Working with Facial Expression Recognition (FER-2013) Dataset

Christian Sutton
Bear, Delaware, USA
csutto16@alumni.jh.edu

***This project studied the performance of neural networks for facial emotion recognition using the fear 2013 data set. Multiple models were trained using AlexNet like architecture, ResNet-34 and Convolutional block attention model modified ResNet-34 architecture. The best performance was obtained in the ResNet architecture models and averaged about 60% accuracy.* This project investigated manipulating the hyperparameters of batch size and learning rate to ascertain the effect on performance of the data set.**

I. Introduction

There is an increasing need to identify and categorize images of human emotion, for instance, in the realm of security and terrorism. Expression of human emotion is very nuanced in its manifestation on a human's face. Detecting human emotion is further complicated by the fact that changes in occlusions pose variations and illumination can make the data difficult to train on.

The work on categorization of human emotion was first touched on by psychologist Paul Ekman Wallace. He recognized six primary human emotions happy, sad, anger, surprise, fear, disgust, and neutral [1]. However, with the increasing computational power of compute resources and the increasing volumes of generated data especially image data, the last decade has shown increasing interest in detecting and classifying human emotion using convolutional neural networks.

Prior to convolutional neural networks the traditional approach was to use a two-step process to extract features and then classify them. It was quickly found out that this approach was not achieving satisfactory results due to high intraclass variation [1]. The rise of deep learning techniques led to the discovery of convolutional neural networks. CNN's provided researchers with a new network architecture that was lightweight with respect to the number of matrix weights needed for the model. A smaller number of weights in the convolutional neural network led to researchers expanding the depth and size of the networks for a given amount of computer resources. Additionally with the increased depth and size of the networks such as AlexNet and ResNet there came increased performance in competitions during around mid-decade 2013. Results of the competitions suggested that CNN's are capable of outperforming hand design features for image classification tasks such as facial expression recognition [2].

This Project aims to test two convolutional neural network architectures on the facial expression recognition data set (FER-2013). AlexNet and ResNet are two popular convolutional neural network CNN architectures for which researchers developed for submittal to image competitions. The goal is to explore the performance of these architectures on the FER-2013 data set with the understanding of a tight time constraint which impacts tuning the models.

This project consisted of implementing the two neural network architectures and adjusting hyperparameters to obtain the best model fit given the time constraints. The types of hyperparameters adjusted are network architecture, batch size learning rate, and number of epochs.

II. Data

In 2013 as part of the ICML 2013 workshop on representation learning a Kaggle competition was launched based on the FER-2013 data set. The data set was created using a Google image API search for images of faces that match a set of emotion related keywords [2]. The competition's obvious goal was to find the best solution for facial recognition from the data set of images with seven emotions as described above. To top three teams in the competition all used convolutional neural networks as solutions to the FER-2013 data set [1].

The dataset consists of 28,709 training images and 3,589 test images. The images depict faces of people displaying one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and a neutral expression. The FER-2013 data set has been widely used in facial expression recognition competitions and has been a common benchmark for evaluating algorithms. The data set was obtained from kaggle.com. The version hosted on the website consisted of a downloadable data set consisting of JPEG images, of size 48 x 48, placed into separate folders which were labeled with the appropriate emotion depicted in the images.

Some problems have been noted with the data set in terms of its generality and trainability. The following are some examples of issues noted in the data set.

- Dataset is highly biased towards certain demographics this can lead to poor generalization.
- The dataset does not include images of complex emotions.
- Some of the images are mislabeled, which can affect the accuracy of the training data set.
- Some of the images have artifacts over top of the faces such as text.

- The images in the data set are of low resolution which makes it hard to capture facial expressions.

Below is an example of the images present in the FER-2013 dataset, see Fig 1.

