Performance Comparison of Deep Learning Methods in Facial Emotion Recognition

Sai Krishna Karthikeya Dulla, Shahidhya Ramachandran,
Sriram Sitharaman

Department of Data Science
School of Informatics, Computing and Engineering
Indiana University - Bloomington

May 1, 2018

1. Abstract

Facial Emotion Recognition (FER) plays a vital role in Monitoring systems, Entertainment, Consumer Marketing, Education, Health, etc. This project presents a comprehensive analysis of the FER performance of various Deep Learning network architectures like AlexNet, VGGNet, ResNet, CapsuleNet and an ensemble of AlexNet, VGGNet and ResNet. The task is to classify am image into one of the seven categories: Neutral, Happy, Sad, Surprise, Anger, Fear and Disgust. Convolutional autoencoders were developed to map each emotion to a latent space which were then used to classify examples based on distance proximity. Though the ensemble network yielded accuracies close to State-of-the art, the memory and run-time requirements were intensive. This was minimized by compressing the ensemble network using it's Dark Knowledge without any deprecation in performance. The models were trained, validated and tested on 50K images from the Affectnet database. The performance of the models were compared in terms of the number of network parameters and accuracy. Ensemble network with dark knowledge gave the best accuracy of 62% with 91% reduction in parameters as compared to the parent ensemble network.

2. Introduction

Human Emotions have been widely studied in order to understand intent. Images, speech and text serve as different components conveying emotion. Two-thirds of human communication is through non-verbal communication [1]. The field of visual detection has been growing rapidly and the market is predicted to grow as much as 7.76 Billion USD by 2022 [2]. One of the key areas in this field is Emotion detection which has aided the development of several innovative applications over the years. Facial Emotion Recognition plays a vital role in consumer marketing, Education, Health, Monitoring systems, Entertainment etc. The rapid growth in Facial emotion detection can be attributed to the development of deep learning in Computer Vision. State-of-the-art systems have been constructed using Convolutional Neural Networks(CNN). Real time systems have also been developed to extract emotion information from video streams [3].

The task of emotion prediction is particularly challenging due to a plethora of reasons. The major challenge is to develop a system that is robust to changes in position of the camera, pose, lighting, presence of occlusion etc. In reality, humans use a wider range of facial expressions than the seven basic expressions, with some expressions being combinations of the seven [4]. The systems trained using posed indoor images do not generalize well on actual real-time images. Further, Deep Learning models are memory and time consuming to deploy in hand-held devices.

In this project, the problem has been solved using multiple approaches and the results have been compared. Existing Neural network architectures like AlexNet, VGGNet, ResNet and CapsuleNet. Additionally, Convolutional Autoencoders, ensembling and dark knowledge techniques were used to achieve higher accuracy and minimize parameters. Capsule Networks did not perform as well as the simpler architectures.

Autoencoder approach was adapted from 'Facial Emotion Detection Using Convolutional Neural Networks and Convolutional Autoencoder Units' [5] where the authors have used Autoencoders to create a unique representation for each emotion. A one-dimensional image vector that disregarded the structural orientation was used to represent the emotion. We attempted to improve the performance of the system by using convolution filters as the encoder and deconvolution filters as the decoder. However, this did not help in improving the accuracy.

3. Existing Work

A plethora of research work exists in the field of human emotion recognition from facial images. Early research in this field, conducted by Michael Lyons and Shigeru Akamatsu (1998) images were coded using a multi-orientation, multi-resolution Gabor filters which were then approximately aligned with the face. This method achieved a rank correlation of 0.67 with the semantic ratings created by human observers [6].

Following this, few researchers form the London University devised a model to improve facial emotion recognition accuracy on the Cohn-Kanade Database. Their system used Local Binary Patterns to represent important micro-patterns of facial images. Caifeng Shan et. all (2005) performed template matching based on weighted Chi square value and used support vector machine to classify facial expressions. They were able to achieve an accuracy of 79.1% for the seven class prediction [7].

