertheless received attention on a smaller scale and far fewer new activation functions were introduced [5]–[10]. Some of these activation functions showed improvements of up to 10% for ELU on ImageNet and of 1% for swish with ResNet. However, these studies are not up to date when we consider recent developments in deep learning models. ELU and LReLU attempts to push the mean activation towards 0 to cancel the bias shift problem that occur in neural network learning. More recently, inspired by how humans retrieve information from a scene, attention mechanisms have been introduced to improve the performance of deep learning models by selectively focusing on the most informative parts of the data [11]. Currently, attention mechanisms are dominating the state-of-the-art in various tasks, such as natural language processing [12] and computer vision [13]–[16].

In this work, we investigate the effect of using different activation functions on deep learning architectures with attention mechanisms and test them on facial expressions recognition tasks. We introduce a new variant of ELU activation and conduct an extensive comparative study of different activation functions (AF) used with different attention models on the aforementioned architectures applied to the facial expression recognition domain. Our study empirically demonstrates the effectiveness of a newly proposed activation function that we call ELU+ and sheds light on the importance of AF in the context of attention mechanisms.

This paper proceeds as follows. We start by reviewing the facial expression recognition domain in sections 2. Then, in section 3, we move to discuss the datasets used and the attention mechanisms applied in the domain. Then we discuss different activation functions and introduce our newly proposed ELU+. Then in section 4, we conclude by showing the empirical results of our extensive comparison study.

II. Related Work

Detecting human emotions automatically is an important research domain and has many applications [17]. While recognising facial expressions constitutes the first step towards autonomous and effective communication. In the context of facial expression recognition, different types of deep learning models have been used to tackle several issues that are common in image classification models. These include bias shift between layers due to activation [7], noise in the captured images, high compute and memory requirements and difficulties in extracting useful features. We categorise these models into two groups. One that does not use attention mechanisms and one that does.

A. DL Models For FER

To tackle some of the difficulties that arise in the context of FER, deeper and more elaborate architectures are usually introduced. However, such deep neural network models imply a high number of parameters that require high computational power and a large memory size to fit those parameters. Hence, one of the major challenges in deep learning for facial expression recognition is producing efficient and lightweight models that can be used in resource-constrained devices. To address this issue, eXnet [18] proposed a lightweight but efficient model for FER. They proposed a CNN network based on a parallel feature extraction process resulting in a lighter model with only a 4.57M parameter compared with the VGG19, which has 14.72M parameters. [19] have also addressed the same issue by proposing a model called EmNet that consists of two similar deep convolutional neural network models and a third one for integrating both models. The model is optimised using a joint-optimisation technique. The EmNet gives three predictions, two from the DCNN models and one from the integrated DCNN models. Those predictions are then fused for final classification. The resultant model has only 4.80M parameters.

In facial expression recognition, factors such as age, ethnicity, culture and gender affect the performance of FER systems since they hold significant variations among individuals. Research has shown that the basic facial expression is neutral, and emotions are an addition to the neutral face. People can distinguish the emotion by comparing the expressed emotion with the neutral face. Based on that idea, [20] proposed an approach called De-expression Residue learning (DeRL). DeRL can extract the expressive component from an expression face image and produce a neutral face. This helps solve the issue of individual variations since we could use the neutral face as a reference.

Considering issues in the available datasets, [21] proposed the FN2EN model that deals with the problem of the small number of available datasets on facial expression recognition by introducing a new distribution function to model the high-level neurons of the expression network. Another problem in the available datasets is the annotation errors and biases in those datasets caused by the human factor during the creation process of the datasets. [22] dealt with such a problem by proposing a framework that trains facial expression models from multiple inconsistently labelled datasets and a large-scale unlabelled dataset.

Furthermore, [23] tackled the issue of inter-subject variation of facial expression recognition. They proposed a model called the Identity-Adaptive Generation method (IA-gen), which consists of two parts. The first part generates the six different expressions of a given subject using six different cGANs, each of which generates one of the six emotions where it keeps the identity features and alleviates the features of the expression. The second part is a facial expression recognition model, where they fed the input and the generated image to a pre-trained CNN model.

Moreover, the main focus of the facial expression recognition models is on frontal face images because pose variation is still a challenging task in deep learning. Preserving the expression with frontalisation is one of those challenges. Frontalisation means altering the head pose from a non-frontal face to a frontal face resulting in a synthesised frontal face. [24] aimed to solve this problem by building a frontalisation system that preserves facial expression. They have developed a multi-task model based on GANs that can preserve the expression while frontalising the face from a profile pose to a frontal pose and recognising the expression.

