ELL888: Assignment 2
Deep Models for Spatial Data

Raunak Gautam 2015EE10467
Harsh Maheshwari 2015EE30517
Ayush Joshi 2015EE30509

1. Abstract:

Deep Neural Networks are highly complex and highly achieving models that have achieved state of the art performance on image classification. Training such deep nets require huge datasets like ImageNet. The goal of this assignment is to get acquainted with deep neural architecture, transfer learning and spatially distributed data, particularly images and hence to classify videos.
To unveil what a neural net learns, various visualisation methods have come up recently and they help us understand a deep neural net better. Through this assignment we also got acquainted with those methods.

2. Introduction:

Problem Statement: Frame-wise identification of dominant speaker in a video given the set of possible speakers.
We approached this problem as human recognition. We did so by employing 3 models (4 layer CNN, fine tuned Inception-net[1] and pre-trained Facenet[2]) and modified them to get good classification.To understand what is being learnt inside the neural network's black box, we performed various experiments to visualise how our neural networks see the world and what they interpret.
Since we had the real world data, preprocessing it to extract relevant information to be learnt was a big challenge. Dealing with noise in training and testing phases was other major challenge.

3. Dataset:

Frames extracted from different short videos of 6 speakers (classes) with variability in his/her attire, location and scale on screen, frame background, lighting and occasional shots of audience and some frames without humans (label noise).
Training Set: ~2000 data points per class
Validation Set: ~500 data points per class
Test Set: 3560 frames similar to Training set
(Test Set 0: 3560 frames, probably perturbed with noise/ some filters)

3.1 Data Preprocessing

The problem of human recognition can further be reduced to a face recognition task as it is the inherent feature of importance rather than the subject's attire or video background.
From the frames, we detected faces using HOG classifier[3] and Dlib's CNN face detector[3] (better performing). Largest face was cropped and aligned using eyes and lower lips as landmarks for uniformity across dataset.

Fig 1. Actual frame and the face aligned after extraction using Dlib CNN face detector.

4. Human Classification: Approaches And Results

4.1 Learning from scratch: 4 layer CNN

Model:

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 222, 222, 32)	896
activation_1 (Activation)	(None, 222, 222, 32)	0
max_pooling2d_1 (MaxPooling2	(None, 111, 111, 32)	0
conv2d_2 (Conv2D)	(None, 109, 109, 32)	9248
activation_2 (Activation)	(None, 109, 109, 32)	0
max_pooling2d_2 (MaxPooling2	(None, 54, 54, 32)	0
conv2d_3 (Conv2D)	(None, 52, 52, 64)	18496
activation_3 (Activation)	(None, 52, 52, 64)	0
max_pooling2d_3 (MaxPooling2	(None, 26, 26, 64)	0
flatten_1 (Flatten)	(None, 43264)	0
dense_1 (Dense)	(None, 64)	2768960
activation_4 (Activation)	(None, 64)	0
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 6)	390
activation_5 (Activation)	(None, 6)	0

Total params: 2,797,990

The model was trained on two datasets, full images and the cropped faces generated by preprocessing. Validation accuracy achieved on these two datasets was ~98% and ~90% respectively.

4.2 Transfer Learning: Inception Net V3

For cropped faces, base model of InceptionNetV3 was used excluding the top fully connected layers. To this base model we added our batch of fully connected layers and trained. After that top 2 inception blocks were fine tuned. Validation accuracy achieved was ~96%.

4.3 Pre-trained Model: Facenet

128-dimensional embeddings were generated for each preprocessed data point using Inception Resnet V1 with triplet loss function trained on LFW dataset. Multiple embeddings were generated by random augmentation of input images with each pass. A linear SVC was trained on the embeddings to give a validation accuracy of 97.30%.

4.4 Comparison of Approaches

To classify the 7th class (junk) we put a threshold on the maximum probability of classification and tuned the threshold.

The best accuracy we achieved on test data is **61.96%** on first test data and **62.7%** on second test data.

Performance of other models is as follows:

1. 4 layer CNN
    1. Complete images: 46.43% and 47.18%
    2. Cropped faces: 48.17% and 50.78%
2. InceptionNetV3
    1. Cropped faces: 56.03% and 61.04%

respectively on first and second test data.

5. Visualisation

5.1 4-layer CNN

To visualise inputs that maximally activate the filters of the our architecture, we used gradient ascent. Fig2.1 and 2.2 show the generated images for the last layer for the CNN trained for cropped and complete images.