In the paper 'Face expression recognition with a 2-channel Convolutional Neural Network' by Hamester et. all (2015), 2-channel CNN was used to achieve an accuarcy of 94.1% on the JAFFE dataset. Channel 1 of the architecture constitutes a Standard CNN and the additional channel which is trained in an unsupervised fashion as a Convolutional Autoencoder [8].

In 2017, Xiu and Hu proposed a Feature Redundancy Reduced (FRR) CNN for Facial expression recognition [9]. FRR-CNN presents a discriminative mutual difference among feature maps of the same layer. This helps in creating less redundant features and more compact representation of images. The transformation-invariant pooling layers help to extract important features. This architecture has produced better accuracies than the state-of-the-art on two facial recognition databases.

4. Data Sources

AffectNet [10] database contains nearly One Million facial images collected from the Internet by querying three search engines using 1250 emotion related keywords in 6 languages. Unlike JAFFE dataset, this dataset does not contain images captured in controlled environments where every person poses with seven different emotions. Rather, it contains images of people from all angles (including side poses), different countries, ethnicities, gender and age. Of these 1 Million images, nearly 420K were annotated for the 6 basic facial expressions and 1 neutral emotion. Figure 1 shows examples of each class from the dataset. For this project, 65K images were used- 48K for training, 12K for validation and 15K for testing. Owing to a variation in images sizes in the database, all of them were re-sized to 96x96.

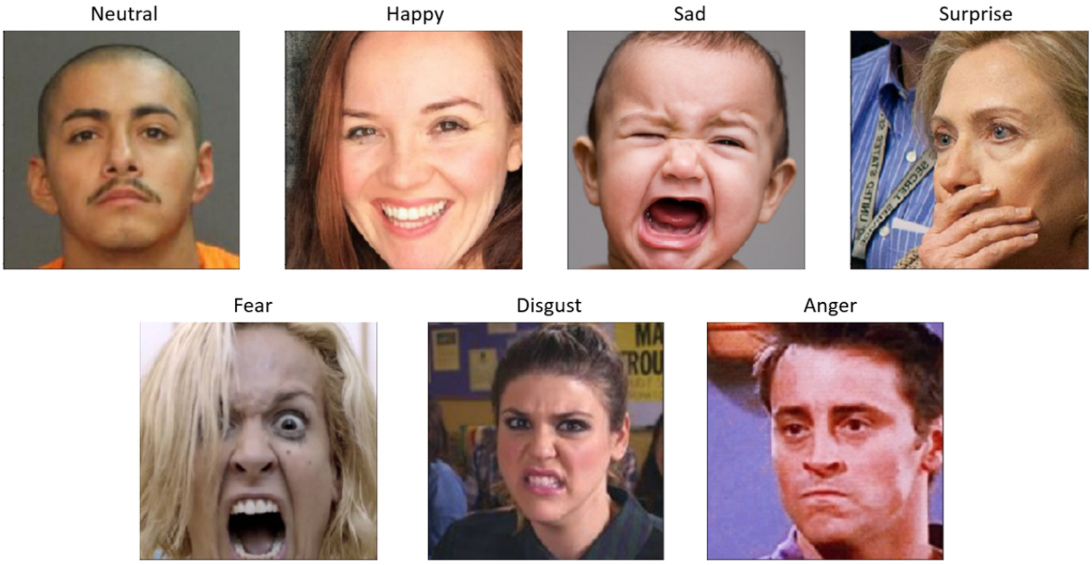


Figure 1: Seven classes of emotion

5. Methodology

5.1 AlexNet

The AlexNet architecture used for training the model is shown in the Figure 2. The images were passed through 5 convolution layers followed by max pooling. The output of the last convolution layer was fed to a fully connected neural network with 2 layers of 156 and 128 neurons each. ReLU activation function was used and a dropout probability of 0.2 was used in each of the Convolutional layers. A batch normalization layer was used prior to each max pool layer.