Feature extraction is also one of the most challenging tasks of facial expression recognition since most of the features lie in the mouth region, which is very detailed. Thus it is challenging for a system to capture those features and not get distracted by other features in the image, such as pose and illumination. [25] argue that FER systems tend to suppress variations in the feature extraction, which yields the performance of such a system. They propose a system called Two-branch Disentangled GANs (TDGANs). The system can disentangle the expression features from other features by transferring the expression.

Although the above models have the edge over general models like ResNet, they are domain specific to FER and do not provide a neutral basis to conduct our comparison. Therefore, we will base our study on more general image classification architectures, such as ResNet, to demonstrate that the effect of choosing a suitable activation function can match bespoke and tailored architectures and to ensure a more generalised applicability of our results.

B. DL Models For FER with Attention

The DDL model presented in [26] deals with two major issues in facial expression recognition, which are datasets-related and feature extraction issues. The available facial expression recognition datasets provide labelling for the expressions only, and some other datasets provide labelling for pose and identity. Other factors, such as age, race, and illumination, are not provided in terms of labelling, which limits the performance of facial expression recognition models. Authors in [26] propose a model that disentangles multiple disturbing factors (other than the expression factor) by multi-task learning and adversarial transfer learning. They followed two stages; in the first stage, they pre-trained a disturbance feature extraction model that performs multi-task learning to classify different disturbance factors on a large-scale dataset. In the second stage, they built a disturbance-disentangled model with three sub-networks, a global shared sub-network and two task-specific networks (one for the expression and one for the disturbance). The purpose of the disturbance-disentangled model is to learn the disturbance-disentangled representation for expression classification. Specifically, the expression sub-network utilises a multi-level attention mechanism to extract an expression's features.

In contrast, the disturbance sub-network uses adversarial transfer learning to extract the disturbance features based on the pre-trained model in the first stage. The DDL model achieved state-of-the-art performance in three lab-controlled datasets, which are CK+, MMI and Oulu-CASIA with **99.16%**, **83.67%**, and **88.26%** classification accuracy, respectively. Additionally, the model achieved state-of-the-art in the RAF-DB, a real-world dataset with **87.71%** classification accuracy.

The ADDL model in [27] states that the DDL model has two limitations. It can not adaptively choose the disturbance factor while training and the disentanglement process of the disturbance factor are not performed explicitly. Therefore, they modified the DDL model by introducing the ADFL module, designed to learn the importance weights of the disturbance factor before performing the adversarial transfer learning, and the MINE module, which minimises the correlation between the expression and the disturbance feature. The ADDL is the current state-of-the-art in many facial expression recognition datasets. The classification accuracy scores in the lab-controlled datasets, which are CK+, MMI, and Oulu-CASIA, are **99.64%**, **86.13%**, and **89.44%**, respectively. In addition, The classification accuracy scores in the real-world datasets, which are RAF-DB and AffectNet, are **89.34%**, and **66.20%**, respectively.

Similar to the IA-gen [23], [28] dealt with the inter-subject variation by removing the identity features from the image resulting in an identity-free image with only the expression features. They propose a GAN model called Identity-Free conditional Generative Adversarial Network (IF-GAN). First, an average expression face is computed using all images of the same expression. Then, the input of the IF-GAN is an image of any subject with an expression and the average expression face. The IF-GAN, then, tries to learn to generate an identity-free image using the pair of those inputs. In other words, the generator learns to transfer the expression of the input image to the average identity.

In the context of real-world datasets, [29] proposed a method for facial expression recognition in-the-wild considering partially occluded faces, which is common in the real-world. They have proposed a convolutional neural network with an attention mechanism that alleviates the occlusion and pays attention to the most important features. Additionally, [30] outlined two observations in a real-world facial expression recognition task: the variations in images' sizes and the sensitivity of CNNs to the input size of the image. Thus, they have proposed a network called Pyramid with Super-resolution to deal with the variation in images' sizes.

Following the extraordinary success of the transformers in natural language processing (NLP) proposed in [11], [31] proposed a framework called Vision Transformers (ViT) that utilises transformers on computer vision without the reliance on convolutional neural networks. They split the input images into small patches of 16x16 size, and then those patches are flattened and passed to a linear layer. Then, the output of the linear projection is embedded alongside a position and an extra learnable class embedding. Finally, the embedding sequence is fed to a transformer. This allows the transformer to deal with the images as it does with words in NLP. ViT is a more general model than CNNs because it has less inductive bias. This is because the transformer part of the model has no clue about the positions of the patches in their two-dimensional space, and they have to be learned.

