Ensemble of Machine Learning Models and Fine-tuned CNN models on FER2013 Dataset

Yamini Dharmasala
114597218
yamini.dharmasala@stonybrook.edu

1. MOTIVATION

The objective of this project is to understand how the ensemble of Machine Learning models performs compared to the ensemble of deep learning models for the problem of facial-emotion recognition. The dataset chosen for this project is FER2013, downloaded from kaggle datasets. The ensemble techniques are chosen from the ensemble methods taught in the class. As this dataset consists of 48X48 images, we need feature map representation of images to pass to the models. CNN architectures have deep convolution layers that does the feature extraction. Some of the famous CNN architectures like Resnets, VGG-nets that performed well on image classification are chosen for this project. Models are developed in torch and exploitted Tensor Processing Units(TPU's) to expedite the training.

2. INTRODUCTION

Facial emotion recognition is the process of detecting human emotions from facial expressions.Humans have the innate ability to recognize and understand the emotions from facial expressions. Now, computers are also able to do the same with the advancements in technology and Artificial Intelligence. It has been applied in many applications related to human-computer interaction, crowd analytics, video analytics, online gaming and biometrics analysis. The basic emotions that are detected using these systems are Happy, Neutral, Angry, Fearful, Surprised, Disgusted and Sad. However recognising human emotions by analysing human face characteristics is really challenging and a difficult task. The difficulty comes from the lightning conditions, pose estimations, occlusions and shadows.

With the advent of Deep Learning Models and Transfer Learning, the state of art CNN models achieved great accuracies with FER2013 dataset. CNN's have convolution layers that extract features at every level and pass down those features to hierarchical layers to form feature maps of images that are used by fully connected(FC) layers for classification.As CNN's can handle large datasets, it's easier to extract features from CNN's compared to the hand-crafted feature extraction methods like SIFT, Local Binary Patterns(LBP), Local Ternary Patterns(LTP),Histogram of Oriented Gradients(HOG) and many.

Fig.1. Face images from FER2013 dataset illustrating variabilities in age, pose, illumination and other factors.

In this project, I used transfer learning to extract features maps of images from FER2013 dataset. The extracted features are passed to the classifiers using different ensemble methods like Voting, Bagging and Boosting. Different pre-trained CNN models like VGG16, Resnet34 and Resnet50 are finetuned by freezing only the CONV layers and trained on FER2013 Dataset. Individual models achieved accuracies of 53.44%, 59.51% and 63.44% on the test dataset respectively. An ensemble of three models with voting approach achieved an accuracy of 65.17% without any extra training data.

3. RELEVANT LITERATURE

With the advancements in CNN architectures and the ensemble of well-performed CNN's have resulted in high accuracies with FER2013 dataset. Quinn et. al. [1] extracted HOG features of FER2013 and CK+ dataset and passed them to OVO SVM classifier which gave an accuracy of 45.95%. With CNN, they achieved an accuracy of 66.67%.

With the single VGGNet architecture, Yousif et. al. [2] rigoursly fine-tuned the hyperparameters by experimenting with various optimizers, schedulers and learning rates. They achieved 73.28% on FER2013, which is state-of-art single network accuracy.

Ensembling of CNN models have been widely used to improve the accuracy than the individual models in image classification. K. Liu et. al. [3] used ensemble of 3 structured CNN models to achieve an accuracy of 65.03% on FER2013 dataset.

In this work Khanzada et. al. [4], used shallow CNN's and pretrained-networks based on SeNet50,ResNet50 and VGG16. By ensembling seven models,data augumentation and using auxilary datasets, they achieved 75.8% on FER2013 dataset.

3.1 DATASET

. The dataset used in this project is FER2013, a large publicly available database with 35,887 face images. It was introduced at ICML in 2013. Since then it became a benchmark dataset for emotion recognition. This dataset is really challenging as the images vary significantly in age, gender, pose,and other factors.It's human level accuracy is at 65.5% and published work have achieved highest accuracy of 76.82%. The dataset is split into train and test sets with 28709 and 7178 samples. All images are grayscale and have a resolution of 48X48 pixels. It's unbalanced with the most label being the happy-8989 and least label being the Disgusted-547