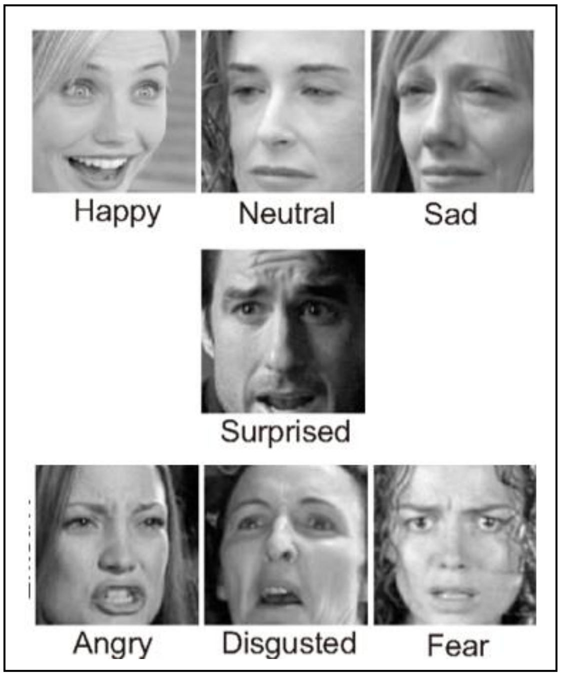


Fig 1. Example images from the FER-2013 Dataset

III. Network Architectures used

A. *AlexNet like Convolutional Neural Network (CNN)*

AlexNet is a well-known CNN architecture introduced in 2012. This network achieves state-of-the-art performance on the ImageNet database. AlexNet consists of five convolutional layers followed by three fully connected layers. This project uses a similar architecture to AlexNet both some modifications to meet the projects needs for a fast turnaround time [3].

For this project the AlexNet like architecture is modified with the following properties, see Fig 2 for a graphical representation.

- All convolving kernels are of size 3 by 3 for this architecture whereas AlexNet uses various size kernels depending on what layer they operate in.
- The input image is constrained to be 32 by 32 for this model whereas the AlexNet architecture had dimensions of 224 by 224.
- Layer 2 has a feature map that is the result of 96 kernels and not 128 like AlexNet has.
- Layer 5 has a feature map that is the result of 256 kernels and not 128 as AlexNet has.
- Layers 6 and 7 are fully connected with 512 and 256 neurons respectively.
- The SoftMax layer is only 7-way and not 1000-way like AlexNet. This matches the scope of the problem – classification of 7 classes of facial expressions.

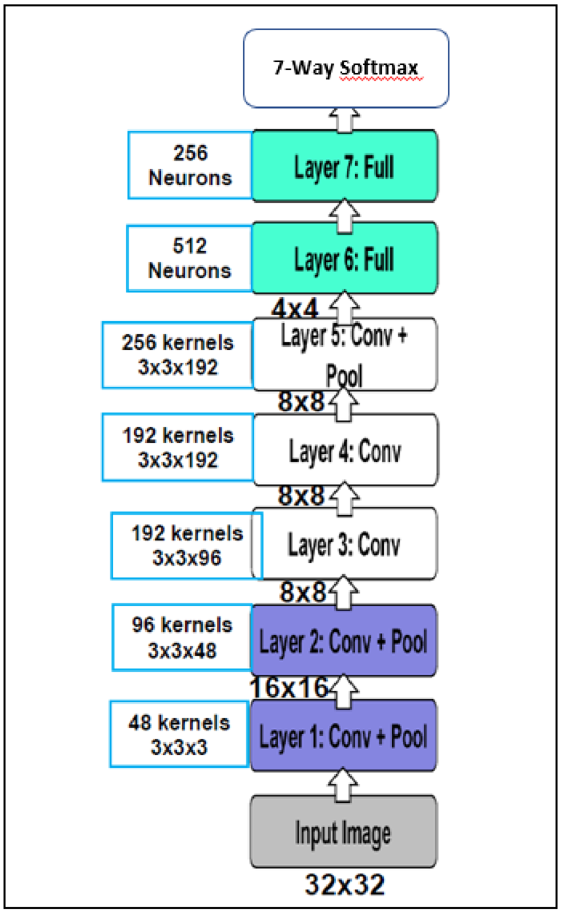


Fig 2. AlexNet like architecture used in project.

B. *ResNet-34*

ResNet was introduced in 2015 as a means of tackling the vanishing gradients problem in deep neural networks. The vanishing gradients problem arises when the gradients of a loss function with respect to the matrix weights of the early layers of network become extremely small. The small matrix weights cause the early layers to train extremely slowly or not at all. VGG net was one of the early deep convolutional neural networks that achieved impressive results in image classification. However, due to the vanishing gradients problem VGG net architecture was limited if a researcher wanted to make the models easily trainable. ResNet, short for residual network, developed by Microsoft researchers, introduced a new architecture that helped the deep neural network train effectively. Resnet implements residual connections that bypass one or more convolutional layers allowing the gradient to flow directly from the input to the output of a block [4]. Resnet achieved state-of-the-art performance in image recognition tests and won that ImageNet large scale visual recognition challenge in 2015 and 2016.

IV. Training proceedures and Experiements

These experiments were deployed on Google Colab using a Pytorch code base which implemented the use of GPUs to train and test on the fear 2013 data set.

A. AlexNet vs ResNet-34 (Batch size 256, Lr=.005, 50 epochs)

This experiment aimed to compare the performance between two convolutional neural network models, AlexNet and ResNet, in terms of accuracy, training loss, and their confusion matrix properties. To achieve this goal, both models were trained in parallel, and their training progress was closely monitored and analyzed.