In the field of facial expression recognition, Vision Transformers (ViT) was utilised in [32], where they used it with the addition of a Squeeze and Excitation (SE) block for facial expression recognition. The ViT module extracts the local features using its attention ability, while the SE module captures the global relations from the extracted features. The addition of the SE module helps optimise the learning process, which has the limitation of dealing with small facial expression datasets. The ViT module is pre-trained on the ImageNet dataset and fine-tuned on the FER-2013 dataset. The proposed model is the current state-of-the-art on CK+ dataset with **99.80%** classification accuracy and achieved competitive results on the RAF-DB dataset with **87.22%** classification accuracy.

We argue that more suitable activation functions can alleviate the need for a deeper and more complex architecture. Nevertheless, we study two types of deep learning architectures, the first is ResNet which is relatively deep and has a high number of parameters, and the second is MobileNet which is more lightweight and has fewer parameters. We would like to shed light on the combination of traditional (legacy) deep models with attention and suitable activation functions. The idea is to provide a basis for comparison and demonstrate the benefit of using different activation functions on these traditional architectures and to prove that they are comparable with other architectures that utilise transformers.

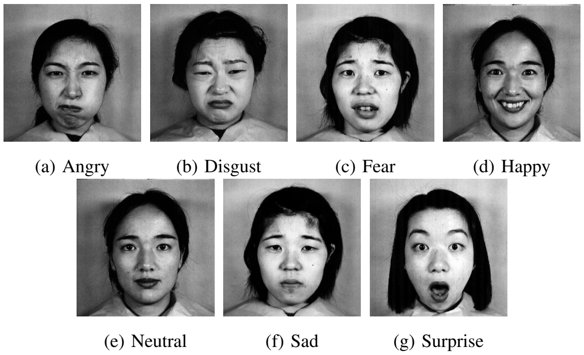
C. Activation Functions

Since replacing the sigmoid activation function used in LeNet with the ReLU activation function in AlexNet, the ReLU became the default choice in deep learning models due to efficiency reasons compared with the sigmoid and tanh activation functions [33]. ReLU deals well with the vanishing gradient problem because its derivative is not contractive. On the other hand, it does not activate neurons with negative input values. This reduces the number of active neurons, which in turn makes learning faster but limits the model's learning capability. In addition, it introduces a bias shift from early layers to later layers because their activation's mean is always positive [7]. This led to the need of other techniques, such as batch normalisation, to help limit the shift.

Modifications to the ReLU activation function have been proposed since then to overcome its limitations. For example, LReLU [5] and PReLU [6] were proposed to overcome such limitations by introducing a slope in the negative direction. In the LReLU, the slope is fixed, while in the PReLU, the slope is a learnable parameter. The ELU activation function proposed in [7] is another modification to the ReLU where the negative values are represented using a log curve instead of a straight line compared with the LReLU and PReLU. The PReLU-net proposed in [6] was the first network to surpass the human performance in the ImageNet challenge. In addition, the ELU activation function [7] enhanced the training speed and led to a better generalisation.

III. Methodology

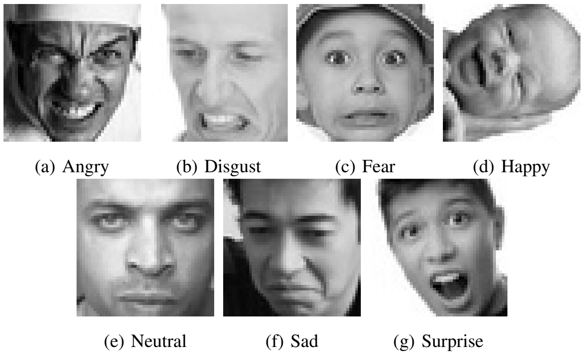
To investigate the effectiveness of adding attention mechanisms to convolutional deep learning models, two complex CNNs models were chosen in this work, namely ResNet50 and MobileNetV2. In terms of attention mechanisms, the focus of this work is on the attention mechanisms that combine channel and spatial attention and are CNN based. Therefore,



(a) Angry (b) Disgust (c) Fear (d) Happy

(e) Neutral (f) Sad (g) Surprise

Fig. 1: Sample images from JAFFE dataset



(a) Angry (b) Disgust (c) Fear (d) Happy

(e) Neutral (f) Sad (g) Surprise

Fig. 2: Sample images from FER-2013 dataset

three-channel and spatial attention mechanisms were chosen, namely CBAM, BAM, and Triplet Attention. The two models were trained with the addition of each attention mechanism. The chosen attention mechanisms are designed as blocks that can be easily integrated into any CNNs model. The attention blocks were plugged before the residual connection in the building blocks of both models.

