Analyzing CNNs for Facial Emotion Recognition

Tejas Patel
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
tp3381@cs.rit.edu

***Abstract*—As the world's computers become more competent and innovative, it is increasingly more important for computers to understand and predict human behaviors and emotions. Facial Emotion Recognition(FER) has been researched ongoing for decades. Still, it shows little understanding of facial expressions when a small obstruction or a profile view obscures the entire face in an image. The rise of CNNs and transfer learning makes it easier to work with the aforementioned challenges as opposed to landmark-based studies. This paper studies different advancements in the field of FER. It proposes a new model dubbed VGG-FER to address the problem of occlusion via augmenting synthetic occlusion during its training phase. VGG-FER is also optimized for the popular FER2013 dataset using a learning rate scheduler and more regularization via dropouts. It achieves a best single model accuracy of 72.39% with a top 2 accuracy of 85.9%. Another part of the study compares different transfer learning models against the variants of VGG-FER and clubbed together to achieve an accuracy of 74.98%, and the top 2 accuracy goes up to 88.13%.**

***Index Terms*—Computer Vision; Deep Learning; Occlusion; Facial Emotion Recognition**

I. Introduction

Predicting facial expressions has always been a challenging subset in sentiment analysis. Facial Emotion Recognition looks to classify these expressions into six basic emotions (angry, disgust, sad, happy, fear, and surprise) and one neutral emotion. Detecting facial expressions can be highly accurate in a lab environment with up to 99% accuracy [1]. However, these frameworks fall short in real- world scenarios, due to challenges such as profile views, occlusion, and illumination differences, are likely to occur. Occlusion is the presence of another object in front of the image, resulting in only a partial face visible to the classifier.

This paper classifies facial emotion recognition on the FER-2013 data set [2], more suited to a real-world environment. It contains multiple occlusion, profile view, illumination, and label noise instances, unlike an in-lab environment. Another factor that makes FER-2013 more practical is its low-resolution grayscale images, similar to what a surveillance camera feed provides. Previous studies on the data set have yielded accuracies between 65-75% using Convolutional Neural Networks (CNNs) to achieve these results [3], [4], [5].. These studies utilize popular techniques such as transfer learning and creating an ensemble of multiple models to achieve these accuracies.

The proposed VGG-FER model architecture(s) used for training draw inspiration from the VGGNet family of CNNs, which are easy to understand and have shown to classify a thousand objects in the Image Net challenge with slight error [6]. Like most deep learning models, the best hyperparameters to maximize test data accuracy get determined empirically through trial and error. Another critical area this paper explores is whether shallow CNNs can perform FER just as well, helping reduce computation power. Owing to the limited training data size, the addition of auxiliary data enriches the model's generalization capabilities further. Finally, after training this model, a saliency map is generated better to understand which parts of the image or the face the CNN model gives more attention to when making its predictions.

Finally, to test the model's robustness, we make use of two other data sets collected in a laboratory environment with posed expressions: JAFFE, CK+ [7], [8] and another dataset containing real-world images: AffectNet[9]. A live feed from a webcam can show how accurately the classifier can detect emotions in a real-world scenario under different conditions. By being able to analyze expressions better, the next generation of robots and intelligent gadgets should understand humans more effectively. The current best applications of FER can be seen in public safety and for diagnosing mental disorders.

II. Related Works

FER is a more significant part of the affect recognition line of studies. Marechal et al.[10] review various studies across different modalities, including sound, image, video, gestures, and physiological signals such as heartbeat and pulse rate. Working with multi-modal data such as sound and video has also been proven to recognize emotions effectively. The practical applications of affect recognition are seen in the crime and healthcare industries primarily, with applications such as lie detection and detecting the onset of mental disabilities.

Studying facial expressions has been a keen research area since the 1960s. All the research back then was based on identifying landmarks on a human face and encoding an expression as a combination of these landmarks. Paul Ekman and Friesen[11] pioneered most of the landmark-based research in this time and released the Facial Action Coding

System (FACS) after more than 14 years of research. It consisted of marking landmarks as Action Units (AUs), and by combining two or more AUs, different facial expressions can be classified. For example, as per the FACS, happiness is a combination of AU6(cheek raiser) and AU12(lip corner puller).

