Comparative Analysis of Deep Learning Models for Facial Emotional Recognition

Kushagra Srivastava
Computer Science & Engineering
KIET Group of Institutions
Ghaziabad
kushagrathisside@gmail.com

Sanjeev Sharma
Computer Science & Engineering
KIET Group of Institutions
Ghaziabad
martin.mmmec@gmail.com

***Abstract*—Emotional Intelligence and Facial Emotion Recognition are some of the fields that come under high focus when computer vision is finding its way to develop one of the smartest AI solutions. Implementation of custom models often requires elaborate and well-defined processes at all stages of model development i.e. dataset preparation, model training, tuning, and optimization to run them effectively in applications. Thus, the use of pre-trained models comes into play, which can significantly reduce the training/tuning cost and time requirements. Thus, analysis of publicly available pre-trained models contributes significantly to their use for building efficient and faster AI solutions.**

Keywords— Sentiment Analysis, Deep Learning, Pretrained Models, Transfer Learning

I. Introduction

Machine Learning has found its way into many applications starting from basic regression-based problems like price prediction or classification problems which include multiple classes as per the requirement of specific use cases. Deep Learning comes into discussion when machine learning steps into the domain of complex problems that require multiple variables that can't be exactly defined in a specific order. Neural Networks within Machine Learning are studied under this subset referred to as Deep Learning. Neural Networks are classified into many types based on the type of architecture and functions. Each type itself has different applications, most of which require some additional techniques for better results. Hence, a comparative study of the existing pre-trained models makes it less time-consuming to build application-ready models. For a better understanding of the above-mentioned terms in the context of the use case of sentiment analysis, some terminologies are provided below.

A. Machine Learning and Deep Learning

Machine learning (ML) is a subfield of artificial intelligence (AI) that enables computers to learn without being explicitly programmed. It involves algorithms that learn from data and improve their performance over time. Deep learning (DL) is a subset of ML that uses artificial neural networks with multiple hidden layers to learn complex patterns from data. DL has achieved significant success in various applications, including computer vision, natural language processing, and speech recognition.

B. Computer Vision and CNNs

Within the expansive domain of deep learning, computer vision (CV) emerges as a subfield enabling computers to perceive and interpret the visual world akin to humans. It tackles intricate tasks like object detection and image classification, drawing heavily on techniques like convolutional neural networks (CNNs). These specialized architectures excel at processing grid-like data like images, extracting valuable spatial features through their unique layered structure. Their ability to learn hierarchical representations of visual information has revolutionized CV, paving the way for groundbreaking applications like self-driving cars and medical image analysis.

C. Facial Emotion Recognition (FER)

Further exploration leads us to sentiment analysis (SA), a branch of natural language processing (NLP) dedicated to automatically deciphering the emotional undertones woven into text. By classifying opinions, feelings, and attitudes as positive, negative, or neutral, SA offers invaluable insights across various domains, from social media analysis to market research.

Diving deeper still, we encounter facial emotion recognition (FER), a subfield of CV focused on deciphering human emotions from facial expressions. By analyzing features like eyebrows, eyes, and mouth position, FER aims to infer emotional states like happiness or sadness. Potential applications span human-computer interaction and psychological research, offering a deeper understanding of human emotional expression through facial cues.

D. Transfer Learning

Unlocking the full potential of deep learning hinges on the powerful technique of transfer learning (TL). This approach leverages knowledge gained from a pre-trained model on a new, related task. Imagine pre-trained models like VGG or ResNet, initially trained on massive datasets for image classification, being fine-tuned for CV tasks like FER. These models have already learned general-purpose features like edge detection or object recognition, acting as a strong foundation for the target task. Fine-tuning the final layers on the specific facial expression dataset allows the model to adapt to the new task, significantly reducing training time and potentially achieving better results compared to building a model from scratch.

E. Pretrained Models in Computer Vision

Commonly used pre-trained models include the following models:

A. AlexNet[2]: Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton won the ImageNet Large Scale Visual Recognition Challenge in 2012 with a test accuracy of 84.6% using this model. This model significantly outperformed the second runner-up (top-5 error of 16% compared to runner-up with 26% error). This was one of the research which had over 60 million parameters and used the 'relu' function.