After analyzing the results, it was observed that ResNet quickly attained near-full accuracy within the first 15 epochs of training. This indicates that ResNet is a powerful neural network model that can achieve optimal results with minimal training time. In contrast, the training pattern of AlexNet was erratic, with varying levels of accuracy achieved over the course of its training. Please see Fig 4 and Fig 5 for details.

The comparison of the two models' training patterns highlights the fact that the ResNet architecture is more robust and can perform better in complex tasks that require deep learning. This is due to ResNet's unique residual block architecture that enables it to train deeper neural networks without encountering the problem of vanishing gradients. On the other hand, AlexNet's training pattern indicates that it may require careful tuning of its hyperparameters to achieve optimal performance.

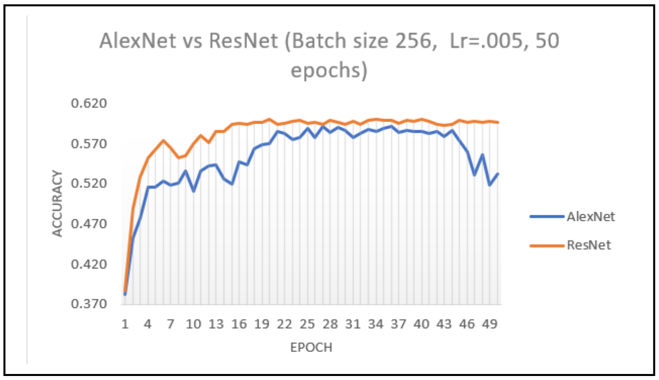


Fig 3. AlexNet vs ResNet Accuracy – Batch 256

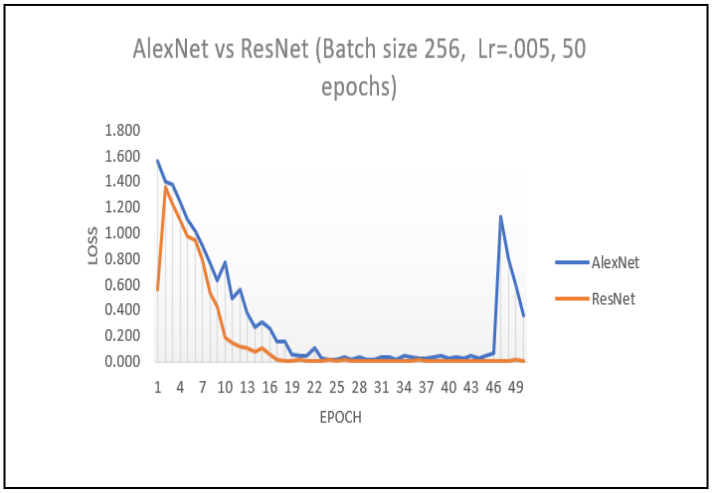
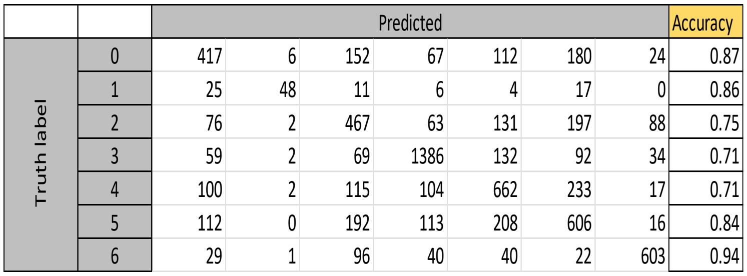


Fig 4. AlexNet vs ResNet Loss – Batch 256

Figure 5 and Figure 6 show the representative confusion matrices for AlexNet and ResNet respectively. These matrices provide a detailed view of the interclass performance of the models. Figure 5 shown below provides class reference for the confusion tables.

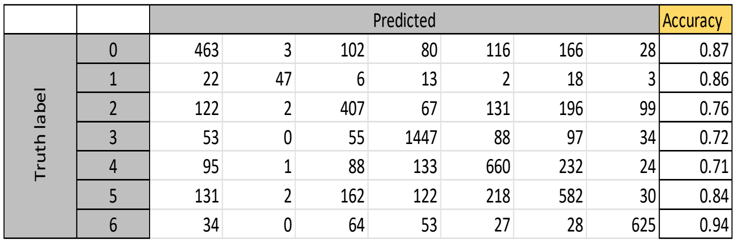
Class						
0	1	2	3	4	5	6
angry	disgust	fear	happy	neutral	sad	surprise