However, the biggest drawback with landmark-based studies is being able to perform FER in the absence of a landmark, either due to occlusion or a profile view of the face. This factor, coupled with the advent of CNNs, migrated the research to using Deep Learning frameworks for FER with more competitive accuracy. However, the problem arises because most facial expression datasets are taken in a lab environment and generally only consist of frontal face images. CNNs work best when a sizeable computation power supplements many training data. Hence, having enough training data with instances of occlusion and profile views is imperative for making more robust classifiers.

Over the last eight years, there have been many advancements in FER using CNNs. Tang et al. [5] used encodings from a CNN model on a linear SVM classifier to much success in the ICML 2013 challenge by Prof. Ian Goodfellow [2]. Mehendale [12] utilizes two CNN models, the first one to perform face detection based on skin tones, used to generate an expressional vector of different face landmarks by using nearest cluster mapping and edge detection. This expressional vector is then passed down to the second part CNN, a series of simple convolution filters, and a final perceptron layer to make its prediction.

Feng and Chaspari[13] cover the use of transfer learning models in the field of Emotion Recognition. It outlines critical factors to consider before applying transfer learning from a source to a target dataset: the method of collection of data and the number of classes in source and target datasets, and the environment of respective data (in lab/real world). In the field of images and video, most prior works opt for deep CNNs such as ResNet and VGG. Akhand et al. [1] use transfer learning on in-lab datasets, JAFFE and KDEF to much success; instead of fine-tuning the entire pre-trained model in one go, this paper gradually fine-tunes each block over incremental steps of training. They obtain their best accuracy from the DenseNet model, which is 161 layers deep and achieves an accuracy of 96.51% on the KDEF dataset and 99.52% on the JAFFE dataset with 10-fold cross-validation.

Jaiswal and Nandi[14] move away from the transfer learning school of thought and advocate for a simpler model with fewer parameters. The paper introduces a model inspired by the inception net architecture, which is less deep but broader, which still achieves a competitive accuracy on these models. Pramerdorfer et al. [4] train eight custom models derived from the VGG, ResNet, and InceptionNet architectures and ensemble them together to achieve an accuracy of 75.2%.

Lee et al. [15] gives a refreshing take on the problem of Facial Emotion Recognition, where it considers the problem from a continuous aspect as opposed to a discrete classification problem. This paper deals with modeling human emotions using a dimensional method, representing continuous emotional states, and then applying linear regression to quantify human emotions as valence and arousal values.

Finally, Khaireddin and Chen[3] demonstrate the effectiveness of a better pre-processing pipeline and rigorous hyper-parameter tuning to achieve the best single model accuracy of 73.28%. This paper uses the standard augmentation techniques of flipping, rotation, and translation for pre-processing. Using another augmentation technique: randomized crop, the image is cropped at random and then resized to 40x40. The SGD optimizer and Cosine Annealing [16] as a learning rate (LR) schedule help eke out the best accuracy on the FER-2013 for the model after comparing various other optimizers and LR schedulers.

III. Data set and Model Architecture

A. Data set

The data set used for this study is FER-2013 which was released as a competition by Goodfellow et al. [2]. There were 28701 training images and 3589 validation set images released initially, after which the test set was used to evaluate the shortlisted candidates with the best validation accuracy. The competition's winner achieved a 71.11% test accuracy, which is the benchmark this study aims to match or beat.

FER-2013 consists of greyscale images of size 48x48 labeled into seven different emotions: anger; disgust; fear; happiness; sad; surprise, and neutral. There is a significant class imbalance observed as the disgust class only consists of 436 images instead of the majority class happiness containing 7215 images. The entire class distribution can be seen in Fig. 2. The images were carefully curated from various sources such as snapshots of movies, candid expressions, and label noise in the form of emojis and drawings for better regularization. The human accuracy on this dataset has been reported to be around 65% [2].

B. Model Architecture

The proposed model VGG-FER Net is based on the VGG-Net family of CNN architectures proposed by Simonyan and Zisserman [6] in the 2014 ILSVR Challenge. The VGGNet model achieves a 71.30% accuracy on the ImageNet dataset and uses 138 million parameters or computations in total. It consists of 2 layers of 3x3 convolution and a 2x2 maxpool layer in each processsiong block of its architecture.