B. VGG 16[3]: This model comes from the paper Very Deep Convolutional Networks for Large-Scale Image Recognition (ICLR 2015). This model achieved 92.7% (top

5) test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014.

C. VGG 19: VGG-19 is a convolutional neural network that is 19 layers deep and can classify images into 1000 object categories such as a keyboard, mouse, and many animals. The model trained on more than a million images from the ImageNet database with an accuracy of 92%. It is an updated version of VGG 16.

D. DenseNet:

The name “DenseNet” refers to Densely Connected Convolutional Networks[7] developed by Gao Huang, Zhuang Liu, and their team in 2017 at the CVPR Conference. It received the best paper award and has accrued over 2000 citations. Traditional convolutional networks with n layers have n connections but DensetNet has n(n+1)/2 connections in total because of feed-forward fashion.

E. EfficientNet

EfficientNet introduces a compound scaling method that simultaneously scales depth, width, and resolution using carefully chosen coefficients. This systematic approach ensures optimal resource utilization while maintaining high accuracy. EfficientNet achieves state-of-the-art accuracy on ImageNet and other benchmarks while requiring significantly less computation and memory compared to VGG-19 and DenseNet. This efficiency makes it well-suited for various applications, including mobile and embedded systems, real-time tasks, and scenarios where resource limitations are a concern.

F. Performance of the State-of-the-Art Models

The term "state-of-the-art (SOTA)" refers to the best-performing models or techniques in a specific field, and it's constantly evolving as research progresses. SOTA’s are usually specific to their area of application and differ based on metrics like speed and accuracy.

For FER-like tasks, while custom models offer exploration opportunities, established deep learning architectures specifically designed for FER consistently achieve superior performance. These models, meticulously crafted and rigorously optimized on extensive datasets, leverage techniques like transfer learning and pre-trained weights, granting them a significant edge in accuracy and generalizability. Thus, fine-tuning these models provides better inputs.

II. Literature Reviews

A. Literature Review from papers

The paper "Gradient-based learning applied to document recognition" by LeCun et al. (1998)[1] stands as a landmark in the history of deep learning, specifically for convolutional neural networks (CNNs). This work introduced a novel architecture inspired by the visual processing hierarchy of the brain, utilizing stacked convolutional layers to extract spatial features from image data. The authors demonstrated the effectiveness of CNNs in handwritten digit recognition, achieving superior performance compared to traditional methods.

Yu et al. (2011) were among the first to explore deep learning for FER, proposing a Convolutional Neural Network (CNN) architecture that significantly outperformed traditional methods based on handcrafted features like Ekman and Friesen's (1978) Facial Action Coding System (FACS). This work paved the way for further research into CNN-based architectures for FER, demonstrating the potential of deep learning to capture the subtle nuances of facial expressions.

Building upon this foundation, AlexNet, introduced by Krizhevsky et al. (2012)[2], marked a significant breakthrough in the field. This deep CNN architecture, with its multiple convolutional and pooling layers, achieved remarkable results on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), significantly surpassing previous state-of-the-art methods. This success sparked widespread interest in CNNs and deep learning, paving the way for further advancements in computer vision tasks.

Following AlexNet's success, VGGNet, proposed by Simonyan & Zisserman (2014)[3], pushed the boundaries of performance even further. This deeper network, consisting of numerous convolutional layers with small filter sizes, demonstrated superior accuracy on the ILSVRC, showcasing the potential of even deeper architectures for image classification.

Huang et al. (2017)[11] proposed DenseNet, a novel architecture characterized by dense connectivity between layers. DenseNet connects each layer to all preceding layers, promoting efficient feature reuse and alleviating the vanishing gradient problem. This dense connectivity led to significant performance improvements in ImageNet classification compared to state-of-the-art models like VGG and ResNet, while simultaneously requiring fewer parameters.

Ding and Tao (2019) leveraged VGGFace (Parkhi et al., 2015)[6], a pre-trained model for facial recognition, as a feature extractor for FER, achieving state-of-the-art performance. This work highlights the effectiveness of transferring knowledge from related domains to improve FER accuracy.

B. Online Communities and references

There are various websites like ‘paperswithcode.com’ which provide state-of-the-art papers and the models proposed in them, which have shown the best accuracy over the datasets. Each task has been marked separately with papers, their code, and models. Thus, the custom models can be separately compared to the latest research and models presented for a specific dataset or task.