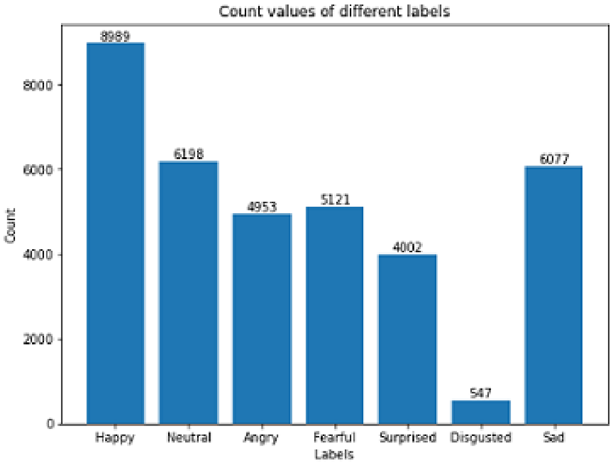


Fig.2. Distribution of labels over the complete dataset

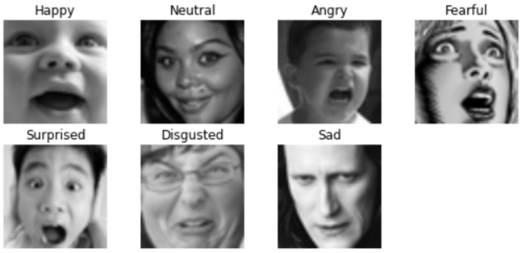


Fig.3. Images representing all the seven face emotions

4. EXPERIMENTS

In this project, Resnet-34 CNN architecture is used to extract the feature maps of FER2013 dataset. Resnet-34 has 4 layers with many residual blocks, an average pooling layer and one fully connected layer. The feature maps from the intermediate layers represent the information about low-level features. The final layer before the average pooling layer gives the feature map that represents the high-level features. The last fully connected layer of Resnet-34 is removed and the feature maps are extracted.

Most of the face expressions can be detected by the position and angles of face parts like eyes, nose, lips and ears. Intermediate layers extract features related to ear,nose,lips and eyes. Final year extract the features related to a happy face,sad face and other emotion related face images. Adding the intermediate layer feature maps to the final layer feature maps can add more weights to the low level features.

4.1 Ensemble Models

Ensemble models combine the predictions from several models in order to improve the overall predictive performance and generalizability over a single prediction model. There are different types of ensembling techniques like Stacking, Voting, Bagging and Boosting.

4.1.1 Voting

In this project, an ensembles of SVM's by major voting approach are used for emotion classification . SVM natively doesn't support binary classification. We use OneVsOne and OneVsRest approaches that divides the multiclass problem into multiple binary classification problems. Sklearn's SVC function also supports both binary and multclass classifictaion. The accuracy achieved with voting approach is 39.48%.

4.1.2 Boosting

Boosting is an ensemble technique that combines a set of weak learners into a stronger learner by increasing predictive power and generalization. In this project, Adaboost classifier is used with hyperparameter tuning. The model achieved an accuracy of 29.92% with adaboost classifier.

4.1.3 Bagging

Bagging also known as Bootstrap aggregating is an ensemble technique that selects random sample of data points (with replacement) and trains several weak learners with chosen random data sets. The results are aggregated by majority or average depending on type of the problem. It significantly raises the performance by reducing the variance. The accuracy achieved using Random Forest classifier is 24.87%.

4.2 CNN Architectures

Over the years, variants of CNN architectures have been proposed due to the advancements in deep learning. Some of the famous architectures are VGG, Resnets , Inception-Net/GoogleNet and DenseNets. Training these deep networks can take several days due to million parameters and also they require large datasets. With Transfer Learning, we can use the weights of a pre-trained model that is usually trained on massive datasets like ImageNet. In this project, variants of VGG and Resnet pre-trained models are used.

4.2.1 Fine Tuning VGG16 Model

VGG16 architecture has 16 layers. It has 13 Convolution(CONV) layers and 3 dense layers. The most unique thing about this architecture is-all the CONV layers have same kernel size-(3,3), stride-(1,1) and padding-(1,1).It also has Batch Normalization layers to reduce over-fitting and improve training performance.The last Dense layers are modified according to the work from [4]. The second dense layer output size has been modified to 1024. The last dense layer output is modified to the number of output classes. All the CONV layers above the dense layers are frozen and the dense layers are kept trainable. After training for 100 epochs using SGD with a 0.01 learning rate and a batch size of 128, the model achieved 53.44% on the test set.

