EmoNet: Facial Expression Recognition using Deep Learning Techniques

Lilly Kumari
Department of Electrical & Computer Engineering
University of Washington Seattle
lkumari@uw.edu

Abstract

In this project, we explore the problem of facial expression recognition (FER) using deep learning techniques where the aim is to recognize seven key emotions: anger, disgust, fear, happiness, sadness, surprise and neutrality. We use the FER-2013 dataset which is a kaggle dataset for this project. Since the focus is on exploring how transfer learning and ensemble learning help in improving the performance of deep models given limited data availability, we start with two state of the art convolutional neural networks, Inception-ResNet v2 and ResNet-50 architectures. To further improve upon our results, we also combine extra information in the form bag of visual words and propose a new deep network which extracts convolutional feature vectors using one of the ResNet blocks and bag of visual words based features such as facial landmarks, and Histogram of gradients. We finally show that our overall accuracy is 73.33% which is almost at par with current best performance of 75.1% which is achieved by training on a corpus of 1.3 million images.

1 Introduction

Non-verbal communication plays an important part in human interactions. In reality, more than 90 % of the human communication is nonverbal. So, understanding and recognizing human expressions is really crucial as they play an important role in interpersonal relations (1). Making systems which are "emotionally intelligent" can have applications in a variety of fields ranging from security and surveillance to recommender systems. Facial emotion recognition can help security systems in filtering violent images and videos. With the growing rise of social media and video sharing platforms, users all over the world upload and share millions of images and videos everyday which provide valuable information about their satisfaction & dissatisfaction about different products and brands. This kind of data can be used by e-commerce companies to change and enhance upon the features that the end-users want. They can also use the satisfaction data to guide their recommendation engines algorithms.

In the area of robotics where human-robot interaction is an important component, facial expressions play a significant role in accurate communications and understanding of the human behavior. It can also be used in clinical practice and behavioral science. So, summing up, automatic and accurate facial expression recognition is really pivotal for a myriad of applications where the facial images could have any orientation and pose.

Previous approaches to this problem (before ConvNets became popular) used to be random forests and Support Vector Machines (SVM) based classifiers (2). With the proliferation of big data and deep learning, there has been some promising results in the FER domain (3). The ability of neural networks to extract undefined abstract features makes them powerful enough to generalize better on unseen test examples. Since the training of very deep neural networks is time consuming and computationally intensive, there has been a growing trend of transfer learning in various computer vision tasks (4).

.

So, in this project, we focus on two state of the art convolutional neural network architectures, Inception-ResNet v2 (5) and ResNet-50 (6) (7) and leverage the kaggle dataset called FER-2013 (8). We choose these two architectures because they have won previous ImageNet based challenges (9) achieving near state-of-the-art performance in terms of prediction accuracy. Besides using these networks for extracting feature representations from raw image pixels, we also experiment with an end-to-end convolutional fully connected neural network which uses either ResNet-50 or Inception-ResNet v2 as the Convolutional block and takes as input a 2728 dimensional engineered feature vector (facial landmarks and HOG features). We show that the network using this extra supplementary information in the form of bag of visual words performs better than the ones trained just on raw image pixels. Finally, using differential evolution algorithm, we design a weighted average ensemble model which combines the predicted output from the possible four model variants and achieves 73.3% accuracy on FER-2013 dataset which is almost comparable to the state-of-the-art accuracy of 75.1%.

2 Related Work

Prior to the deep learning era, researchers used several traditional machine learning classifiers such as bayesian networks, hidden markov models (10) , support vector machines (2) etc in order to build automatic facial expression classifiers. Some of them used to classify the face into a known set of established emotions such as sadness, anger, happiness, surprised etc (11). Other systems in order to provide an objective description of face, worked on recognizing individual muscle movement produced by the face (12). They proposed a psychological framework called Facial Action Coding System (FACS) (13) that classified facial movements using Action Units, where each expression is a combination of several action units (14).