The proposed model, as seen in Fig. 1 has two layers of convolution in each block, followed by relu activation and Batch Normalization to speed up model training [17] .

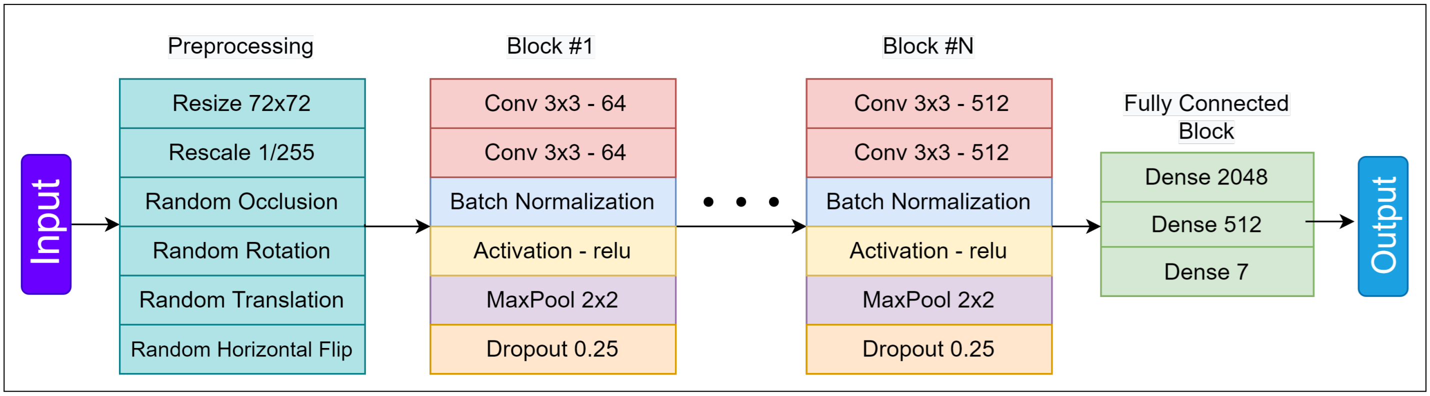

Fig. 1. VGG-FER Architecture

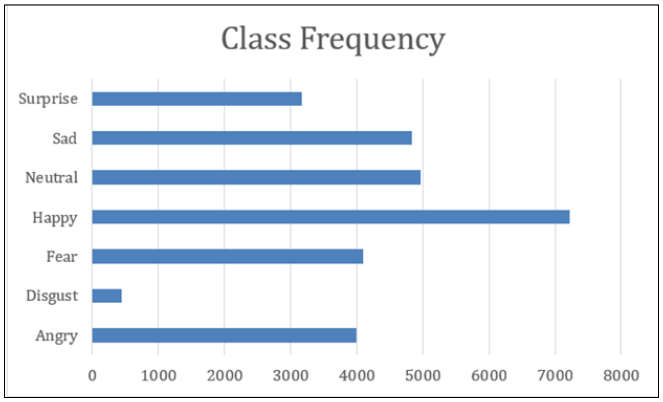

Fig. 2. Data Imbalance in FER-2013

Finally, a maxpool layer of varying kernel sizes per different model variations reduces the dimensions for the next layer by taking only the maximum activations. After maxpooling, a dropout layer is utilized to drop 25% of the inputs before feeding it to the following processing block. The dropout layer [18] stays in effect only during the training phase.

The proposed model also performs its own preprocessing. The image is resized and re-scaled before numerous data augmentations are applied during the training phase, including the custom synthetic occlusion. After the preprocessing and convolutions, the inputs are flattened to a single dimension and passed through two dense layers with relu activation and a final softmax layer to get the class probabilities or predictions.

IV. Methodology and Experiments

A. Baseline model

For initial experimentation, a baseline model in a vanilla CNN is used with a CONV-POOL-Dense block to make model predictions on raw data with no pre-processing pipeline placed. This baseline model is used to form a general idea of what techniques help improve model accuracy and generalization and what techniques are deterrent to the same.

For transfer learning models, the baseline is formed using predictions made on image net weights, classified into one of the seven emotions through a softmax dense layer attached at the end. The general idea here is to find the efficacy of how suited a pre-trained architecture trained over a different domain adapts to the domain of FER.