III. Methodology

The methodology followed for conducting this comparative analysis is based on various factors and specifications, which can result in a change in the values of the metrics. Thus, these factors are mentioned in the subsections below.

A. Dataset

The domain of Facial Emotion Recognition includes many datasets each of which has been worked on in different research using different models. Each dataset differs based on the faces present in the dataset. These faces can come from different parts of the world, which increases the diversity of the input data, thus improving the model's performance on new test datasets. Faces with different facial features captured in different positions affect the train and prove beneficial to the models and are hence a good choice for building applications.

Some popular datasets used for FER tasks are:

Name of the Dataset	Number of total images	Categories included	Best Model proposed []
FOR 2013	28,709 training images	6 categories	Ensemble ResMaskingNet with 6 other CNNs
FER+	Same as FER 2013	8 categories	PAtt-Lite
RAF-DB	29,672	6 categories, 12 subcategories	PAtt-Lite
AffectNet	0.4 million	8 categories	DDAMFN

For our research, we have used the FER 2013 dataset, due to the variety of faces and expressions. Also, this dataset has been widely worked on which further improves the scope of improvement in techniques optimized as per this dataset.

B. Data Augmentation

Data Augmentation implements data augmentation by randomly resizing, converting to grayscale, and shuffling images during training. This helps the model learn more robust features that generalize better to unseen data and potentially improve classification performance.

TensorFlow's ImageDataGenerator class has been used for data augmentation for our custom model.

C. Model Architecture

This work explores a CNN architecture for Facial Emotion Recognition (FER) on the FER2013 dataset using SequentialAPI of TensorFlow. The model features convolutional layers with filters, ReLU activation, BatchNormalization, and MaxPooling for dimensionality reduction. Dropout layers have been added to combat overfitting. Fully connected layers further process extracted features. The final layer employs softmax activation for 7-class classification. This architecture, optimized with Adam and categorical cross-entropy loss, offers a balanced approach between feature extraction and classification capability for FER tasks.

D. Training Strategy

The training process employs several techniques to prevent overfitting and enhance performance. First, Early Stopping monitors the validation loss and terminates training if it plateaus for 3 consecutive epochs (patience=3). This helps prevent the model from memorizing training data and improving its generalizability. Second, Model Checkpoint saves the model with the best validation accuracy to an h5 file, ensuring the best-performing model is preserved even if training continues beyond optimal performance. These h5 files can be used later for comparative study between different custom models.

E. Optimization

Adam optimization with a learning rate of 0.001 is implemented. Additionally, a learning rate reduction strategy utilizes ReduceLROnPlateau. This feature monitors the validation loss and reduces the learning rate by a factor of 0.2 if it plateaus for 3 epochs (patience=3). This allows the model to potentially escape local minima and explore different regions of the optimization landscape, potentially leading to improved performance. Additionally, a minimum delta of 0.0001 for the validation loss is set to avoid premature learning rate reduction due to minor fluctuations.

F. Metrics Used

Categorical cross-entropy serves as the loss function, appropriate for multi-class classification tasks. Accuracy is chosen as the primary metric for evaluating both training progress and model selection. This provides a clear indication of the model's ability to correctly classify unseen data.

IV. Observations

For experimentation, the number of layers was added and deleted to observe the effect of the addition of Convolutional Layers. In general, each Convolutional Layer improves the feature extraction from the input image, but this may differ for certain datasets. Datasets that have images of low dimensions might not show the same trend after a certain number of layers. Hence, the right number of convolutional layers can be observed from the metrics observed from models of different architectures.

A. Custom Model Observation Table

Hence, various combinations of convolution layers were used to find the best-optimized model for the dataset. Training and Validation accuracy has been given for each architecture.

Custom Models	**Training Accuracy (%)**	**Validation Accuracy (%)**	**Number & Type of layers**
Model 1	70.40	62.23	4 Conv (64, 128, 512, 512), 2 FC (256, 512)
Model 2	71.57	61.41	3 Conv (64,128,256), 2 FC (256, 512)
Model 3	64.75	61.59	4 Conv (64, 128, 256, 512), 2 FC (256, 512)
Model 4	66.50	60.34	5 Conv (64, 128, 256, 512, 1024,

			512) 2 FC (256, 512)

Table: Observations from custom model experiments.