With the advancement of GPU computability, recent work in FER has been focused on using different convolutional neural networks. (15) proposed a network comprising of two convolutional layers followed by max-pooling with four inception layers. They trained their model on seven facial expression datasets such as FER-2013, SFEW, FERA etc. Their method was able to achieve 66.4% accuracy on the FER-2013 dataset. (16) proposed a de-expression learning procedure to extract information of the expressive component of face. The de-expression was performed by training a conditional generative adversarial network, and their method learnt the residue that contained the expressive component. They also trained their model on seven publically available datasets, such as CK+, Oulu-CASIA, MMI, BU3DFE, BP4D+ etc. (17) proposed to focus on the second order statistics, covariance in order to capture the distortions in regional facial features. They used manifold preserving networks in conjunction with convolutional networks for spatial pooling in an end to end learning procedure via training on two datasets, SFEW (static facial expressions in the wild) and RAF (Real World Affective).

Kaggle Competition The FER-2013 dataset was released as a part of ICML 2013 workshop on "Challenges in Represenatation Learning". The winner of this challenge used a deep neural network (using CIFAR-10 weights) to extract feature based representations and then used an SVM for multi-class classification while achieving an overall accuracy of 69.76%.

3 Methods

3.1 Convolutional Networks

ConvNets can successfully capture the spatial and temporal dependencies in an image via the application of relevant filters, along with the benefit of reduced parameters to be learnt. Both presented networks make use of residual (or skip) connections (7) to make the network more dynamic so that it can optimally tune the number of layers during training instead of treating the number of layers in the network as an hyper-parameter to tune.

3.1.1 Inception-ResNet v2 - Model A

Inception-ResNet v2 (5) combines two state-of-the-art architectures, inception networks and residual networks which have shown very good performance with relatively low computational cost. Inception modules make the network wider than deeper by applying filters with multiple sizes on the same

level. The computational speed is improved by factorizing bigger convolutions to several smaller convolutions. Since neural networks performance is better when convolutions don't alter the dimensions of the input drastically, a drastic reduction in the dimensions may lead to loss of information, known as a "representational bottleneck". So, this problem is eliminated by expanding the filter blocks. The network also utilizes "Reduction Blocks" to change the width and height of the input grid.

Residual connections are introduced to add the output of the convolution operation of the inception module to the input. Each Inception block is followed by a filter-expansion layer (1x1 convolution) that is used to scale up the dimensionality of the filter bank before the addition in order to match the depth of the input. This is needed to compensate for the dimensionality reduction brought upon by the Inception block.

Residuals are scaled before adding them to the previous layer activation in order to stabilize the training. Overall, the network has three types of Inception modules and two types of Reduction blocks based on the filter factorization and depth at which residual connections are placed. Figure 1 shows the overall architecture of Inception-ResNet v2 with schematic diagram of each Inception and Reduction block. Pooling operations are present only in the Reduction block.