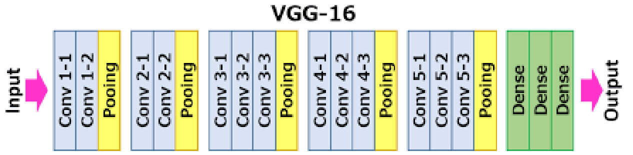


Fig.4. VGG-16 Architecture

4.2.2 *Fine Tuning Resnet-34 Model*

Resnet34 architecture has 34 layers - 33 Convolution(CONV) layers and one fully connected layer. It has residual layers that takes output and pass them to subsequent layers. This solves the vanishing gradients problem in deep networks. In this project, the last fully connected is modified by adding two additional dense layers of size 4096 and 1024. All the 33 convolution layers are frozen and the last three dense layers are made trainable. After training for 100 epochs using SGD with a 0.01 learning rate and a batch size of 128, the model achieved 59.51% on the test set.

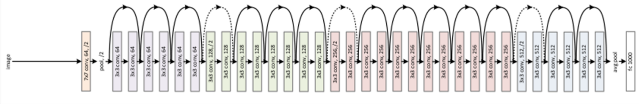


Fig.5. Resnet-34 Architecture

4.2.3 *Fine-Tuning Resnet-50 Model*

Resnet50 is a variant of Resnet architecture and has 50 layers - 48 Convolution(CONV) layers, one average pooling layer and one fully connected layer.The last fully connected layer is modified according to the work in [5]. The original output layer is replaced with two Fully connected layers of sizes 4096 and 1024 respectively. The initial 48 CONV layers are frozen and the modified fully connected layers are kept trainable. After training for 100 epochs using SGD with a 0.01 learning rate and a batch size of 128, the model achieved 63.44% on the test set.

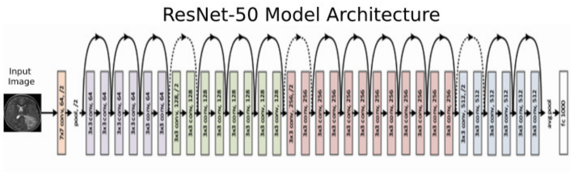


Fig.6. Resnet-50 Architecture

4.2.4 *Ensemble Model*

Ensemble model is build with the predictions based on fine-tuned VGG16, Resnet34 and Resnet50. The output probabilities of the three models are combined and the max probability class will be the output(unweighted soft voting). After training the model for 10 epochs using SGD with learning rate of 0.01 and batch size of 128 gave an accuracy of 65.17% on the test set.

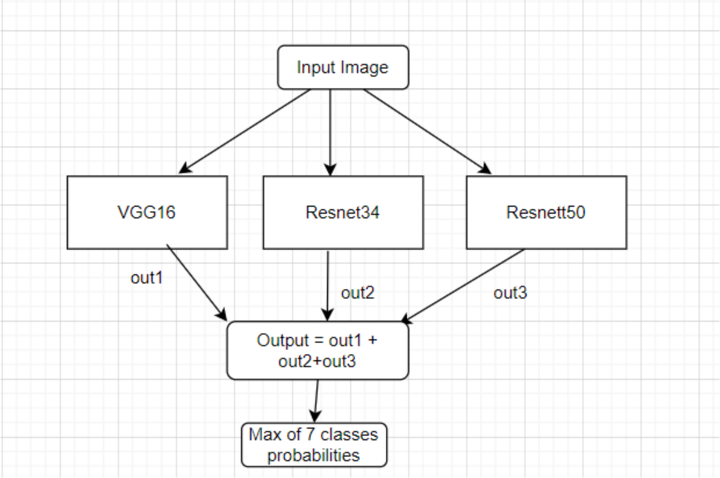


Fig.7. Ensemble Voting Model

5. RESULTS

Table I gives the performance metrics of ensemble of machine learning models. Table II gives the accuracies from the fine-tuned models

Table 1: PERFORMANCE OF ML MODELS ON FER2013

Model	Performance Metrics		
	Accuracy	**Precision**	**F1 Score**
Voting Approach	39.48	63.69	42.49
Boosting	29.92	30.60	29.95
Bagging	24.87	95.36	37.92