Additionally, the two models were tested using different activation functions to investigate their role. By inception, ResNet50 base model uses ReLU activation functions throughout its layers. This was replaced by ELU. Similarly, MobileNetV2 uses the ReLU6 activation function throughout its layers. This was replaced by ELU6, a new activation function created by modifying ELU to match ReLU6. In addition, further modifications of ELU are proposed resulting in two new activation functions that we call ELU+ and ELU7+. The performance of each of the above models, when equipped with a combination of each of the three activation functions (ReLU, ELU and ELU+), along with each of the three different attention mechanisms (CBAM, BAM and TA), are compared with the base models. For training and testing, we used 10-folds cross-validation (CV) for the CK+ and JAFFE datasets, while the FER2013 dataset has a substantially dedicated testing set which alleviates the need to use CV. CK+ is a fairly easy dataset, and we achieved perfect and near-perfect scores without even applying any of our suggested modifications. Therefore, in order to make the CK+ dataset more challenging and use it for benchmarking, we have applied the following data augmentation techniques: Color Jitter, Random Solarize, Random Affine, and Random Horizontal Flip. The following subsections explain the datasets, attention mechanisms and activation functions used in this work.

A. Datasets

In [34] they have created a lab-controlled dataset for emotions recognition, which is the Extended Cohn-Kanade Dataset (CK+). The set of emotions presented in this dataset are Angry, Disgust, Fear, Happy, Sadness, Surprise, and Contempt. CK+ consists of 593 video sequences from 123 different individuals. Each video shows a transition from a neutral face to the expressed emotion. Only 327 videos are labelled with one of the emotions set, aka emotion class. CK+ is one of the widely used datasets in emotion recognition. In [35], authors created a lab-controlled dataset for emotion recognition, which is the Japanese Female Facial Expression (JAFFE). Ten different female Japanese individuals were asked to express the six basic facial expressions plus a neutral face. A total of 213 8-bit grayscale images were taken and viewed by 60 different Japanese individuals to come up with an average semantic rating for each image. In [36] they have introduced a real-world dataset for emotions recognition, which is the Facial Expression Recognition 2013 (FER2013). The FER2013 dataset consists of 35,685 examples in a grayscale format. The images are classified into one of the six basic emotions plus neutral.

B. Attention Mechanisms

[14] proposed the bottleneck attention module (BAM) that increases the receptive field using a dilated convolution. The module consists of two branches: the channel attention branch and the spatial attention branch. The channel attention branch computes the channel attention using average pooling, then an MLP, and finally applies batch normalisation. At the same time, the spatial branch computes the attention map using a bottleneck-structured convolution with a dilation. The results of the two branches are reshaped to match the dimensions of the input feature maps. Next, they are added together and passed to the sigmoid function.

[16] proposed the convolutional block attention module (CBAM) that has two subsequent operations: channel attention followed by spatial attention. The channel attention is similar to the SE block but captures the global information using average and max pooling in parallel. The spatial attention module generates the attention map using a convolutional layer.

CBAM and BAM compute the spatial and channel attention independently, which might result in the loss of discriminative information across different dimensions. [15] proposed the triplet attention block, which considers cross-dimension interactions. The triplet attention block has three branches; two branches capture the interactions between the channels and one of the spatial dimensions (the height or the width), while the third branch captures the spatial attention. The input data is rotated (90^o rotation anti-clockwise) along the desired spatial axes in the first two branches. Then the results are passed to a z-pool layer that reduces the dimensionality of the other spatial dimension to two. In other words, for the first branch that capture the cross-dimension interaction between the channels and the height dimension, the dimensions of the input tensor are transformed from $(W \times H \times C)$ to $(2 \times H \times C)$ by computing and concatenating the average and the max pooling across the dimension. The results are then convolved with batch normalisation, resulting in an output of the shape $(1 \times H \times C)$ that is passed to a sigmoid function to produce the attention weights. The attention weights are then applied to the rotated input, and the results are rotated along the desired dimension (90^o rotation clockwise). The same process is followed for the second branch but on the W axis. For the last branch, the channels are reduced to two using the z-pool layer, and then the results are convolved and batch normalised. Finally, the results are fed to a sigmoid function to produce the attention weights, which are then applied to the input. The final result is computed by averaging the results of the three branches.