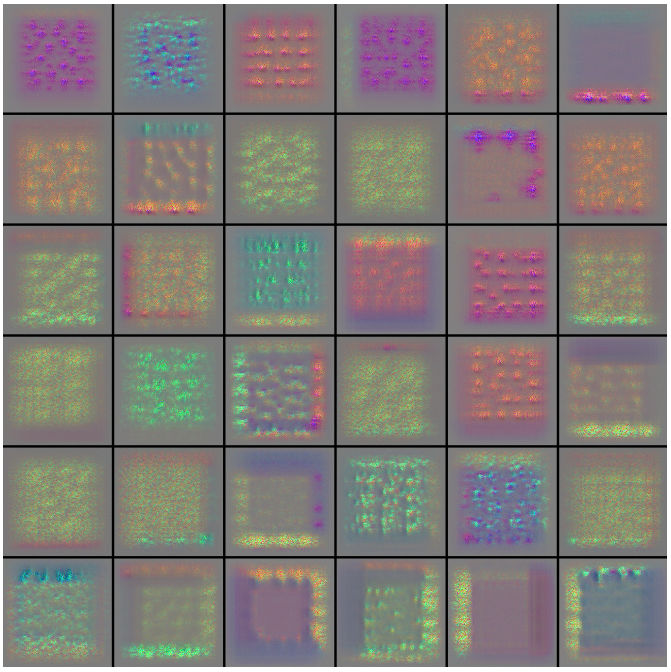


Fig 2.1

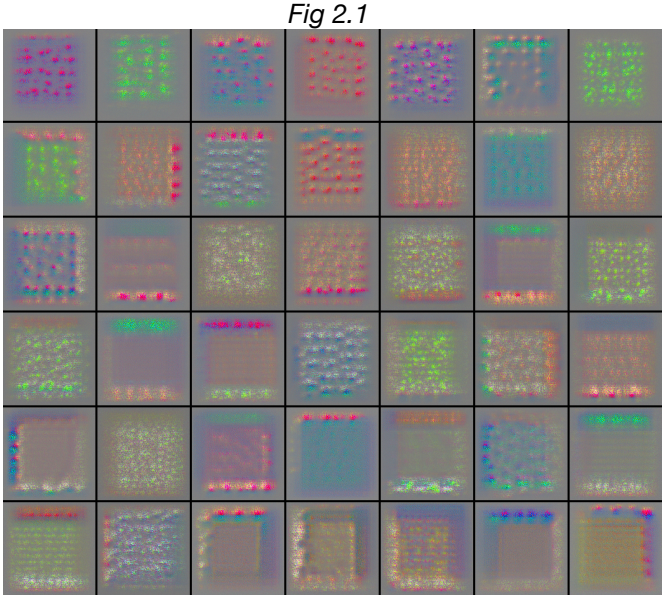


Fig 2.2

Fig 2.1 and 2.2 show the images that maximally activate filters of last convolutional layer of CNN trained on cropped and complete images respectively.

To understand what our network learns from complete images, we performed occlusion experiments[4] by occluding a part of image and then checking its probability of being classified to its class.

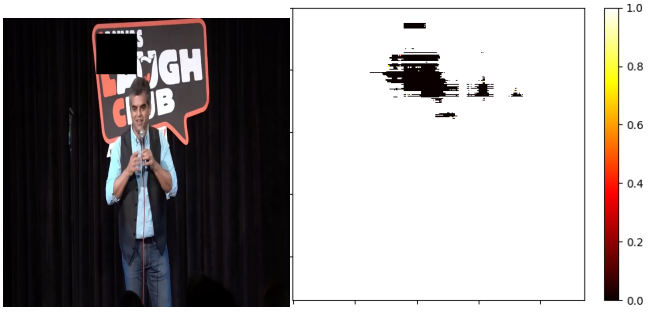


Fig 3.1

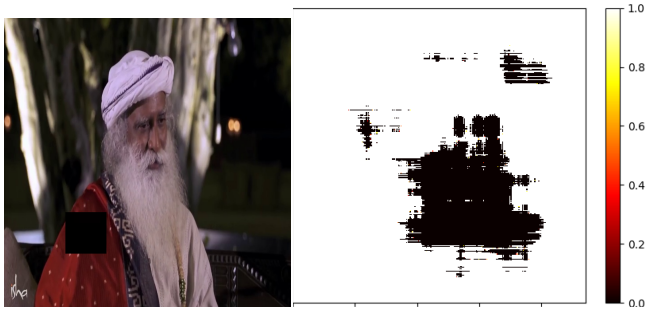


Fig 3.2

Fig 3.1 and 3.2 show results of occlusion experiments. The heat-map values are the probability of classification after occluding that region

We can see that for complete images, the network is probably learning the background in case of Atul Khatri and Sadguru's clothes or beard.