Fig 5. Class Reference



		Predicted							Accuracy
Truth label	0	417	6	152	67	112	180	24	0.87
	1	25	48	11	6	4	17	0	0.86
	2	76	2	467	63	131	197	88	0.75
	3	59	2	69	1386	132	92	34	0.71
	4	100	2	115	104	662	233	17	0.71
	5	112	0	192	113	208	606	16	0.84
	6	29	1	96	40	40	22	603	0.94

Fig 5. AlexNet Confusion Matrix – Batch 256



		Predicted							Accuracy
Truth label	0	463	3	102	80	116	166	28	0.87
	1	22	47	6	13	2	18	3	0.86
	2	122	2	407	67	131	196	99	0.76
	3	53	0	55	1447	88	97	34	0.72
	4	95	1	88	133	660	232	24	0.71
	5	131	2	162	122	218	582	30	0.84
	6	34	0	64	53	27	28	625	0.94

Fig 6. ResNet Confusion Matrix – Batch 256

The total accuracy for the ResNet architecture was 59.75% while the total accuracy for the AlexNet architecture was 58.35%. with these two architectures when training during this test program resulted in similar performance, however, as discussed above the resonant architecture attains near full accuracy within the first 15 epochs.

B. *AlexNet vs ResNet (Batch size 128, Lr=.005, 50 epochs)*

This experiment aimed to investigate the effect of batch size on training performance of two neural networks. To maintain consistency in the experiment the architectures of both models were held constant. ResNet was trained for only 35 epochs due to the length of training time required per epoch.

Batch size is an important parameter in deep learning. By keeping the architectures and learning rate the same the experiment aimed to isolate the effect of batch size on the training performance between the two models. As can be seen in figure seven and eight below, the smaller batch size had a negligible effect on the ResNet training and accuracy. Resonant architecture again attained near full accuracy within the first 15 epochs whereas the Alex net architecture repeatedly struggled to obtain similar performance over 50 epochs.

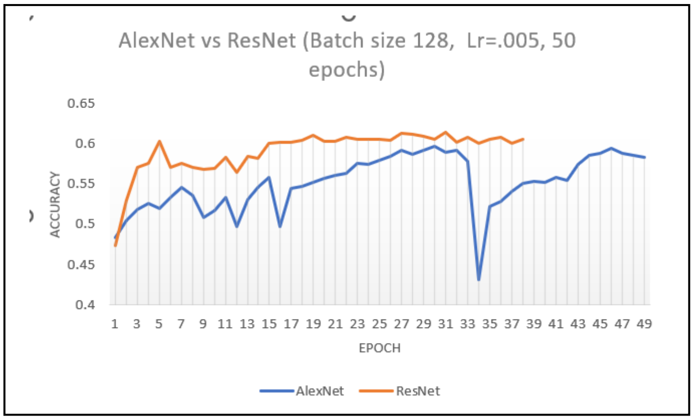


Fig 7. AlexNet vs ResNet Accuracy – Batch 128

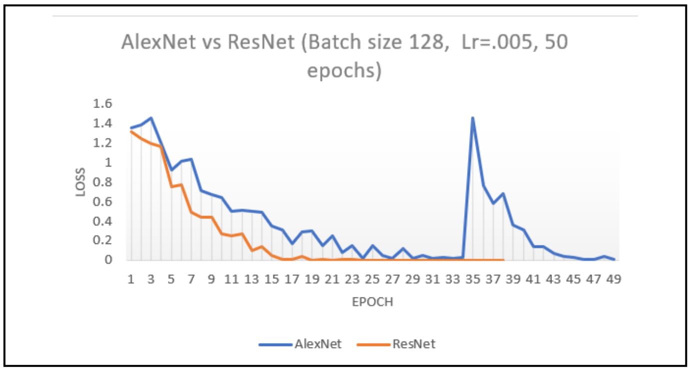


Fig 8. AlexNet vs ResNet loss– Batch 128

The final accuracy performance of the ResNet and AlexNet architectures was compared in the experiment, with ResNet achieving an accuracy of 60.36% and AlexNet achieving an accuracy of 58.63%. Figure 9 and Figure 10 show the representative confusion matrices for ResNet and AlexNet, respectively. These matrices provide a detailed view of the interclass performance of the models.