C. ELU: Better Fit Activation Functions

Given the discussion of the above-mentioned attention mechanisms, one can conjecture that an activation function that is smooth and uses exponentiation would be more suitable than the simple ReLU. This is because attention needs to peak naturally for specific excitation, quite sharply and quickly, while it should dampen and neutralise other irrelevant excitation. Also, the sigmoid function which has been used in all of these mechanisms in its core depends heavily on specific exponentiation formula. ELU activation function satisfies these traits (smooth and uses exponentiation) and forms a better fit for the discussed attention mechanisms, although we will show how to amend it in the next section to get even better AF. ELU activation function applies the exponential linear unit. It returns x if x is greater than 0; else, it returns $e^x - 1$, where e^x is the exponentiation of x, multiplied by α which is a hyperparameter with a default value of 1. See (1) and Fig. 3.

$$ELU(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{if } x \leq 0 \end{cases} \tag{1}$$

Just like the ReLU6 activation function, the ELU activation function was modified to return the minimum value between x and 6. See (2) and Fig. 3.

$$ELU6(x) = \begin{cases} min(x, 6), & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{if } x \leq 0 \end{cases} \tag{2}$$

D. ELU+: Best Fit Activation Functions to Support Attention and Normalisation

A further modification of the ELU function was carried out to produce the ELU+, which returns x if x is greater than β, where β is a hyperparameter, see (3) and Fig. 4. The motivation for this change is that we realised that activation functions in general, and ELU in particular, work better for attention if we allow it to model non-linearity on the input beyond 0. Given that the majority of deep learning models perform normalisation in one way or the other, specifically batch normalisation, we know that the model sensitivity towards the range $[0, 1]$ should be made higher than other inputs in the range $[1, \infty]$ or $[1, n]$. Therefore, we opted to keep the function as close as possible to its original form but increase its sensitivity beyond 0 to the interval $[0, \beta]$, where β can be tailored towards the application and problem domain's needs. We have tried to add the linear part (straight line) at the end of the non-linear curve that resides along the $[0, \beta]$ interval (i.e. on the top of the little peak that can be seen in Fig 4) to enhance its smoothness, but that did not produce good results. In addition, similar to ReLU6 used in MobileNetV2, the ELU7+ is introduced to replace the ReLU6, which returns the minimum value between x and seven if x is greater than β. Seven was specified via trial and error, which seems to work better than other values between 1 to 10. See (4) and Fig. 4.

$$ELU+(x) = \begin{cases} x, & \text{if } x > \beta \\ \alpha(e^x - 1), & \text{if } x \leq \beta \end{cases} \tag{3}$$

$$ELU7+(x) = \begin{cases} min(x, 7), & \text{if } x > \beta \\ \alpha(e^x - 1), & \text{if } x \leq \beta \end{cases} \tag{4}$$

IV. Results

This section shows the results of the previously explained experiment. Table I illustrates the results of ResNet50 after adding attention mechanisms with three activation functions, ReLU, ELU, and ELU+ on the CK+, JAFFE, and FER2013, respectively. For this architecture (ResNet50), the first observation is that the ELU activation function outperformed the ReLU in every dataset with all attention mechanisms. This suggests that ELU is, in fact, a better fit for models that use attention mechanisms than the traditional ReLU. This insight can be interpreted due to both the smoothness of the ELU in comparison with ReLU, particularly when transitioning from negative to positive input, and due to allowing negative input to take negative values, which helps in reducing the bias shift. Furthermore, our ELU+ activation function outperformed the ELU activation function in the majority of comparisons (except for BAM on the CK+ dataset and Triplet Attention on the FER2013 dataset). This shows that the proposed activation function offers an important alternative that can gain a performance boost for attention mechanisms, particularly for BAM and CBAM.

The accuracy achieved on the CK+ without using attention is already high when moved from ReLU to ELU. The model gained 6%. This seems to have saturated the possibility of raising the accuracy by using our ELU+ activation function. On the CK+ dataset with attention, the best performance achieved is **96.23%** for CBAM. The best classification accuracy on the JAFFE dataset is **94.87%** after adding CBAM to the ResNet50 using ELU+ with $\beta = 0.7$. On the FER2013, the best performance achieved is **60.90%** after adding triplet attention to the ResNet50 using the ELU activation function.