5.2 InceptionNetV3

5.2.1 Gradient Ascent-Filter Visualisation

Fig 4 shows results similar gradient ascent for inception net's deeper layer. One thing we can note is that the color doesn't play an important part. This means that deeper layer's don't learn the color but some abstract features.

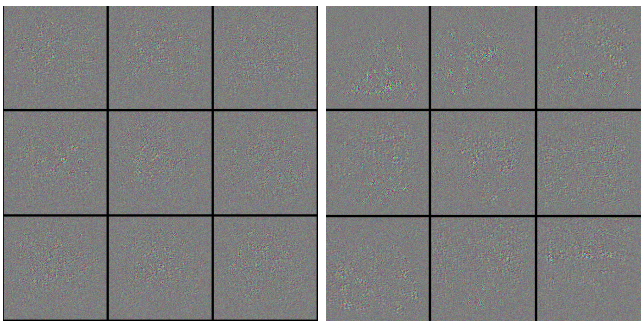


Fig 4.1 *Fig 4.2*

Fig4.1 and 4.2 show results of gradient ascent on InceptionNetV3 with fine tuning for convolutional layer 48 and 93 respectively. Please zoom in to see clearly.

5.2.2 Occlusion Experiments

We also performed occlusion experiments for cropped images for InceptionNetV3 to know which part is important for this net to classify a face.

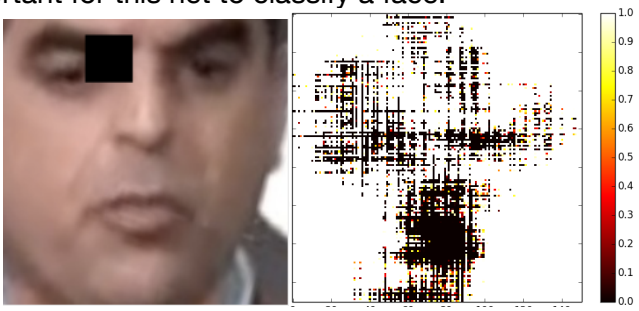


Fig 5.1

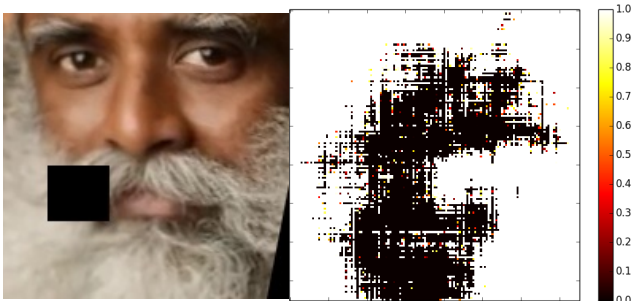


Fig 5.2

Fig 5.1 and 5.2 show results of occlusion experiments. Heat-map values are probability of the image being classified to its class after occluding that region.

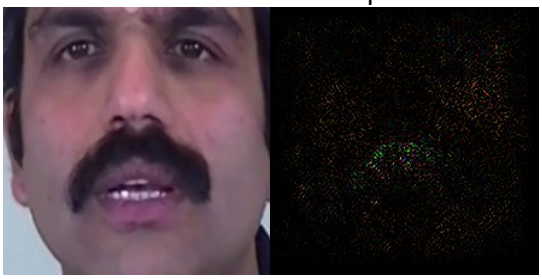
We can see from Fig 5 that there are some areas which are certainly more important than others. In general cheeks are not important to classify an image, but regions near eyes and nose are much more important and occluding them reduces probability drastically.

5.2.3 Gradient Visualisation with backpropagation

Backward pass of the activation of a single neuron after the forward pass from the network gives the gradient of activation with respect to image. The resulting image gives the areas of input images which cause maximum change in the activation of given neuron.

Class Saliency maps[5]:

Given an input image and the class label, we find the importance of pixels in determining the class i.e. gradient of output w.r.t input pixels. As we can see, moustache of Flute Raman is important.



5.2.4 Inverted Image Representation[6]

Starting from white noise image, generate an image minimising the loss between representations of generating and the target image at a particular layer of the network. This is essentially finding the optimal input which minimises loss for a layer given the model and output, which is inverse of generating optimal output given the input and model.