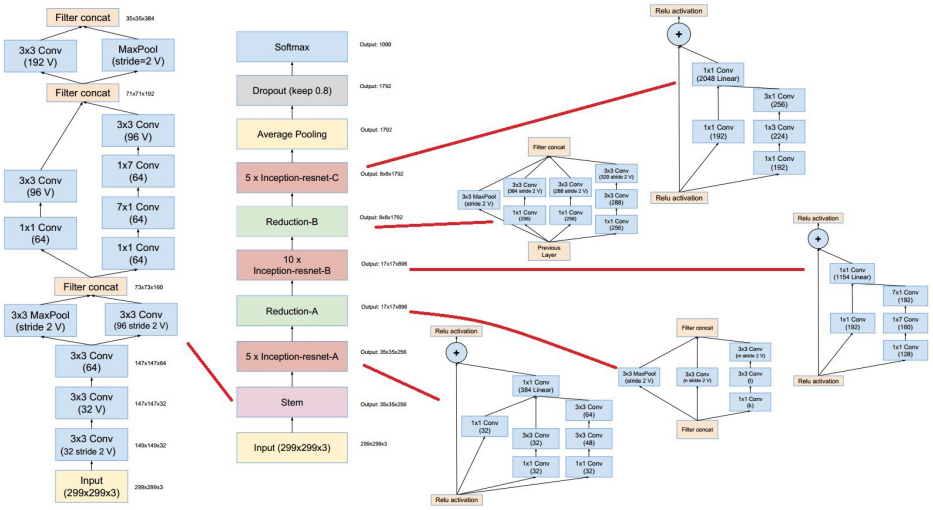

Figure 1: Inception-ResNet v2 architecture

In our experiments, we have dropped the last dense layer of this network. The input image should be greater than 139x139 and the image must have 3 channels.

3.1.2 ResNet 50 - Model B

ResNet-50 is another state of the art convolutional network architecture which is very much similar to the VGG-16 network. It uses skip connections (7) as shown in Figure 2 to eliminate the issue of vanishing gradients by providing another shortcut path for the gradient to flow through. This makes it convenient for the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer. It also enables the model to circumvent a vanilla convnet weight layer if the current layer is not sufficient.

It uses 1x1 convolutions to match the dimension of the residual to the input. This network overall consists of five stages, each with a Convolution block and an Identity block. Each convolution block has 3 convolution layers and each identity block contains 3 convolution layers with altogether 23 million trainable parameters. For our experiments, we have dropped the last dense layer with softmax activation. The input image should be greater than 197x197 and the image must have 3 channels. Figure 3 shows the schematic diagram of the ResNet-50 network.

3.2 Proposed Methodology

Prior to the popularity of deep learning, feature engineering from raw pixels data was pretty common. In this project, while simultaneously experimenting with one of the above mentioned ConvNets

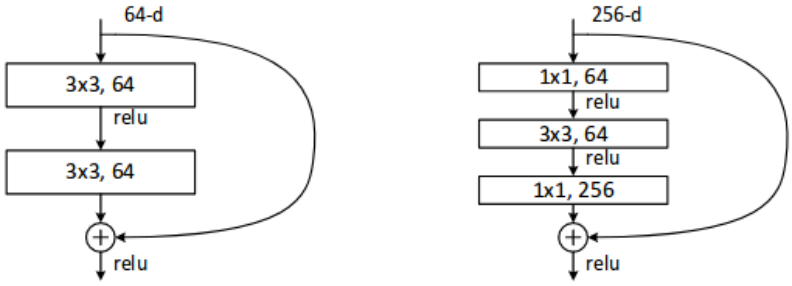

Figure 2: ResNet residual block with skip connections

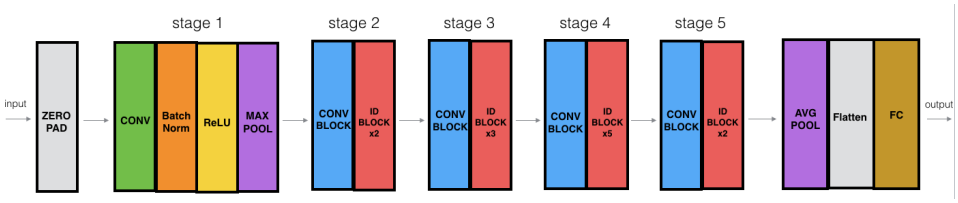

Figure 3: ResNet-50 architecture

for extracting abstract features from raw image pixels, we also utilize additional information in the form of facial landmarks and histogram of gradient features which we call as "Bag of Visual Words" - BOVW similar to what words mean in the context of a sentence. In order to make use of this additional information in the form of BOVW, we propose a new network architecture as shown in Figure 5. This network takes two inputs, first - a raw image which is resized according to the Conv Block requirement (either 139 or 197) and stacked up to comprise 3 channels; second - a 2728 dimensional feature vector comprising of 68 facial landmarks (x and y co-ordinates, so 68*2 features) and 2592 dimensional Histogram of Gradients binned into 8 unsigned orientations. Example shown in Figure 4.

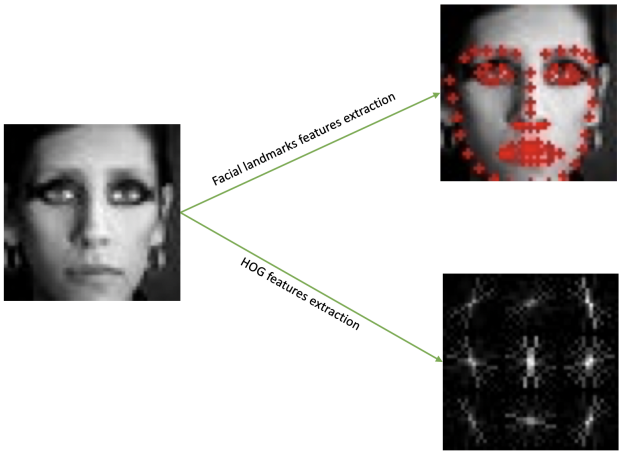

Figure 4: Bag of Visual Words (BOVW) Feature Extraction

So based on the type of Conv Block and on the usage of additional BOVW features, we have overall 4 different network variants which we train in an end to end fashion.

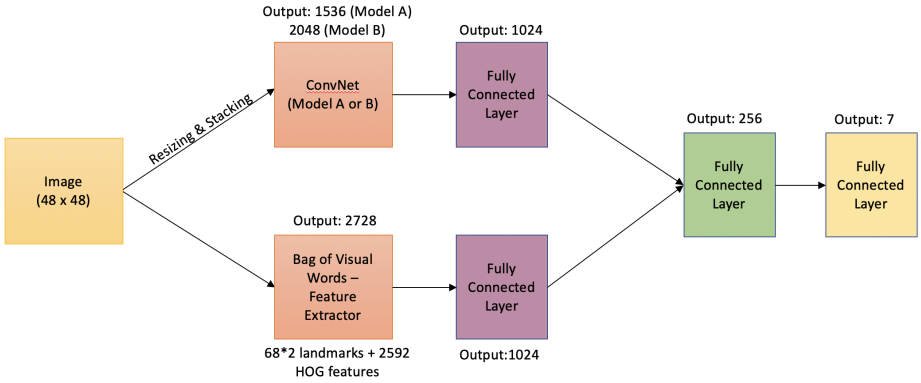

Figure 5: Proposed Network Architecture

Transfer Learning Transfer learning (4) is a technique in which learnt weights of a particular model are used to train another neural network by fixing the weights of various layers and fine-tuning the remaining layers. It is very popular in the field of computer vision, specifically for object recognition and classification tasks. In this project, we utilize this technique by initializing our network in Figure 5 with ImageNet trained weights when the Conv Block is Model A i.e. Inception ResNet v2 and with VGGFace2 (18) trained weights when the Conv Block is Model B i.e. ResNet-50. The remaining part of the network except for the Conv block is trained completely. After few epochs, all the fixed weights are unfrozen and the entire network is fine-tuned further to improve upon the training set performance.

Ensemble Learning In machine learning, ensemble methods that boost the predictions of two or more networks are very popular (19). These ensemble methods take as input the predictions of individual models and assign an weight to each model's predictions by learning it over some held-out dataset while minimizing some loss objective. In our project, because of limited data availability (28k images), we leverage this method to boost the performance of our final model by learning weighted average across all 4 network variants.

4 Dataset

In this project, we use FER-2013 dataset provided by Kaggle (8) which was launched during ICML 2013 workshop on "Challenges in Representation Learning". This dataset consists of 35,887 grayscale images of size 48x48, most of them in wild settings, as shown in Figure 6. These images are processed such that the face is mostly centered and each face occupies roughly the same amount of space in each image. The training set comprises of 28709 images while the validation and testing set contain 3589 images each. The dataset is uniformly distributed with respect to gender, race and ethnicity of the subjects. The images are annotated with 7 key emotions: *happiness, sadness, anger, disgust, fear, surprise & neutrality*, with individuals posing at various angles. FER-2013 has almost an even distribution of emotions across the subjects except for the "disgust" emotion as shown in Figure 7. This dataset is challenging as some of the images are incorrectly labeled since it was collected using the Google image search API. So, this makes the training hard for the models as they need to generalize well on new samples and be robust to the incorrect labels which are nothing but noises.

Data preprocessing: The images are normalized via mean-centering i.e. via subtracting the mean of the training images from each image. For the purpose of data augmentation, we perform random rotations in the range of 10 degrees, random shear transformations, random zooming inside the images, and mirroring them by flipping them horizontally.

5 Implementation Settings

- openCV and dlib libraries used for face detection in order to filter out images containing zero or more than one faces.

Figure 6: Sample images from FER 2013 dataset

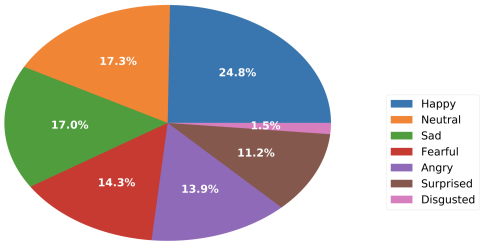

Figure 7: Distribution of 7 emotions in FER 2013 dataset

- Data processing: Mean centering, standard resizing and stocking, dlibś pretrained facial landmark detector used to identify 68 landmarks, HOG feature extraction with a sliding window size of 24 and a window step of 6
- *tensorflow* deep learning framework used for all neural network architectures.
- *Transfer Learning* leveraged for Convolutional Feature Extractor Networks - network initialized from ImageNet weights in Model A and VGGFace2 weights in Model B
- For first 5 epochs with a batch size of 64, the pretrained convolutional blocks are frozen and the remaining network is trained using an Adam optimizer with a learning rate of 0.001 with categorical cross entropy loss.
- For next 50 epochs with same batch size, the entire network is fine tuned using a SGD optimizer with nesterov momentum factor of 0.9 and a learning rate of 0.001.
- Learning rate scheduler is used to shrink the learning rate by monitoring validation loss changes and early stopping criterion is used to prevent the model from overfitting on the training set.
- Used differential evolution algorithm built in *scipy* for learning weighted ensemble of different model outputs.

6 Results

6.1 Qualitative Results

In Figure 8, the third image in both second and third row have wrong ground truth annotations/emotions. The child clearly doesn't have a facial expression of fear; it resonates more with a neutral expression which is the model's prediction. The woman in the 3rd image in 3rd row doesn't have a happy expression, while the model's prediction of sadness syncs well with the image. So, in a way, the final model is robust to incorrect ground truth emotions present in some images.

6.2 Quantitative Results

We report the performance of our models with respect to overall accuracy as well as per-class precision, recall and f1-score metrics. The results show that weighted average ensemble model learnt using both Inception ResNet v2 and ResNet-50 models outputs perform the best.

Figure 9 and 10 show the classification report corresponding to the network using Model A (Inception-ResNet v2) as convolutional feature extractor without and with BOVW features respectively.

Figure 11 and 12 show the classification report corresponding to the network using Model B (ResNet-50) as convolutional feature extractor without and with BOVW features respectively. Figure 13 shows the classification report (for each emotion) corresponding to the ensemble model learnt using all 4 models's predictions. The ensemble weights learnt over the 4 models using the differential evolution algorithm on the held-out i.e. validation dataset are shown in Table 1

Table 2 shows the overall accuracy of all 4 trained models as well as the ensemble models on the test set.

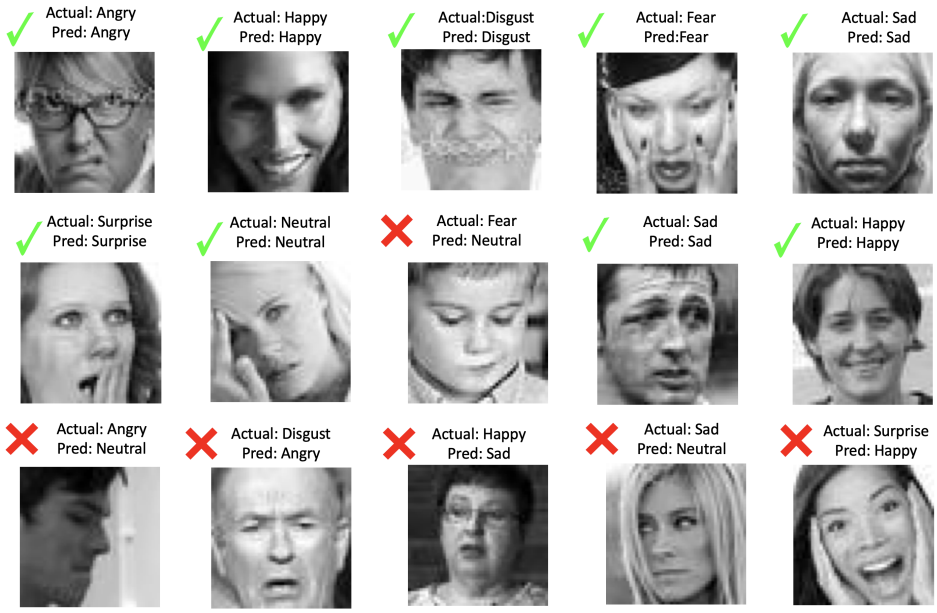

Figure 8: Sample predictions on FER-2013 using the final ensemble model

	Learnt weight
Inception-ResNet v2 (pixels only)	0.146737
Inception-ResNet v2 (pixels + BOVW)	0.210536
ResNet-50 (pixels only)	0.285459
ResNet-50 (pixels + BOVW)	0.357268

Table 1: Ensemble Weights for 4 trained models on FER-2013 dataset

	Accuracy
Inception-ResNet v2 (pixels only)	62.75%
Inception-ResNet v2 (pixels + BOVW)	64.55%
ResNet-50 (pixels only)	68.92%
ResNet-50 (pixels + BOVW)	70.97%
Weighted Ensemble Learning	73.39%

Table 2: Performance of different models on FER-2013 dataset

It can be clearly seen that the model trained with additional BOVW performs better than their respective counterparts. Also, the final weighted ensemble model is able to achieve a test accuracy of 73.39% which is comparable to the best current accuracy of 75.1% on FER-2013 dataset which is achieved via augmenting three datasets together in order to increase the training data size and compensate for the small size of FER images (20).

7 Conclusion

We explore two state of the art convolutional neural networks based on Residual connections and show that supplementing additional information in the form of facial landmarks and HOG features can improve the model's performance with almost negligible increase in the number of trainable weights i.e. computational cost. Also, an ensemble learning approach combining the output of multiple models can assist in further boosting the overall performance, thus eliminating the need for a large corpus of training images.

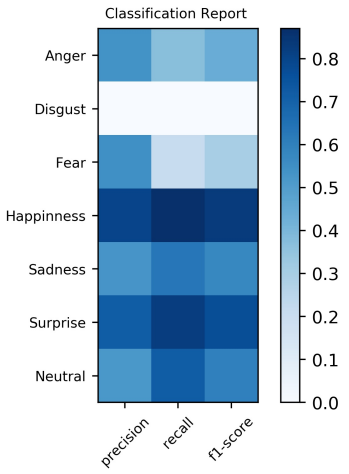

Figure 9: Network using Inception-ResNet v2 Model (A) for raw pixels only

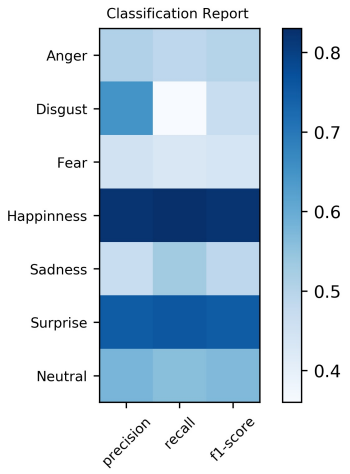

Figure 10: Network using Inception-ResNet v2 Model (A) for raw pixels + BOVW

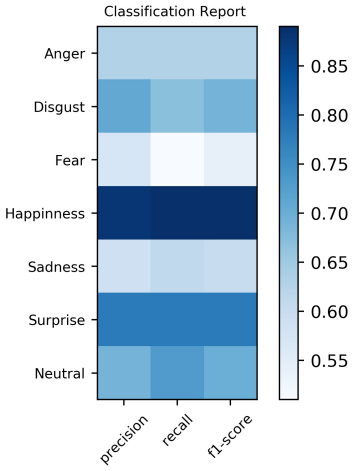

Figure 11: Network using ResNet-50 Model (B) for raw pixels only

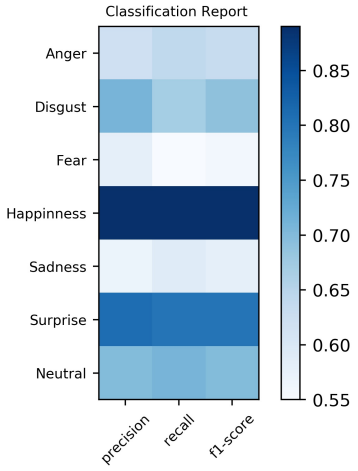

Figure 12: Network using ResNet-50 Model (B) for raw pixels + BOVW

Acknowledgments

We are thankful to Prof. Eli Shlizerman and our course T.A. Kun Su for their valuable feedback and suggestions.

References

[1] A. Lonare and S. V. Jain, "A survey on facial expression analysis for emotion recognition," *International journal of advanced research in computer and communication engineering*, vol. 2, no. 12, 2013.

[2] P. Michel and R. El Kaliouby, "Facial expression recognition using support vector machines," in *The 10th International Conference on Human-Computer Interaction, Crete, Greece*, 2005.

[3] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Deep learning for emotion recognition in faces," in *International Conference on Artificial Neural Networks*, pp. 38–46, Springer, 2016.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

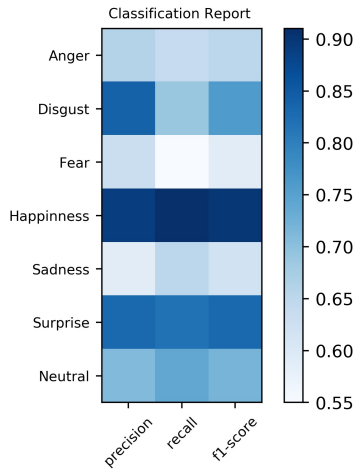

Figure 13: Ensemble Learning using Model A and Model B predictions with and without BOVW features

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[8] "https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge,"

[9] "http://image-net.org/challenges/lsvrc/,"

[10] J.-J. J. Lien, *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity*. University of Pittsburgh, 1998.

[11] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 80–80, IEEE, 2004.

[12] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, J. R. Movellan, *et al.*, "Automatic recognition of facial actions in spontaneous expressions.," *Journal of multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[13] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[14] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1449–1461, 2004.

[15] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10, IEEE, 2016.

[16] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168–2177, 2018.

[17] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 367–374, 2018.

[18] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

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