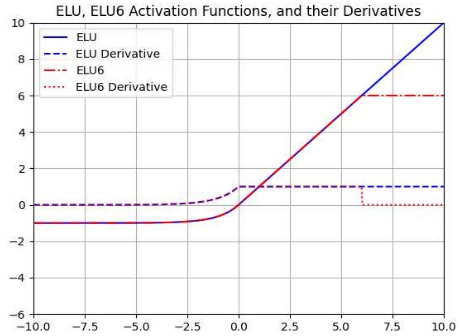


Fig. 3: ELU, ELU6 Activation Functions, and their Derivatives

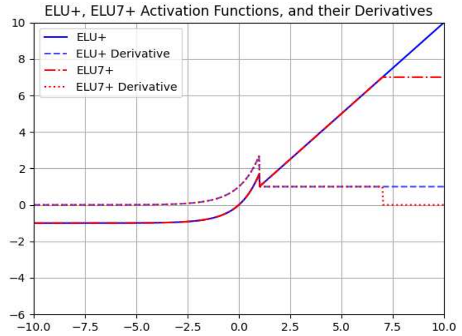


Fig. 4: ELU+, ELU7+ Activation Functions, and their Derivatives

However, the ELU+ outperformed the baseline models using CBAM, and BAM. The second best performance achieved is **60.50%** with CBAM using ELU+ with $\beta = 0.7$. All of the best classification performances were achieved using the ELU+ activation function except on CK+ with BAM and on FER2013 with Triplet Attention, which proves the validity of replacing the ReLU activation function with the newly proposed ELU+ activation function.

The ELU+ activation function shows a significant performance, especially with the FER2013 dataset, which is a real-world dataset and more challenging due to the variations in viewpoints and illuminations. We can infer that the model's capability to deal implicitly with such variations has improved using such an activation function.

Furthermore, the results of the more lightweight deep learning architecture MobileNetV2 after adding attention mechanisms with the three activation functions, ReLU6, ELU6, and ELU7+ are illustrated in tables I. We can see that the newly created ELU7+ activation function has significantly improved the performance of the MobileNetV2 of +25% on the JAFFE dataset, from **75.56%** to **90.63%**. The performance of the baseline MobileNetV2 model using ReLU6 with CBAM was increased by +31% after changing the ReLU6 activation function with the ELU7+ activation function from **61.58%** to **92.94%**. The same observation holds for the other models using the BAM and Triplet Attention. The performances were improved by a significant margin. Fig. 5 shows the training and testing curves of the best-performing model compared with the baseline models. We can clearly see that the testing accuracy (shown in red) started to increase around the tenth epoch, leaving a noticeable margin compared with the baseline models. Such results are extremely encouraging and open the window for more investigation of the ELU7+ activation function. It should be noted that although ELU6 performed well in training, but in testing, it performed worse than all other AFs due to overfitting. On the CK+ dataset, the best performance achieved is **99.59%** after adding Triplet attention to the MobilNetV2. The best classification performance in the JAFFE dataset is **94.35%** after adding BAM to the MobileNetV2 using ELU7+ with $\beta = 1$. In the FER2013, the best performance achieved is **65.30%** after adding triplet attention to the MobileNetV2 using the ELU7+ activation function with $\beta = 1$. Fig. 6 shows the training and testing curves of the best-performing model compared with the baseline models. Even though the model using the ELU7+ activation function was slower in terms of convergence, it learned to generalise better. Similar to ResNet50, all of the best classification performances were achieved using the ELU+ activation function except on CK+ and FER2013 with BAM, which proves the validity of replacing the ReLU6 activation function with the newly proposed ELU7+ activation function. In fact, it is more encouraging with MobileNetV2 due to the significant improvement achieved on the JAFFE dataset. The results using the ELU+ and the ELU7+ activation functions are worth more intensive investigations since they seem to have significant potential. One possible justification for their performance is noticeable by looking at their derivatives. In Fig. 4 the solid blue line shows the ELU+ activation function, while the dashed blue line shows its derivative. We can see that the function gave more importance to the values between zero and one, which -in our opinion- allowed achieving such results. In contrast, when we look at the ELU activation function and its derivative, it gives similar importance to all values that are greater than 0, see Fig. 3.

V. Conclusion

Simple but effective modifications to the internal structure of neural net components can lead to significant improvements. In this work, we proposed a new activation function that we call ELU+. We empirically demonstrated the effectiveness of the newly proposed ELU+ activation function compared to ReLU and ELU in ResNet50 infused with three attention mechanisms, CBAM, BAM, and Triplet Attention. We have also demonstrated the effectiveness of the newly proposed ELU7+ activation function compared with ReLU6 and ELU6 in MobileNetV2 infused with the same attention mechanisms. The significant increase of up to +30% of models' performance demonstrated a significant potential for using the ELU+ activation function in future deep learning models that utilise attention. In the future, we plan to conduct more investigation to obtain further insight into the newly proposed activation

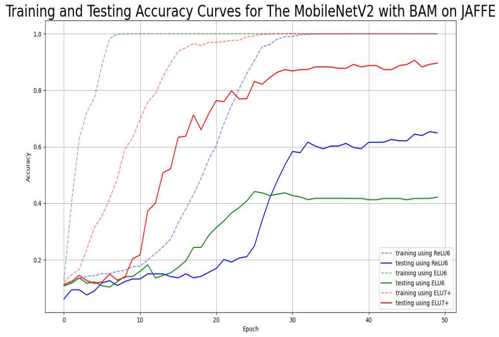


Fig. 5: Training and Testing Accuracy Curves for MobileNetV2 with BAM on JAFFE. (Best Performing Model on JAFFE)

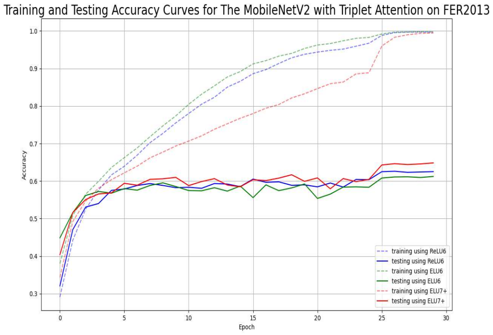


Fig. 6: Training and Testing Accuracy Curves for The MobileNetV2 with Triplet Attention on FER2013. (Best Performing Model on FER2013)

TABLE I: Evaluation of ReLU/ReLU6, ELU/ELU6, and ELU+/ELU7+ Activation Functions using the **accuracy** metric on ResNet50/MobileNetV2 w/o attention mechanisms and with CBAM, BAM, and Triplet Attetnion (TA) on the CK+, JAFFE, and FER2013 dataset. The subscripted values indicate the value of β used in the ELU+/ELU7+ activation function.

Dataset	Activation	**ResNet50**								**MobileNetV2**							
		w/o AM		CBAM		BAM		TA		w/o AM		CBAM		BAM		TA	
		%	σ	%	σ	%	σ	%	σ	%	σ	%	σ	%	σ	%	σ
CK+	ReLU/6	91.95	4.9	87.47	4.9	86.75	2.6	91.03	4.3	97.35	0.8	96.43	1.4	97.35	1.5	96.43	1.1
	ELU/6$_{(0)}$	**97.66**	0.9	93.38	3.0	**94.50**	2.2	90.62	1.8	**99.80**	0.4	98.37	1	**99.70**	0.9	99.50	0.9
	ELU+/7+$_{(0.3)}$	96.94	1.7	94.90	1.8	91.24	0.9	92.46	2.3	99.69	0.5	**99.39**	1.2	92.76	2	99.49	0.9
	ELU+/7+$_{(0.5)}$	94.09	3.1	95.81	1.8	91.23	3.2	**95.93**	2.5	94.90	1.4	**99.39**	0.9	96.23	1.6	99.39	0.9
	ELU+/7+$_{(0.7)}$	88.38	1.8	**96.23**	2.7	82.47	4.0	90.42	2.2	98.98	1.4	99.29	0.9	97.14	2	**99.59**	0.7
	ELU+/7+$_{(1)}$	84.81	5.3	91.12	2.6	83.69	4.7	91.54	3.6	97.66	1.4	98.98	1.2	97.25	1	99.18	1.2
JAFFE	ReLU/6	80.30	12	82.23	9.5	80.32	8.4	79.42	8.5	75.56	7.2	61.58	14	70.04	9.9	68.59	11
	ELU/6$_{(0)}$	89.72	6.8	90.67	5.4	87.85	7.1	89.74	7.4	58.70	6.4	55.43	8.8	47.90	8.5	49.74	6.1
	ELU+/7+$_{(0.3)}$	90.15	8	91.58	6.5	89.22	6.3	91.58	6.8	57.66	9.3	60.06	7.2	61.60	10	52.58	6.8
	ELU+/7+$_{(0.5)}$	**93.87**	6	92.99	6.0	**92.46**	5.7	93.48	6.6	78.38	7.4	60.17	7.6	80.76	4.9	64.72	10
	ELU+/7+$_{(0.7)}$	89.68	7.2	**94.87**	4.4	84.50	9.5	**94.35**	4.6	89.22	3.6	82.16	7.5	89.15	6.7	86.93	7
	ELU+/7+$_{(1)}$	88.24	4.8	91.52	6.7	84.00	9.1	91.04	6.2	**90.63**	7.3	**92.94**	3.8	**94.35**	4.6	**92.14**	8.9
FER13	ReLU/6	56.00	n/a	57.00	n/a	54.60	n/a	57.80	n/a	**64.30**	n/a	62.00	n/a	**64.50**	n/a	63.30	n/a
	ELU/6(0)	59.00	n/a	59.50	n/a	58.70	n/a	**60.90**	n/a	61.70	n/a	61.80	n/a	62.10	n/a	62.50	n/a
	ELU+/7+$_{(0.3)}$	59.40	n/a	57.90	n/a	50.50	n/a	59.20	n/a	62.00	n/a	61.30	n/a	62.90	n/a	62.10	n/a
	ELU+/7+$_{(0.5)}$	**60.80**	n/a	59.80	n/a	**59.50**	n/a	60.30	n/a	62.10	n/a	62.20	n/a	64.10	n/a	62.90	n/a
	ELU+/7+$_{(0.7)}$	59.40	n/a	**60.50**	n/a	56.50	n/a	60.20	n/a	63.40	n/a	**63.70**	n/a	64.10	n/a	63.40	n/a
	ELU+/7+$_{(1)}$	56.00	n/a	59.20	n/a	51.80	n/a	58.50	n/a	63.20	n/a	**63.70**	n/a	51.80	n/a	**65.30**	n/a

function as well as automating the process of optimising the threshold β.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” 2014. [Online]. Available: https://arxiv.org/abs/1409.4842

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[5] A. L. Maas, “Rectifier nonlinearities improve neural network acoustic models,” 2013.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” 2015. [Online]. Available: https://arxiv.org/abs/1502.01852

[7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” 2015. [Online]. Available: https://arxiv.org/abs/1511.07289

[8] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” 2016. [Online]. Available: https://arxiv.org/abs/1606.08415

[9] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” 2017. [Online]. Available: https://arxiv.org/abs/1710.05941

[10] K. Mai Ngoc, D. Yang, I. Shin, H. Kim, and M. Hwang, “Dprelu: Dynamic parametric rectified linear unit.” New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3426020.3426049

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057. [Online]. Available: https://proceedings.mlr.press/v37/xuc15.html

[13] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, “Attention mechanisms in computer vision: A survey,” *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, mar 2022. [Online]. Available: https://doi.org/10.1007

[14] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, “Bam: Bottleneck attention module,” 2018. [Online]. Available: https://arxiv.org/abs/1807.06514

[15] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” 2020. [Online]. Available: https://arxiv.org/abs/2010.03045

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam:

Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[17] R. Santhoshkumar and M. Kalaiselvi Geetha, "Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks," in *Procedia Computer Science*, vol. 152. Elsevier B.V., jan 2019, pp. 158–165.

[18] M. N. Riaz, Y. Shen, M. Sohail, and M. Guo, "exnet: An efficient approach for emotion recognition in the wild," *Sensors*, vol. 20, no. 4, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/4/1087

[19] S. Saurav, R. Saini, and S. Singh, "Emnet: A deep integrated convolutional neural network for facial emotion recognition in the wild," *Applied Intelligence*, vol. 51, no. 8, p. 5543–5570, aug 2021. [Online]. Available: https://doi.org/10.1007/s10489-020-02125-0

[20] Yang, Huiyuan, and Ciftci, Umur and Yin, Lijun, "Facial expression recognition by de-expression residue learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

[21] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, 2017, pp. 118–126.

[22] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *ECCV (13)*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Springer, 2018, pp. 227–243.

[23] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 2018, pp. 294–301.

[24] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 2018, pp. 263–270.

[25] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2359–2371, 2021.

[26] D. Ruan, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2833–2841. [Online]. Available: https://doi.org/10.1145/3394171.3413907

[27] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *Int. J. Comput. Vision*, vol. 130, no. 2, p. 455–477, feb 2022. [Online]. Available: https://doi.org/10.1007/s11263-021-01556-7

[28] J. Cai, Z. Meng, A. S. Khan, J. O'Reilly, Z. Li, S. Han, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1344–1348.

[29] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[30] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[32] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021. [Online]. Available: https://arxiv.org/abs/2107.03107

[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323. [Online]. Available: https://proceedings.mlr.press/v15/glorot11a.html

[34] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[35] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset," Apr. 1998. [Online]. Available: https://doi.org/10.5281/zenodo.3451524

[36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. H., "Lee," *Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. Neural Networks*, vol. 64, pp. 59–63, 2015.