Layer 2 Layer 5 Layer 11

Layer 2 Layer 5 Layer 11

Fig 6: Inverted image representations of images belonging to class 2 (SG) and 5 (SK)

As visible from the figure, the inverted representations become more representative as we go deeper into the layer. The lumination information, which is not of importance in this example, is being diminished and the face color become more uniform. Similarly, Sadhguru's beard becomes more prominent.

5.2.5 Style Transfer

While inversion, if we start from a image other than noise, across iterations we will get images which will be a 'mix' of target and original image. The higher-level 'abstract' features which represent the class are transferred to other image. This can be another approach to transfer style.

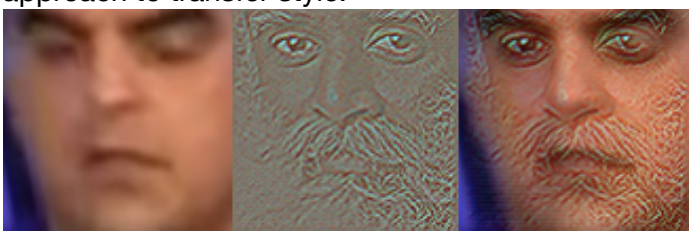


Fig 7. Transferring features of Sadhguru to Atul Khatri Sadhguru opened his eyes (Pun intended)

5.2.6 Deep Dream[7]

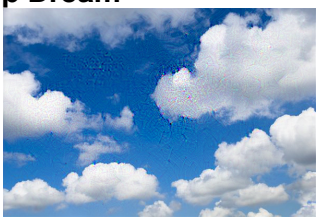


Fig 8 Deep dream (Zoom to view)

Visualising filters starting from a non-noise image gives an image with

5.3 Facenet

5.3.1 Dimensionality Reduction

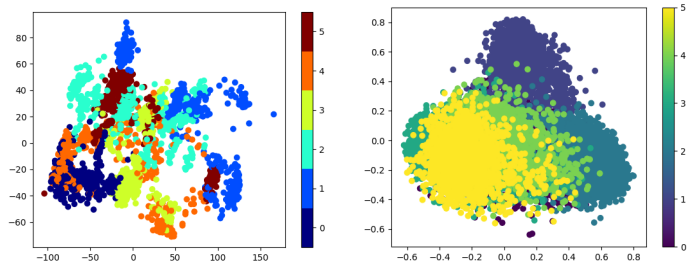


Fig 8.1 2-component PCA representation of complete images and embeddings from Facenet

PCA representations of raw images are scattered throughout the space. This is due to large variability in frame backgrounds which dominate clustering to give poor generalisation results. PCA of cropped images has less intraclass separation across all the videos and thus, better results on unseen data.

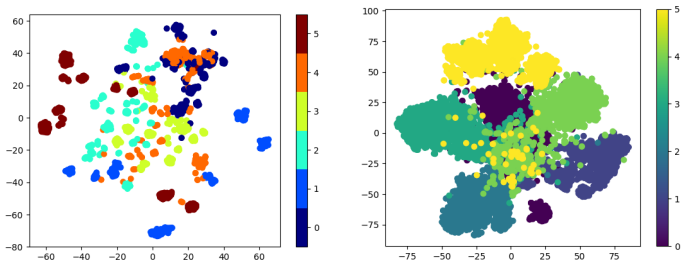


Fig 8.2 2-component t-SNE representations of complete images and embeddings from Facenet

From the t-SNE representations of cropped data, 6 distinct clusters are visible which come from the triplet loss function of Facenet which maximises interclass separation and minimises intraclass.

5.3.2 Euclidean distance across embeddings

Corresponding to one image, we found images corresponding to maximum and minimum euclidean distances in the space of 128 dimensional embeddings.

Fig 9 Distances of various images from the center image in the space of embeddings

It can be observed that the images of same person are closer than that of other people, irrespective of pose and luminance variances.

6. Conclusion

The best performing model is pertained FaceNet with SVM classifier in the end for face recognition. InceptionNetV3 is also performing comparably with fine tuning on imagine weights. The visualisation experiments help us in keeping faith in our neural network as we can now say with confidence that the neural network is learning something meaningful and something which we expected it to learn.

7. References

[1] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[2] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Dlib, http://dlib.net/python/index.html
[4] Matthew D Zeiler, Rob Fergus: “Visualizing and Understanding Convolutional Networks”, 2013

[5] K. Simonyan, A. Vedaldi, A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, https://arxiv.org/abs/1312.6034

[6] A. Mahendran, A. Vedaldi. Understanding Deep Image Representations by Inverting Them

[7] Google Research Blog: Inceptionism: Going Deeper into Neural Networks, June 17, 2015