B. Upsampling images

To treat class imbalance, as mentioned in the previous section, Random oversampling provides more instances of the minority classes in the image. Other techniques like SMOTE and undersampling were also experimented with within this paper. Random oversampling adds duplicate instances of minority images to the training data so that the model learns better on classifying these images. Since augmentation is performed at every training epoch, the model sees a slightly altered image at every training epoch and learns to generalize itself better on less training data.

The only caveat to this approach is a higher chance of overfitting minority classes. However, a CNN model requires more and more training data to work well, and hence other techniques such as undersampling and no sampling prove less beneficial to model performance.

C. Pre-processing and Augmentation

As seen in Fig. 1, all pre-processing is done within the model's architecture itself. The first step is to resize the image to 72x72 from the initial 48x48. This step is crucial as the image gets to train on more features during training and makes it easier to work with unstructured image dimensions during the evaluation phase. Following this step, all images are re-scaled to a range of 0-1 from 0-255, as GPUs work can make fast computations on floating-point data.

During training, since the data size is insufficient and has a low percentage of images with occlusion, image augmentation is utilized to enable the model to generalize better. A novel augmentation technique in the form of synthetic occlusion is introduced in this paper. The proposed RandomOcclusion layer introduces occlusion in the form of a black box with a random center, length, and breadth at a 50% chance.

Synthetic occlusion ensures the model does not grow overly reliant on the presence of a particular feature in the image. Furthermore, horizontal flipping, random rotation of up to ±20 degrees, and random translation of up to ±10% are applied to the training images. After pre-processing the entire set of images, the resultant batch of images is normalized to have a mean of 0 and a standard deviation of 1, which aids in standardizing the inputs for subsequent layers.

D. Training

Model training was done on three different model variants: VGG-FER_v4, VGG-FER_v5, and VGG-FER_v6. Apart from this, four popular image-net trained architectures were fine-tuned for model comparison and to assemble results. Grid searches were performed to find the best hyperparameters for the VGG-FER_v4 model, such as the learning rate, batch sizes, and dropout rates.

During training, a custom learning rate scheduler based on Cosine Annealing was used for model training and the early stopping callback. Cosine Annealing is a learning rate scheduling technique where the learning rate is decreased gradually, then spiked up at a longer interval to help the model reach global minima. The spikes are known as warm restarts as the model effectively starts retraining with learned weights, thereby escaping the local minima and converging to a better minimum.

Another addition is the early stopping callback, which monitors model checkpoints over a specified number of epochs and stops training when no improvement is observed over the determined period. It also restores the model to the checkpoint with the best validation accuracy, significantly decreasing the likelihood of an overfit. Fig. 3 shows the decay and spikes in the learning rate during training with an initial learning rate of 1.

E. Proposed model Variants

The proposed model, as seen in Fig. 1 is flexible in terms of how deep it can be. This paper performs training on the proposed architecture's four-block, five-block, and six-block variants. These models are termed as VGG-FER_v4, VGG-FER_v5, and VGG-FER_v6, respectively. Here, the idea is that a deeper model is more equipped to make non-linear relations to aid its classification. At the same time, a deeper model requires more training data to show a visible improvement in accuracy and consequently needs more computation power and time to train.

All the variants can get a competitive accuracy on the dataset while using lesser parameters than the pre-trained ImageNet architectures. The five-block model achieves an accuracy of 72.39%, the best of any models trained for this study. The proposed architecture works on greyscale images(1 channel) instead of RGB (3 channel) images required by pre-trained architectures, which gives it more relevance to the FER2013 dataset and makes it simpler at the same time.