		Predicted							Accuracy
Truth label	0	466	4	85	74	123	178	28	0.86
	1	26	47	11	3	2	18	4	0.86
	2	122	2	436	55	102	204	103	0.76
	3	60	2	50	1462	85	87	28	0.73
	4	97	5	83	126	662	234	26	0.72
	5	151	4	142	99	238	596	17	0.84
	6	30	0	77	48	29	21	626	0.94

Fig 9. ResNet Confusion Matrix– Batch 128

		Predicted							Accuracy
Truth label	0	498	4	82	84	127	144	19	0.86
	1	32	54	7	5	4	8	1	0.86
	2	121	2	408	69	144	202	78	0.77
	3	66	1	47	1414	121	87	38	0.72
	4	118	2	75	102	714	209	13	0.71
	5	152	3	144	100	265	559	24	0.85
	6	34	1	65	47	47	25	612	0.95

Fig 10. AlexNet Confusion Matrix– Batch 128

C. *AlexNet vs ResNet (Batch size 128, Lr=.007, 20 epochs)*

This experiment was designed to examine the impact of the learning rate on the performance of two neural networks. The learning rate is a crucial parameter and a field of deep learning as it determines the magnitude of adjustments to the weights during the training process. If the learning rate is too high or too low the weights will update sub optimally leading to training convergence problems. Shown below in figures 11 and 12 are the performance characteristics of the AlexNet architecture in conjunction with the ResNet architecture. Figure 11 shows the accuracy of both neural networks and figure 12 shows the training loss of both neural networks.

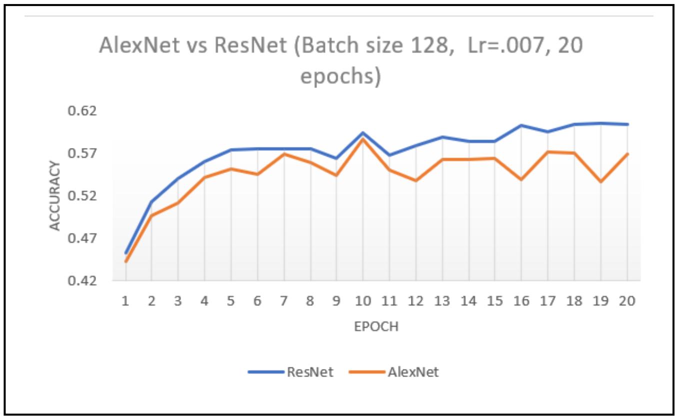


Fig 11. AlexNet vs ResNet Accuracy– Batch 128

Fig 12. AlexNet vs ResNet Training Loss– Batch 128

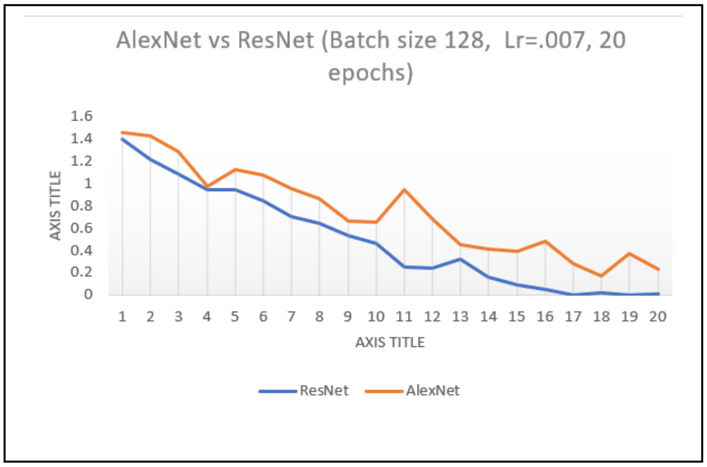
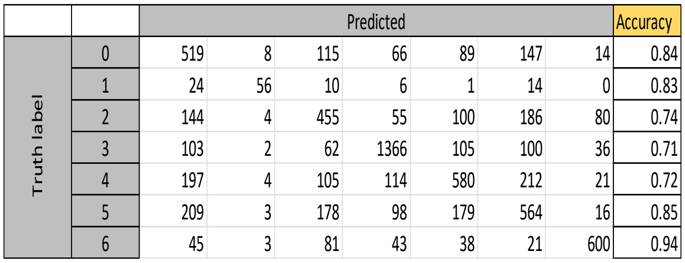
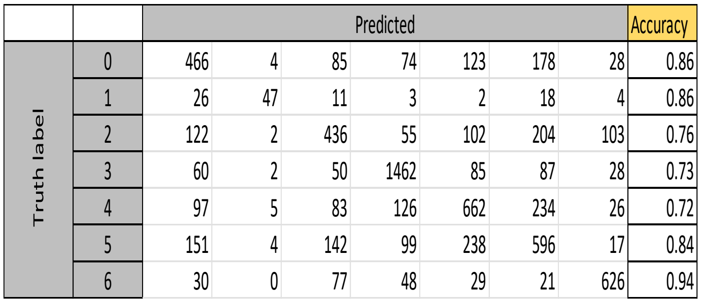


Figure 13 and Figure 14 show the representative confusion matrices for AlexNet and ResNet, respectively. These matrices provide a detailed view of the interclass performance of the models. Figure 5 shown below provides class reference for the confusion tables.