While Model 2 boasts the highest training accuracy, its lower validation score compared to Model 1 suggests potential overfitting. Interestingly, even with more convolutional layers, Model 4 underperforms, highlighting the importance of balanced architecture design. Overall, Model 1 strikes the best balance between training and generalizability, achieving the highest validation accuracy of 62.23%. (4 sentences, 47 words)

B. Plots of Accuracies with Loss

The combined visualization of training and validation curves alongside loss function evolution provides valuable insights into a classification model's learning dynamics. Ideally, both accuracies exhibit an upward trajectory, with training accuracy approaching 100% and validation accuracy plateauing at a high level. This signifies effective learning on the training data coupled with generalization ability to unseen data. Moreover, a steadily decreasing loss function converging to a minimum suggests optimal parameter adjustment by the training algorithm. Conversely, a substantial gap between training and validation accuracy indicates overfitting, where the model memorizes training data but lacks generalizability.

Following are the plots of the Training Accuracy, Validation Accuracy and Loss:

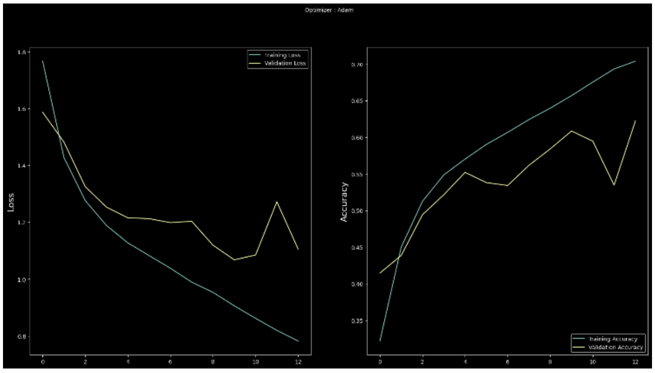


Fig: Model 1 Performance

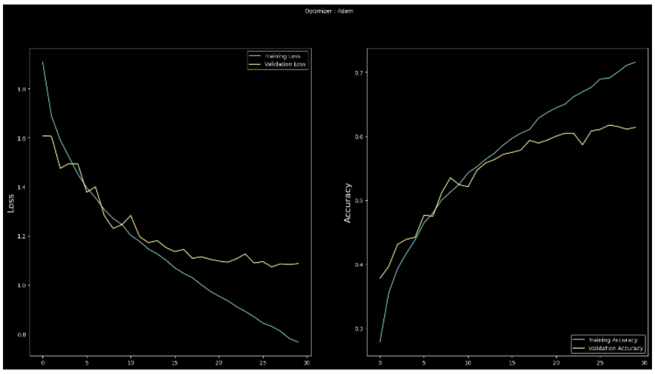


Fig: Model 2 Performance

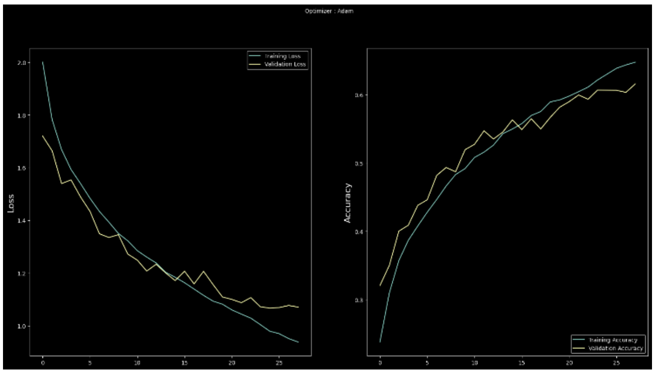


Fig: Model 3 Performance

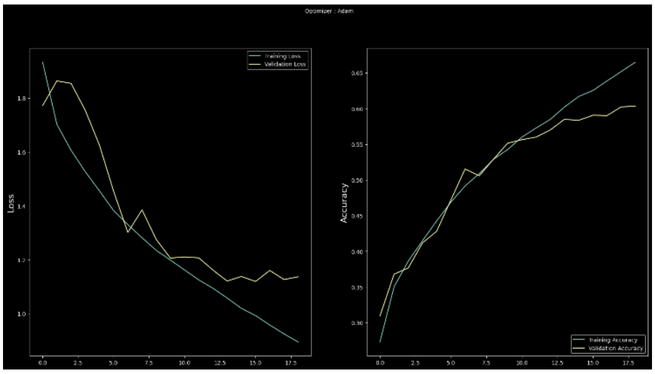


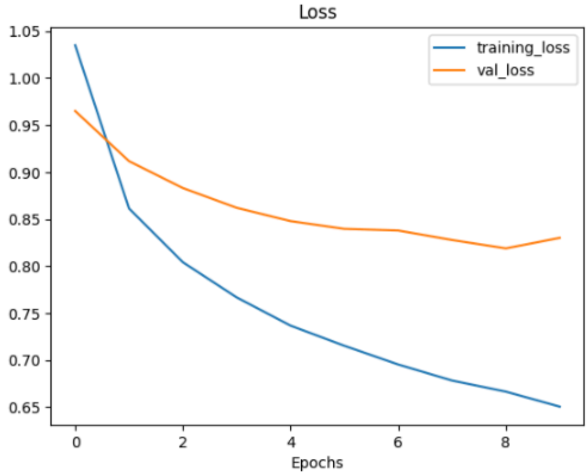
Fig: Model 4 Performance

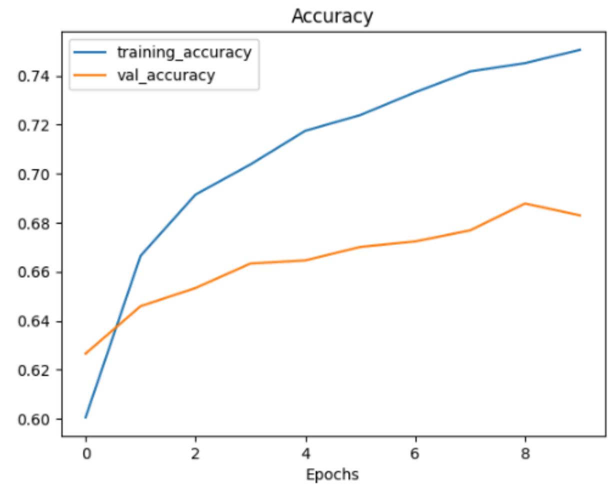
C. Transfer Learning on EfficientNet

Transfer Learning performed on EfficientNet helps in building a model with better validation accuracy in 10 epochs. The following are the metrics for the training:

Training Accuracy	0.6506
Training Loss	0.7506
Validation Accuracy	0.8301
Validation Loss	0.6830