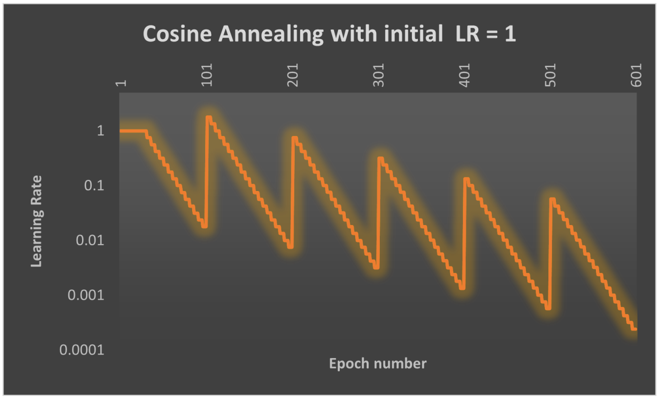

Fig. 3. Cosine Annealing

F. Pre-trained architectures

Transfer learning is a valuable technique for building highly efficient image classification models. The basic idea is to take a model trained on a similar classification domain and retrain its weights to adapt to our classification problem. Some of the most popular architectures for transfer learning are VGG16, VGG19 [6], ResNet50 [19], and InceptionNet_v3 [20]. After being trained on the ImageNet Large Scale Visual Recognition Challenge, these architectures have weights that classify 14,197,122 images into one of 1000 classes. Hence, the weights from these models can be easily adapted to classify seven emotions.

G. Assembling different architectures

Throughout its training, with any deep learning architecture, certain biases kick in that make it impossible to correct itself. The final layer of these architectures generally predicts a list of probabilities for each class which adds up to one. By summing these probabilities from each model, as seen in fig **??**, we can assemble a new classifier that mitigates the said biases of a single architecture. This acts as a jury and serves to classify with more confidence instead of using a single architecture.

V. Results

A. Dropout rates

Dropout layers are present at the end of each convolution processing block in the model, as seen in Fig 1, and are in effect only during the training phase of the model. Stripping the data with a random percentage of inputs has a regularizing effect on the model's training as it learns to do more with fewer data. The ideal dropout was determined empirically for the model. A dropout rate of 0.25 is the most ideal for VGG-FER_v5.

TABLE I
VGG-FER_v5 TRAINED FOR DIFFERENT DROPOUT RATES

Dropout Rate	0.0	0.1	0.2	0.25	0.3	0.4	0.5
Accuracy	71.35%	71.47%	71.91%	72.39%	72.04%	71.62%	71.23%