		Predicted							Accuracy
Truth label	0	519	8	115	66	89	147	14	0.84
	1	24	56	10	6	1	14	0	0.83
	2	144	4	455	55	100	186	80	0.74
	3	103	2	62	1366	105	100	36	0.71
	4	197	4	105	114	580	212	21	0.72
	5	209	3	178	98	179	564	16	0.85
	6	45	3	81	43	38	21	600	0.94

Fig 13. AlexNet Confusion Matrix– Batch 128



		Predicted							Accuracy
Truth label	0	466	4	85	74	123	178	28	0.86
	1	26	47	11	3	2	18	4	0.86
	2	122	2	436	55	102	204	103	0.76
	3	60	2	50	1462	85	87	28	0.73
	4	97	5	83	126	662	234	26	0.72
	5	151	4	142	99	238	596	17	0.84
	6	30	0	77	48	29	21	626	0.94

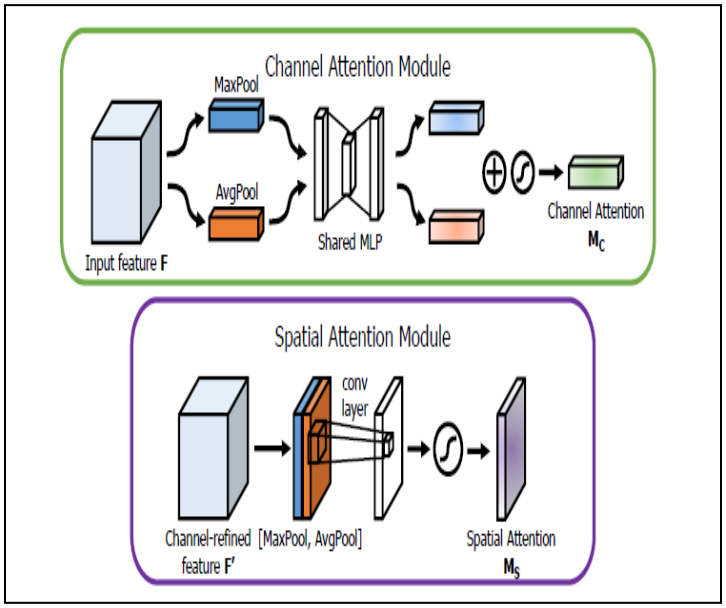
Fig 14. ResNet Confusion Matrix– Batch 128

The total accuracy for the ResNet architecture was 60.93% while the total accuracy for the AlexNet architecture was 57.67%. with these two architectures when training during this test program resulted in similar performance, however, as discussed above the ResNet architecture attains near full accuracy within the first fifteen epochs.

V. Additional Architecture research

Convolutional block attention modules (CBAM) are a type of lightweight attention mechanism used to improve the performance of a convolutional neural network. Leaning on the modular features of successful networks such as ResNet and ResNeXt, the CBAM aims to increase accuracy and efficient efficiency during training by helping information flow within the network. The CBAM module contains two components that can be used either in series or parallel fashion. The first component is called the channel attention module and utilizes a one dimensional attention map that incorporates average pooled and Max pooled features simultaneously. The aim of the channel attention module is to exploit the relationship of features across channels in the image data. The second component is called the special attention module, it generates two-dimensional maps representing average and Max pooled features across a single channel. The aim of the special attention module is to highlight information regions in the images. Figure 15 below depicts the architecture of a convolutional block attention module.

Fig 15. Convolutional Block attention Module (CBAM)



The architecture from the convolutional block attention module can be easily attached to state-of-the-art networks such as resNet by adding a CBAM model after each residual block layer. As a final comparison the C band model was added to the resnet architecture used in this paper, specifically the resnet dash 34 architecture. Due to GPU memory constraints the CBAM-resnet 34 architecture could only be trained on bat sizes less than 256. Therefore, the bat size chosen for the CBAM model were 128 and 64. In figures 16 and 17, the comparative performance of the cbam architecture to Alex net and resnet are shown below. As can be seen in the figures the C bam architecture seems to have a slower training performance for a comparatively similar that size of 128 images.