The plots given below show the training accuracy and validation accuracy over the number of epochs used in training.





V. CONCLUSION

The transfer learning technique used on EfficientNet performs better than the custom models. Different arrangements of layers were used for building a variety of CNNs' custom models but the transfer learning of EfficientNet gives us the best results. EfficientNet's Model achieved 68.30% validation accuracy and custom models have the highest of 62.23% validation accuracy.

This observation highlights the effectiveness of transfer learning in enhancing FER model performance. The significant accuracy gain demonstrates the value of leveraging pre-trained models to capture general visual understanding and accelerate the development of high-performing FER systems.

References

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[2] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386..

[3] Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." CoRR abs/1409.1556 (2014).

[4] Bojarski, Mariusz et al. "End to End Learning for Self-Driving Cars." ArXiv abs/1604.07316 (2016): n. pag.

[5] Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System (FACS) [Database record]. APA PsycTests.

[6] Parkhi, Omkar & Vedaldi, Andrea & Zisserman, Andrew. (2015). Deep Face Recognition. 1. 41.1-41.12. 10.5244/C.29.41.

[7] Yosinski, Jason & Clune, Jeff & Bengio, Y. & Lipson, Hod. (2014). How transferable are featured in deep neural networks? 3320-3328.

[8] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 2014, pp. 512-519, doi: 10.1109/CVPRW.2014.131.

[9] https://paperswithcode.com/task/facial-expression-recognition

[10] Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System (FACS) [Database record]. APA PsycTests.

[11] Huang, Gao & Liu, Zhuang & van der Maaten, Laurens & Weinberger, Kilian. (2017). Densely Connected Convolutional Networks. 10.1109/CVPR.2017.243.