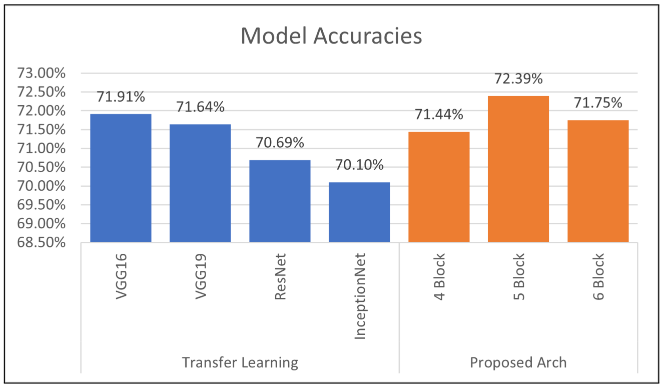

Fig. 4. Accuracy obtained with different architectures

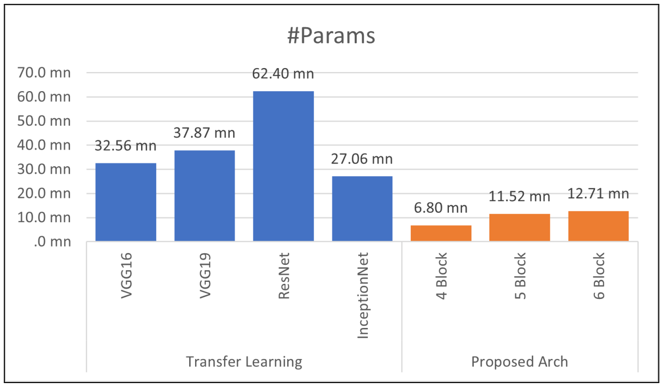

Fig. 5. Total # of parameters for different models

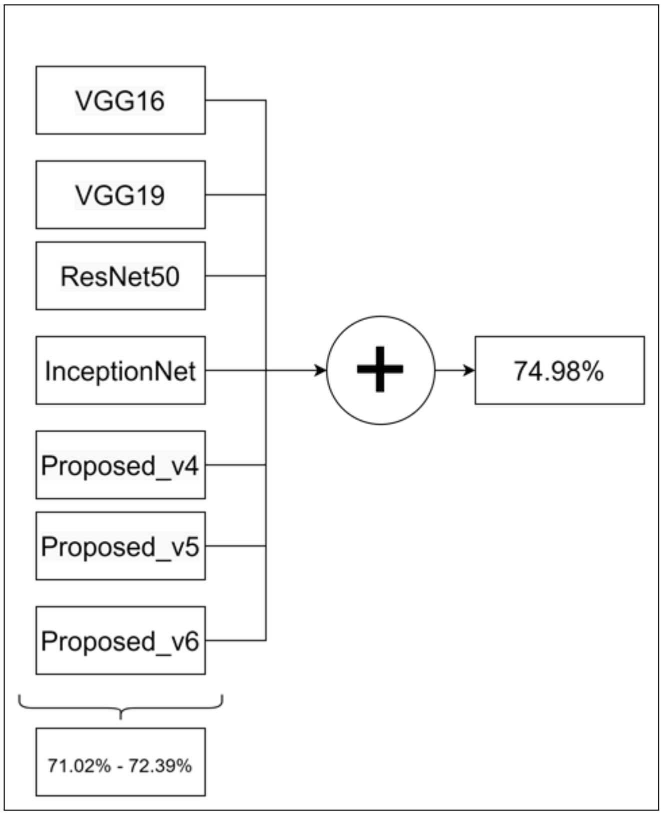

Fig. 6. Assembling 7 different architectures

B. *Single model accuracies*

The accuracies achieved by different variations of VGG-FER and the transfer learning models are listed as shown in Fig 4. The five-block variant of VGG-FER(VGG-FER5̌) achieves the best single model accuracy of 72.39% from all the seven models. Some of the classes have inter-class similarities of more than 15% (sad-neutral or disgust-angry) as seen in Fig 7. Hence top-2 accuracy is proposed to be another way to evaluate the model's efficiency, which sees if the model's second-best guess makes the correct prediction. For VGG-FER5̌, the top-2 accuracy of the model is 85.9% which shows reasonable confidence in the model training correctly.

A vast difference in accuracies is not observed however the pre-trained models fall short of the VGG-FER due to two significant reasons:

- FER-2013 contains greyscale(1 channel) images, whereas pre-trained architectures with ImageNet weights require a 3-channel RGB image to make its predictions. The only workaround to meet the three-channel constraint is to duplicate the greyscale image two more times to get a three-channel image.
- The VGG-FER model is more simple and needs to perform lesser computations. It uses less than 67% parameters on average compared to the pre-trained architectures, as seen in Fig 5.

C. *Assembling model architectures*

The proposed and pre-trained models are assembled to give an accuracy of 74.98%, which is far better than any single architecture. The biggest drawback of assembling seven different models is that the time it takes to make predictions can be much longer, which would be difficult for real-time video data FER, although feasible for image data emotion recognition. The top-2 accuracy of the assembled model is 88.13%, so the model is indeed second-guessing considerably better by using a jury of models.

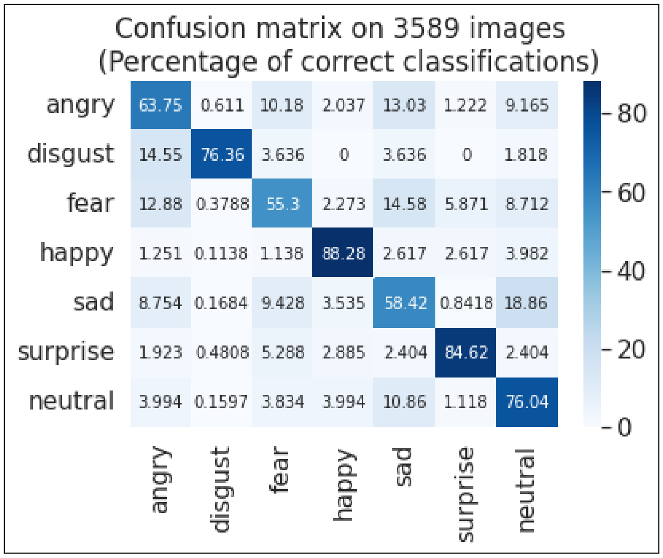

Fig. 7. Confusion matrix of VGG-FER_v5

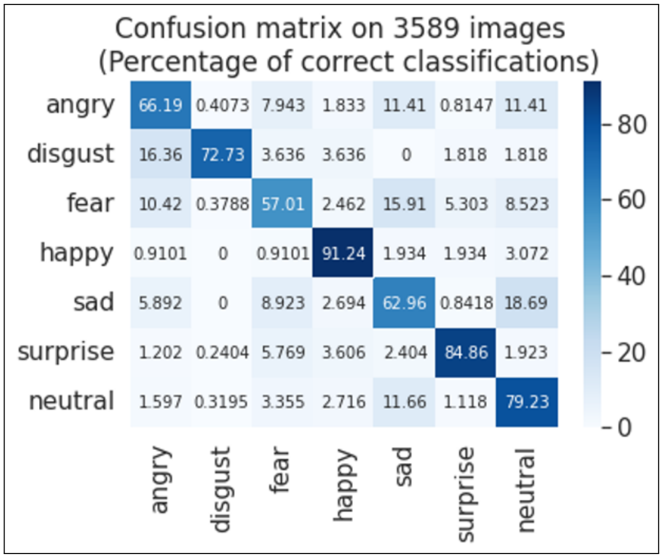

Fig. 8. Confusion matrix of the assembled model

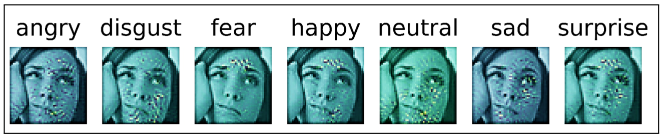

Fig. 9. Saliency map showing activations for different classes

D. Confusion Matrix

A confusion matrix, as pointed out by Lucey et al. [8] is a superior metric to accuracy. It is possible to see inter-class confusion for specific classes and offers more data on what emotions are predicted correctly in more instances. The diagonal values depict the accuracy of each class predicted correctly.

Figure 7 depicts the confusion matrix for VGG-FER_v5, which achieves the best single model accuracy, and figure 8 shows the confusion matrix for the assembled model. The figures show that the assemble model improves the percentage of images classified correctly for each class, except for the disgust class. Also, the assembly of models helps mitigate false negatives during evaluation by tempering the bias creeping into any single model architecture.

E. Saliency maps

Saliency maps help visualize what the model observes when making its predictions for each class. The saliency map takes a maximum of the gradients from the computations at every model layer, obtaining regions of maximum activation for each class and superimposes it on the original image. In Fig 9, just like humans, VGG-FER concentrates on certain parts of the face to make its predictions, where it completely disregards the hair and focuses on regions where the cheeks, brows, and the mouth are present predominantly.

From these maps, we can see exactly what parts of the face the model focuses on when determining each class. As seen in Fig 9, almost half of the face is occluded. The model does well in this scenario to still be able to find the left cheek and focuses a lot more on the eyebrows or the chin area for different classes. Fig 9 shows that the model can generalize well even when it deals with occlusion and validates the use of adding random occlusions during training has helped the model detect occlusion.

F. Testing the model on other datasets

The performance of VGG-FER is trained and evaluated on three other popular FER datasets as shown in Table II. The CK+48 and JAFFE were relatively more minor datasets, and a 10-fold cross-validation was performed over the dataset to get the final model accuracy.

AffectNet [9] is a larger dataset with over 350,000 images classified into eight emotions. However, only seven emotions were considered for the model's study, and the benchmark accuracy compared in Table II is also for the seven emotion training data. Due to computation limitations, this paper had to downsample the training images to only 85,000 and achieved a 62.09% with only a third of the total data.

VI. Future Work

Future improvements on the model can be made in the following areas for VGG-FER and also for research in the

TABLE II
Model performance on other datasets

Data set	#Training Images	VGG-FER_v5 ACC	Benchmark ACC
CK+48	913	98.67%	99.3%
JAFFE	213	94.79%	94.83%
AffectNet	85000	62.09%	65.69%

field of FER:

1) The model training could do with more training data with instances of occlusion and profile views. The current model is trained over only 469 images of the disgust classes and would benefit considerably from including more instances of the minority classes.

2) Some classes show high inter-class similarity, which is akin to human error in understanding facial expressions. Hence, the top-2 accuracy is an excellent metric for benchmarking model evaluations.

3) Use of higher resolution images for training will provide VGG-FER with more features to form non-linear relations, leading to accurate model predictions. Imagenet had images with an average resolution of 469x387 pixels for training in comparision.

4) Due to computing power limitations, only the variants of VGG-FER were tuned for the best hyperparameters through grid searching. The other pre-trained models used for the assembly were only tuned via intuition, which is an area for improvement. Also, including other architectures such as DenseNet might help increase the assemble model accuracy further.

5) The changes introduced in the paper can also be used to derive custom models based on other popular architectures such as ResNet50 and InceptionNet. These models can also provide better accuracies than their pre-trained counterparts and would be simpler again.

AffectNet [9] shows promise in this regard since it contains more training instances and has a higher resolution at 224x224. However, only the training and validation sets have been released since 2017, with the proposed test set held back in hopes of keeping a competition on a later date. There can be a significant leap in the field of FER if a contest does come to fruition.

VII. Conclusion

The proposed paper studies Convolution Neural Networks for Facial Emotion Recognition. It weighs the use of popular transfer learning architectures versus a simpler architecture, and VGG-FER, derived from VGG-Net and introduced for this paper, proves to be the best solution. By incorporating pre-processing in the model architecture, the model deals with a dynamic range of images and performs a novel augmentation technique called random occlusion to deal with the challenge of occlusion in FER. For training, the use of Cosine Annealing for learning rate scheduling and different dropout rates for regularization provide small gains. This paper achieves its best single model accuracy of 72.39% on a five-block variant of VGG-FER, which is also assembled with the other variants and transfer learning models studied to get an accuracy of 74.98%.

The study aims to improve accuracy, although not at the cost of generalization. FER-2013 data set has its own set of challenges such as data imbalance, occlusion and profile views, as it aims to mimic real-world scenarios. This paper comes up with its own model to counteract these challenges, via a new data augmentation technique and dropouts. It shows that the problem is not as complex, and can be dealt better by training a new model as opposed to using transfer learning for FER. Future work can be done to use a bigger data set in terms of resolution and instances of each, and also to derive similar simpler models from other popular architectures.

References

[1] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, “Facial emotion recognition using transfer learning in the deep cnn,” *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[2] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, “Challenges in representation learning: A report on three machine learning contests,” in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[3] Y. Khaireddin and Z. Chen, “Facial emotion recognition: State of the art performance on fer2013,” *arXiv preprint arXiv:2105.03588*, 2021.

[4] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: state of the art,” *arXiv preprint arXiv:1612.02903*, 2016.

[5] Y. Tang, “Deep learning using linear support vector machines,” *arXiv preprint arXiv:1306.0239*, 2013.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[7] M. Lyons, M. Kamachi, and J. Gyoba, “The Japanese Female Facial Expression (JAFFE) Dataset.” Zenodo, Apr. 1998, The images are provided at no cost for non- commercial scientific research only. If you agree to the conditions listed below, you may request access to download. [Online]. Available: https://doi.org/10.5281/zenodo.3451524

[8] 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*. IEEE, 2010, pp. 94–101.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[10] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, “Survey on ai-based multimodal methods for emotion detection.” *High-performance modelling and simulation for big data applications*, vol. 11400, pp. 307–324, 2019.

[11] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.

[12] N. Mehendale, “Facial emotion recognition using convolutional neural networks (ferc),” *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, 2020.

[13] K. Feng and T. Chaspari, “A review of generalizable transfer learning in automatic emotion recognition,” *Frontiers in Computer Science*, vol. 2, p. 9, 2020.

[14] S. Jaiswal and G. C. Nandi, “Robust real-time emotion detection system using cnn architecture,” *Neural Computing and Applications*, vol. 32, no. 15, pp. 11 253–11 262, 2020.

[15] H.-S. Lee and B.-Y. Kang, “Continuous emotion estimation of facial expressions on jaffe and ck+ datasets for human–robot interaction,” *Intelligent service robotics*, vol. 13, no. 1, pp. 15–27, 2020.

[16] I. Loshchilov and F. Hutter, “SGDR: stochastic gradient descent with restarts,” *CoRR*, vol. abs/1608.03983, 2016. [Online]. Available: http://arxiv.org/abs/1608.03983

[17] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

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

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567