Fig 16. CBAM vs ResNet and AlexNet - Accuracy

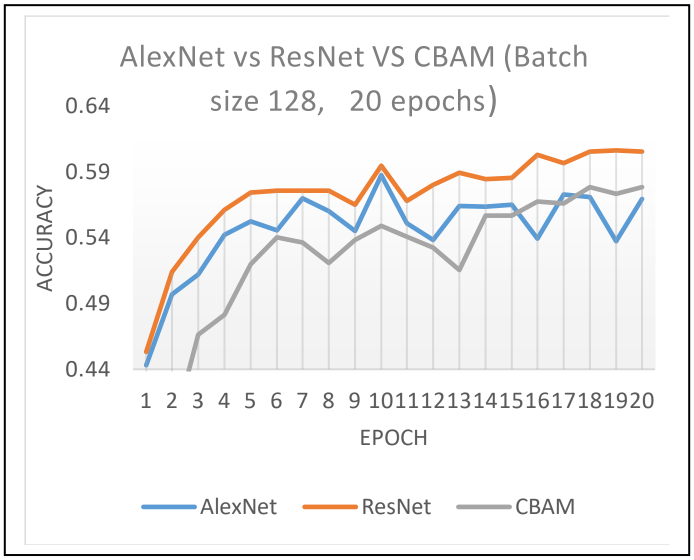
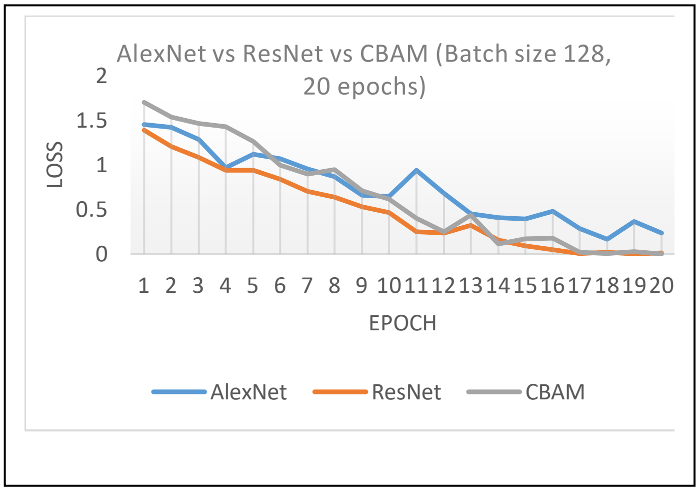
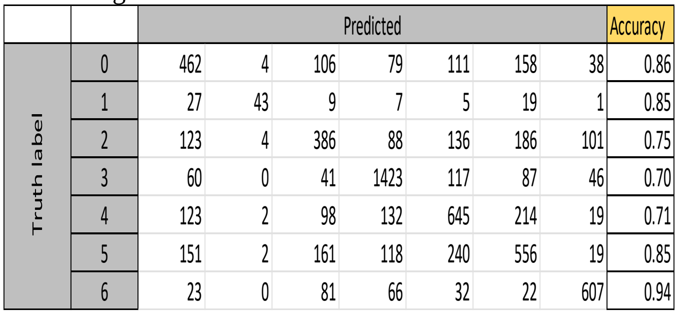


Fig 17. CBAM vs ResNet and AlexNet - Loss



Shown in figure 18 below, the confusion matrix was used to evaluate the inner class performance of the C band modified resnet 34 architecture on the fear of 2013 data set. As can be seen in the table the performance of the C bam architecture is in line with the resnet and Alex net architecture as shown previously in this paper.

Fig 18. CBAM vs ResNet and AlexNEt – Loss



		Predicted							Accuracy
Truth label	0	462	4	106	79	111	158	38	0.86
	1	27	43	9	7	5	19	1	0.85
	2	123	4	386	88	136	186	101	0.75
	3	60	0	41	1423	117	87	46	0.70
	4	123	2	98	132	645	214	19	0.71
	5	151	2	161	118	240	556	19	0.85
	6	23	0	81	66	32	22	607	0.94

Finally, for a longer training of 50 epochs, several runs were created to try to prarmeter search the best performing CBAM model on the FER-2013 dataset. The modified ResNet architecture performed in-line with the AlexNet and ResNet architecture on longer training runs but slight slightly worse the pure ResNet architecture when training is limited.. In figure 19 the training performance is shown against ResNet and AlexNet.