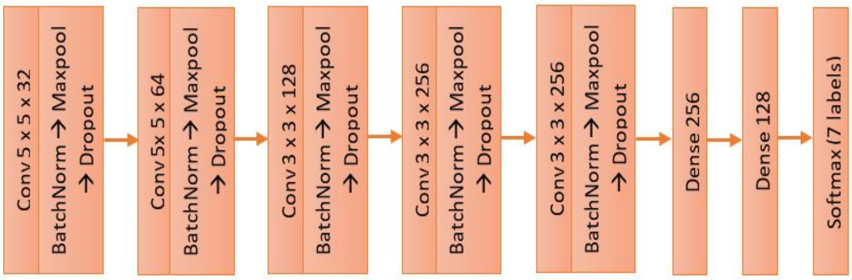

Conv 5 x 5 x 32
BatchNorm → Maxpool → Dropout
Conv 5x 5 x 64
BatchNorm → Maxpool → Dropout
Conv 3 x 3 x 128
BatchNorm → Maxpool → Dropout
Conv 3 x 3 x 256
BatchNorm → Maxpool → Dropout
Conv 3 x 3 x 256
BatchNorm → Maxpool → Dropout
Dense 256
Dense 128
Softmax (7 labels)

Figure 2: AlexNet Architecture

5.2 VGGNet

The design of the network architecture in this scenario was inspired from the original VGGNet paper [11]. The network uses simple 3x3 convolution layers with the same padding and ReLU activation,stacked on top of each other in increasing number of filters with a max pooling layer to reduce the volume size. A batch normalization layer was used prior to each max pool layer and a dropout layer (with a dropout probability of 0.2) was used after every max pool layer to alleviate the overfitting problem. In the end there were two fully connected layers; one with 512 units and other with 256 units, with a softmax layer at the end which predicts a 7-dimensional one-hot vector.Total number of parameters included in this architecture are 3.3 million.

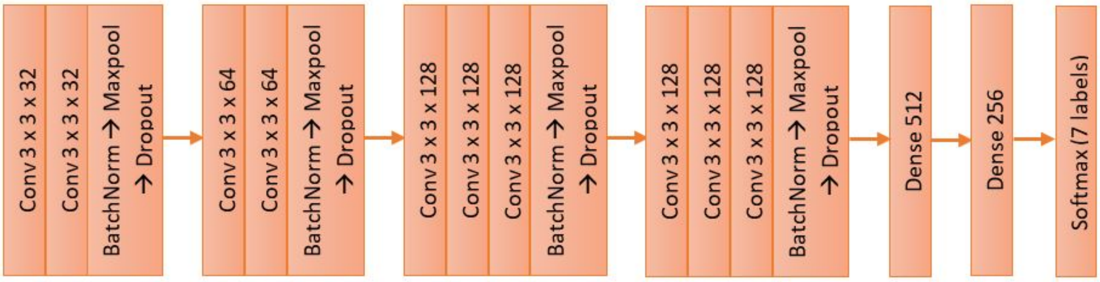
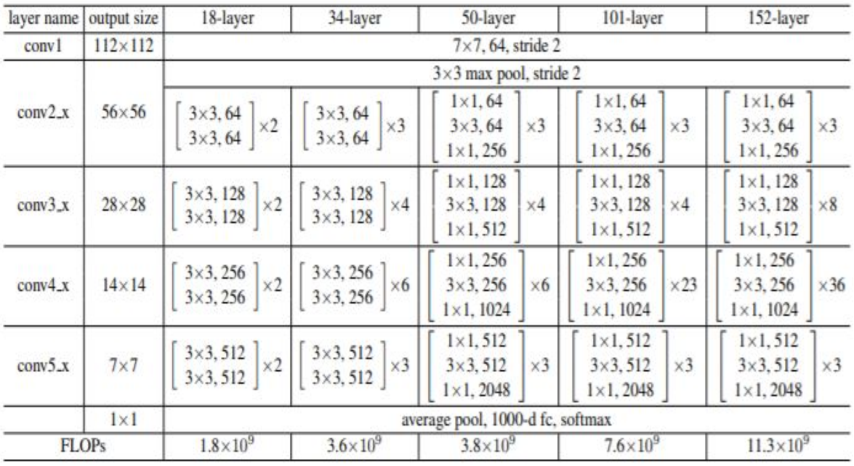


Figure 3: VGGNet Architecture

5.3 ResNet

The architecture used for training the ResNet model is shown in Figure 4. It is a deep Convolutional Neural Network with 18 layers. Batch Normalization, dropout regularization and MAx-pooling was performed after every convolution layer. The implementation from [12] was used for training the model.



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64 3×3, 64] ×2	[3×3, 64 3×3, 64] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3
conv3_x	28×28	[3×3, 128 3×3, 128] ×2	[3×3, 128 3×3, 128] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×8
conv4_x	14×14	[3×3, 256 3×3, 256] ×2	[3×3, 256 3×3, 256] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×23	[1×1, 256 3×3, 256 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512 3×3, 512] ×2	[3×3, 512 3×3, 512] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 4: ResNet Architecture [13]

5.4 Capsule Network

The same architecture mentioned in Geoffrey E. Hinton's 'Dynamic Routing between Capsules' [14] was used for training the network. The first layer was a convolution layer with 256 5 x 5 filters. The second layer was the Primary capsule layer with a squashing function. The primary layer was then followed by the capsule layer where dynamic routing takes place. In the decoder network, the ground truth labels were used to mask the output of the capsule layer. It contains two dense layers with 128 and 256 units respectively. The implementation from [15] was used for training the model.

5.5 Ensemble Netowrk

An ensemble of the trained networks VGGNet, AlexNet and ResNet were considered to predict the output label. The vote from each network was recorded for a test sample. Each examples was assigned to the label with highest number of votes. Incase of a tie the label corresponding to the one predicted by best model was used.

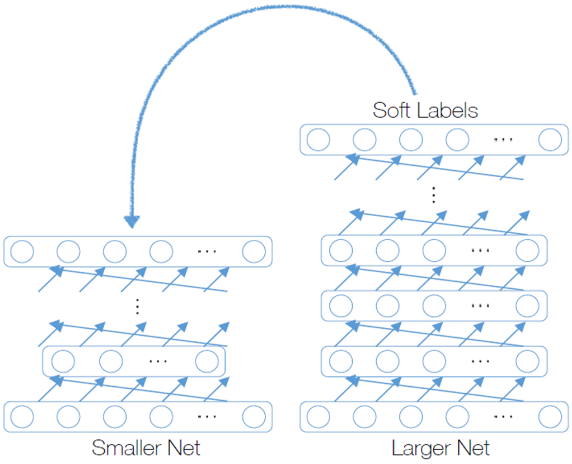


Figure 5: Dark-Knowledge Architecture (Image Source: ENGR-E 533 "Deep Learning Systems" Lecture 09: Network Compression)

5.6 Smaller Network using Dark Knowledge

The structure of the smaller network is as shown in Figure 6. The softlabels (probability vectors) for the training data were computed using the Ensemble network mentioned earlier. The new ground truth labels were computed by averaging the softlabels and the original ground truth labels to help the network resolve any confusion between top few classes. These new labels were passed as the ground truth for a much smaller network, which was designed based on the VGG model. The new model used 3x3 convolution layers with same padding and ReLU activation, stacking up convolution layers in each stage was avoided. The Batch normalization, Maxpool, and Dropout layers were used similar to the large VGG network, and the number of units were reduced to 256 in the dense layer. Total number of parameters included in this architecture are 1.5 million, which is way less than the original network (3.3 million).

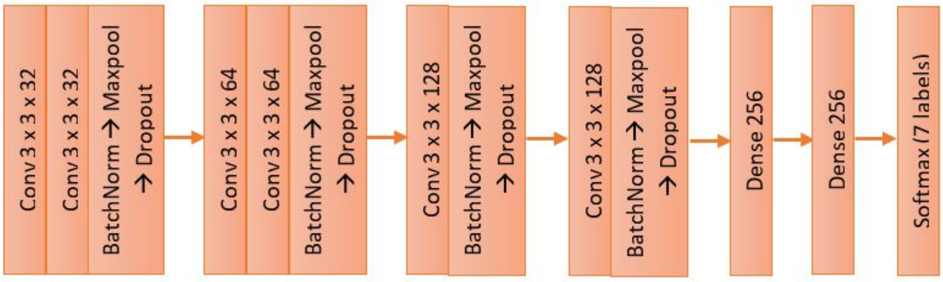


Figure 6: Small Network Architecture

5.7 Convolutuional Autoencoders

Multiple convolutional auto encoders were trained with data specific to each class and the corresponding average latent embedding was computed from the training data. During the testing phase, each test sample was passed to the 7 auto encoders to compute the embedding vectors of that sample in latent space. The sample was then assigned to the label with most similar embedding vector among all the other vectors. In the structure of the auto-encoder, there were three 3x3 convolution layers with 32, 16, 8 filters respectively stacked on top of each other for the encoding part. The final convolution layer output was flattened into dense layers with 1152 units, and 500 units respectively. The last layer was used to capture the embedding. The decoder part had a dense layer with 1152 units attached on top of the embedding layer, and De-Convolution layers were in the reverse order with 8, 16, 32 filters respectively attached on top of the dense layer. The final convolution layer used sigmoid activation function. Binary cross-entropy was used as the loss function in the reconstruction of the image.

6. Experiments and Results

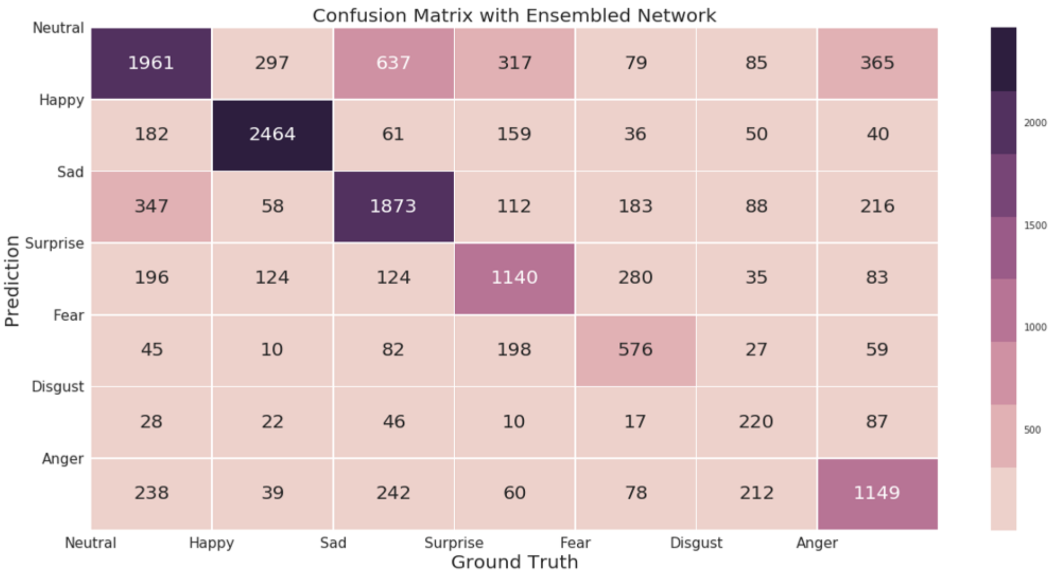
As mentioned in the Data Sources section, Affectnet data was used to the train the emotion prediction model. Images were considered in their RGB format and were resized to 96X96. A random sample of around 75K images was considered for this experiment which was split into 80%-20% for creating training and testing datasets. Training data was again split in the same ratio to create the validation dataset. In order to ensure, this random sampling experiment is controlled , a seed was chosen which is kept constant for all the models built. Following were the model architecture built on Affectnet data:

- Alexnet
- ResNet
- VGGNet
- CapsuleNet
- Ensemble Network
- Dark Knowledge network using softmax predictions from AlexNet, ResNet and VGGNet
- Convolutional Autoencoders

Table 1 shows the accuracy, top-2 accuracy and number of trainable parameters for each of the above mentioned model architectures. It is notable that increase in size of the network is not considerably improving the performance of the emotion prediction. Alexnet with just 2.26 parameters is performing better than ResNet and just falling behind 2% when compared with VGGNet which has 46% more parameters. Capsulenet is lagging behind the best model by a margin of almost 12%. Capsulenet was designed to capture the structural relationships between different parts of the image. Since variations in emotions are very minute, we found that capsulenet is finding it difficult to capture them.

Model	**Accuracy**	**Top-2**	**Parameters (in million)**
AlexNet	59.09%	79.24%	2.26
VGGNet	61.40%	81.38%	3.3
ResNet	56.72%	76.58%	11.18
CapsuleNet	50.17%	65.38%	16.25
Representational Autoencoders	32.56%	48.30%	0.7
Ensemble Network	62.40%	82.05%	16.74
Dark Knowledge	62.04%	82.01%	1.5

Table 1: Comparison of Accuracy, Top-2 Accuracy and no. of Trainable parameters across the Model architectures.

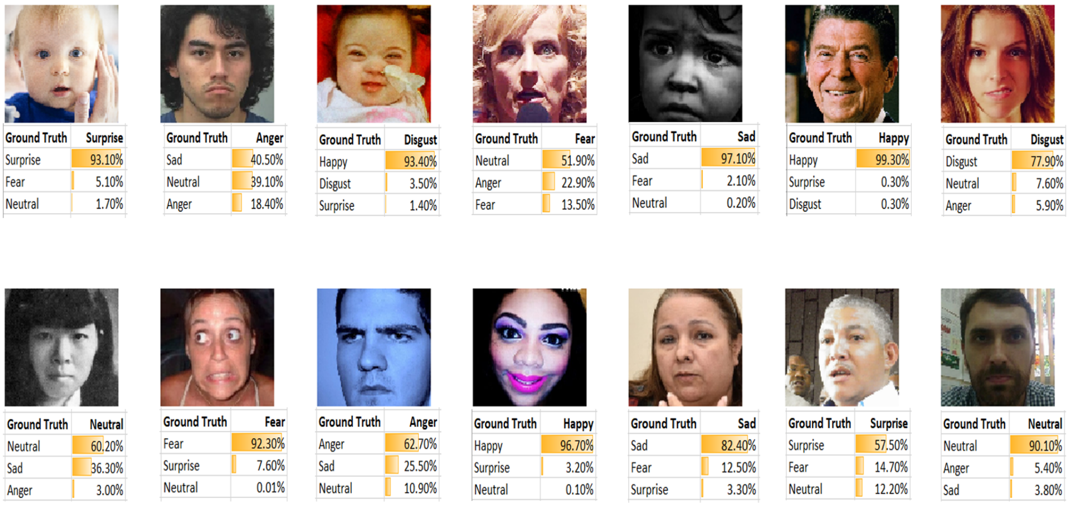


Prediction \ Ground Truth	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
Neutral	1961	297	637	317	79	85	365
Happy	182	2464	61	159	36	50	40
Sad	347	58	1873	112	183	88	216
Surprise	196	124	124	1140	280	35	83
Fear	45	10	82	198	576	27	59
Disgust	28	22	46	10	17	220	87
Anger	238	39	242	60	78	212	1149

Figure 7: Confusion Matrix from the Ensembled Network model

One of the observations from this experiment was from the Dark knowledge network that was built by using the softmax labels obtained from an ensemble of AlexNet, VGGNet and ResNet. The softmax labels were combined with the original one hot encoded ground truth and a smaller network was trained that had just 1.5 million parameters. This had the least number of parameters among the models built and almost 91% less parameters than the best performing ensemble network(Alexnet+VGGNet+ResNet). This network was giving an accuracy of 62.04% which is just behind a few decimal points compared to the ensemble network. The main objective behind building this network was to ensure that such models could be deployed in scenarios where real time predictions could be done with less time and computation.

Another notable observation was that the top-2 accuracy of the models were almost on an average 20% more than the accuracy. When seeing the confusion matrix (as shown in Figure 7), it is evident that the model is getting confused between neutral and other emotions owing to the subtle variations among them. Also the model is getting confused between anger and disgust emotion.



Ground Truth	Surprise
Surprise	93.10%
Fear	5.10%
Neutral	1.70%

Ground Truth	Anger
Sad	40.50%
Neutral	39.10%
Anger	18.40%

Ground Truth	Disgust
Happy	93.40%
Disgust	3.50%
Surprise	1.40%

Ground Truth	Fear
Neutral	51.90%
Anger	22.90%
Fear	13.50%

Ground Truth	Sad
Sad	97.10%
Fear	2.10%
Neutral	0.20%

Ground Truth	Happy
Happy	99.30%
Surprise	0.30%
Disgust	0.30%

Ground Truth	Disgust
Disgust	77.90%
Neutral	7.60%
Anger	5.90%

Ground Truth	Neutral
Neutral	60.20%
Sad	36.30%
Anger	3.00%

Ground Truth	Fear
Fear	92.30%
Surprise	7.60%
Neutral	0.01%

Ground Truth	Anger
Anger	62.70%
Sad	25.50%
Neutral	10.90%

Ground Truth	Happy
Happy	96.70%
Surprise	3.20%
Neutral	0.10%

Ground Truth	Sad
Sad	82.40%
Fear	12.50%
Surprise	3.30%

Ground Truth	Surprise
Surprise	57.50%
Fear	14.70%
Neutral	12.20%

Ground Truth	Neutral
Neutral	90.10%
Anger	5.40%
Sad	3.80%

Figure 8: Ground truth and ensemble model predictions for random test images

Figure 8 shows the ground truth and ensemble model predictions for random test images. It can be seen from the predictions that the model is confused between certain emotions(Sad vs Neutral, Anger vs Sad) as discussed above.

7. Conclusion & Future Work

It can be concluded that, the performance of a Deeper network (VGGNet) was not significantly greater than a shallow network (Alexnet). The smaller net with Dark Knowledge was performing similar to the larger net with 91% reduction in network parameters. The Top-2 Accuracy was almost 20% greater than the top-1 accuracy. Thus, Fine-grained classification models can be employed to capture subtle differences between expressions in the future.

Bibliography

[1] Kathrin Kaulard, Douglas W Cunningham, Heinrich H Bülthoff, and Christian Wallraven. The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PloS one*, 7(3):e32321, 2012.

[2] Market value of visual detection.

[3] Dan Duncan, Gautam Shine, and Chris English. Facial emotion recognition in real time.

[4] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[5] Prudhvi Raj Dachapally. Facial emotion detection using convolutional neural networks and representational autoencoder units.

[6] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[7] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–370. IEEE, 2005.

[8] Dennis Hamester, Pablo Barros, and Stefan Wermter. Face expression recognition with a 2-channel convolutional neural network. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.

[9] Siyue Xie and Haifeng Hu. Facial expression recognition with frr-cnn. *Electronics Letters*, 53(4):235–237, 2017.

[10] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] raghakot. Residual networks implementation using keras-1.0 functional api.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[15] XifengGuo. A keras implementation of capsnet in nips2017 paper ”dynamic routing between capsules”.