Table 2: PERFORMANCE OF FINE-TUNED CNN MODELS ON FER2013

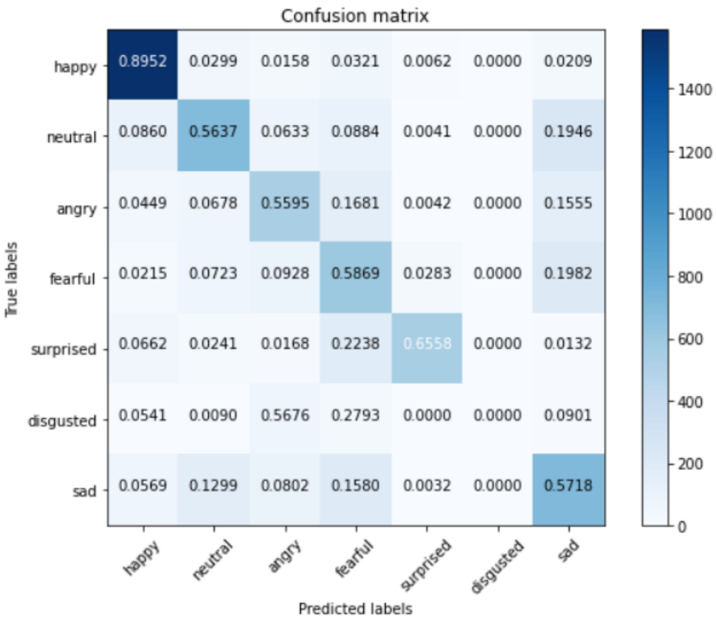
Model	Performance Metrics		
	Accuracy	**Precision**	**F1 Score**
VGG16	53.44	54.60	52.00
Resnet34	59.51	61.32	58.90
Resnet50	63.44	64.51	62.92
Ensemble Voting	**65.17**	**65.76**	**64.99**

6. CONCLUSION

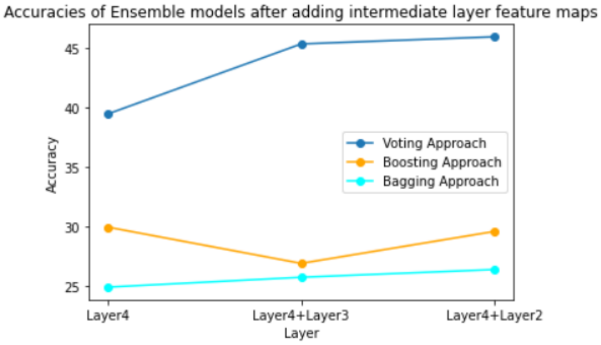
The goal of this project is to understand how the ensemble models works. From the results and experiments, it is clear that ensemble models performs better than their individuals. The purpose of choosing the FER2013 grayscale image dataset is to analyze how the ensemble machine learning models like OVO-SVM, OVR-SVM, Adaboost and RandomForest perform on feature maps representations that have high dimension. For the future work, I would like to explore applying PCA on feature map representations and also combining it with other hand-crafted features. I should also optimize the ensemble network and include more datasets, pre-processing images and data augmentation techniques.

7. REFERENCES

[1] Quinn, M.-A., G. Sivesind, and G. Reis, *Real time emotion recognition from facial expressions.* Standford University, 2017.



Happy label have the highest prediction. All the other labels except disgusted have crossed over 50% in prediction. The model is predicting disgusted as mostly angry due to disgusted label having low data.



CNN extracted feature representations only represent the high-level features. By adding the information about the low-level features have boosted the accuracies of models.

[2] Khaireddin, Yousif, and Zhuofa Chen. ”Facial Emotion Recognition: State of the Art Performance on FER2013.” *arXiv preprint arXiv*:2105.03588 (2021).

[3] K. Liu, M. Zhang and Z. Pan, ”Facial Expression Recognition with CNN Ensemble,” *2016 International Conference on Cyberworlds (CW)*, 2016, pp. 163-166, doi: 10.1109/CW.2016.34.

[4] Khanzada, Amil, Charles Bai, and Ferhat Turker Celepcikay. ”Facial expression recognition with deep learning.” arXiv preprint arXiv:2004.11823 (2020).

[5] Brechet P., Chen Z., Jakob N., Wagner S., “Transfer Learning for Facial Expression Classification”