Figure 19 – CBAM vs Alexnet Vs ResNet

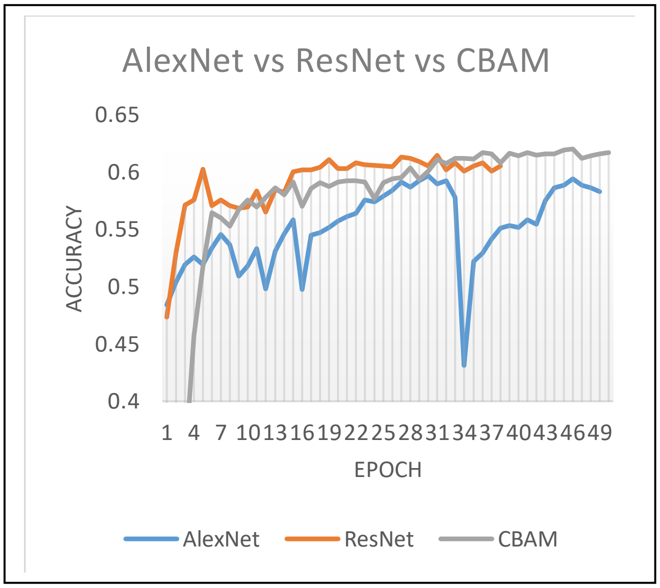
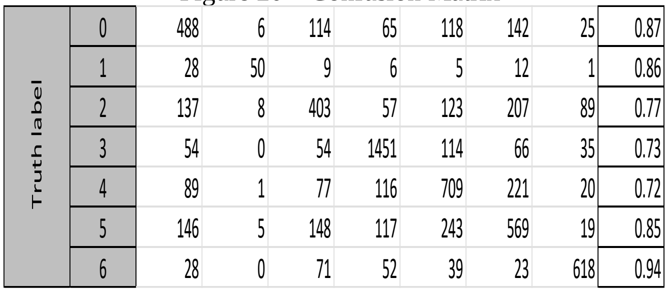


Figure 20 – Confusion Matrix

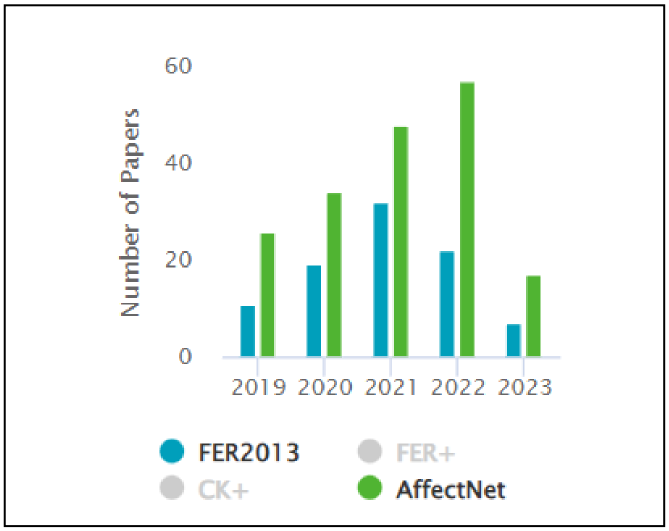


Truth label	0	488	6	114	65	118	142	25	0.87
	1	28	50	9	6	5	12	1	0.86
	2	137	8	403	57	123	207	89	0.77
	3	54	0	54	1451	114	66	35	0.73
	4	89	1	77	116	709	221	20	0.72
	5	146	5	148	117	243	569	19	0.85
	6	28	0	71	52	39	23	618	0.94

VI. CONLUSION

The findings of this research paper indicate that Alex net, resnet and the CBAM modified resnet have relatively similar performance on the FER-2013 data set. The accuracy measures and confusion matrix tables show that no significant differences in terms of facial expression recognition we're seen. However, the loss and accuracy curves reveal that the CBAM modified ResNet does have a slower training performance compared to other models. While the FER-2013 data set appear simple on the surface, the work here shows that the data has subtleties that lend to lower performance of these algorithms. Facial expression recognition Is it complex task due to the fact that subtle movements in the face can convey conflicting signals to an observer. This complexity coupled with images taken at an angle likely leads to lower performance as shown in this paper. Finally it seems researchers have moved on to a newer data set called AffectNet. This data set contains one million images as compared to the FER-2013 data set which only contains on average 30,000 images. Even with the expanded data the researcher in [5] suggests the analysis of human facial behavior is a complex problem. Even human annotators on the affect net project only agreed on 60% of the categories of facial expressions.

Fig 19. FER-2013 vs AffectNet



REFERENCES

[1] R Vamshi N and B Raja S “Facial Expression Recognition using Deep Learning,” Vellore Institute Of Technology

[2] I Goodfellow, A Courville, D Erhan, D. Lee Challenges in Representation Learning: A Report on Three Machine Learning Contests, Article in Neural networks: the official journal of the International Neural Network Society · July 2013.

[3] A Krizhevsky, I Sutskever, G Hinton, ImageNet Classification with Deep Convolutional Neural Networks, University of Toronto.

[4] K He, X Zhang, S Deep, J Sun, Residual Learning for Image Recognition, Microsoft Research

[5] A. Mollahosseini, B. Hasani, M.Mahoor, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild