Homework 3

Submitted by Vishantan Kumar
April 15, 2024

Part 1: Short Answers

Ans 1. A DenseNet is a kind of CNN that involves connected each layer to every other layer in a feed-forward fashion. That is, each layer receives additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers.
Advantages: Better feature propagation, feature reuse, enhanced flow of gradients. They are also much less prone to overfitting
Disadvantages: Increased computational complexity due to an increase in number of layer inputs, potential for redundancy due to reuse of features. Can also have scaling issues due to the inter-woven nature of the network

Ans 2. In an inception module, the network is allowed to choose from an array of different convolutional filters of different size within the same layer. It works by inputting the outputs from all the filters into the subsequent layer, allowing the network to automatically select the best fit for the result.
Advantages: Multi-scale processing allows for the model to process various contexts simultaneously. This also allows use of 1x1 kernels for dimensionality reduction while the diverse paths give better generalisation.
Disadvantages: Complex design and limited explain-ability of the model. Increased compute cost due to additional inputs as well as tedious optimization processes.

Ans 3. 3 loss functions that can be used to train VAEs are as follows:

Cross Entropy: One of the most widely used Loss functions for classification problems. Cross entropy loss is defined as $-\Sigma_{i=1}^{n} x_i log(\hat{x_i})$. Particularly useful when the input refers to the probability of belonging to a certain class

Mean Squared Error: Particularly useful for continuous input data and is defined as MSE = $\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \hat{x_i})^2$. t assumes that the reconstruction error is normally distributed.

Kullback-Leibler Divergence: Also known as KL Divergence. This can also be used as a regularisation term and measures the divergence between 2 probability distributions. It is defined as $\Sigma p(x) \log \frac{p(x)}{q(x)}$. This is useful in ensuring that the encoded distribution approximates the target distribution.

Ans 4. A sparse autoencoder is a variant of AEs that enforces a sparsity constraint on each hidden layer, encouraging the network to learn input representation via only a few neurons at a time. This sparsity can be enforced using a regularization technique during the training process which encourages the activity of the neurons in the hidden layer to be close to zero.

Ans 5. Denoising Auto Encoders (DAEs) are a variation of AEs that take in a noisy image as an input and aim to reconstruct the undistorted image at the output. By training on noisy data and attempting to recover the original undistorted data, denoising autoencoders can learn robust features and perform effective denoising. Typical loss functions used are MSE or Bianry Cross Entropy.

Ans 6. Image segmentation is a broad term which refers to the practice of partitioning an image into subsets which follow a specific principle. The goal of segmentation is to simplify or change the representation of an image into something that is more meaningful and easier to analyze.

Semantic Segmentation is a specific type of segmentation wherein each pixel is assigned to a predefined category for e.g. - Does a pixel belong to a road or the sky. It is not just object detection but also includes delineating a boundary.

Instance segmentation is another type of segmentation that can distinguish between different instances of the same category, i.e. each object is identified and segmented as a separate instance. For example, counting the number of people in an image, differentiating between 2 cars on the road etc.

Ans 7. 3 Major approaches for object detection in computer vision are as follows. We're not discussing any specific model architecture since they have been dealt with in subsequent problems:

a. Two stage detectors: This approach divides the object detection process into two distinct stages. The first stage focuses on generating regions of interest (proposals) where objects might be located, and the second stage classifies the content of these regions and refines their bounding box coordinates. This method is generally more accurate, as it carefully analyzes each region for potential objects. However, it tends to be slower due to the sequential nature of proposal generation and object classification. Exmaple - Faster-RCNN

b. Single stage detectors: These simplify the object detection workflow by eliminating the proposal generation step. Instead, they directly predict bounding boxes and class probabilities across the whole image in a single pass. They

are faster than two-stage detectors because they eliminate the need for a separate proposal generation step. However, they can be less accurate Example - YOLO

c. Transformation based detectors: This approach involves transforming the problem of object detection into a form where known algorithms can be applied efficiently, such as transforming it into a segmentation problem. Another transformation approach involves using anchor boxes of various shapes and sizes as references at different positions across the image. Each anchor box is adjusted to better fit the objects in the image. Examlple - Mask-RCNN

Ans 8. Non Max-suppression is a post processing technique in object detection used to eliminate duplicate detection by selecting the best bounding box out of all the possible bounding boxes around the object. It works on the principle of score-sorting wherein all proposed bounding boxes are first sorted based on their confidence scores — the likelihood that they contain the target object. The box with the highest score is added to a final list of bounded boxes and the process is repeated till all possible boxes have been extracted.

Ans 9. RCNN or Region Based CNNs are primarily used for object detection. It has a few variants such as Fast-RCNN and Faster-RCNN:

RCNN: It uses selective search to generate candidate bounding boxes which are then processed individually by a CNN to extract features which are then fed into an SVM. This can be slow due to the process of re-running the CNN for each bounding box

Fast-RCNN: This improves upon RCNN by using a CNN at the first step to extract the feature maps from the image. These maps are then used to generate the candidate bounding boxes using fully connected layers and probability maps. Though faster than RCNN, it can still be bottlenecked by the selective-search to generate the region proposals.

Faster-RCNN: This method introduces as a new Region Proposal Network (RPN) which inputs the convolution features into a detection network. The RPN predicts object bounds and objectness scores at each position simultaneously following which the high-scoring proposals are fed into the same ROI pooling layer as in Fast-RCNN. Despite its effective speed, the model can be fairly complex due to the dual functionality of sharing features.

Ans 10. YOLO or You Only Look Once is an object detection algorithm that diverges from the other traditional bounding box methods in that it frames object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. YOLO uses a single CNN to predict multiple bounding boxes and class probabilities simultaneously by dividing the image into a $S \times S$ cell grid with each cell predicting multiple bounding boxes and the likelihood that each box contains a specific class. YOLO has been seen to be extremely fast but can struggle with detection of small or grouped objects.

Ans 11. There are several metrics that can be used to check the performance of an object detector-

Precision, Recall and F1 Score: Precision is the accuracy of Positive Predictions, Recall is the ability of the detector to identify true positives while F1-Score is the Harmonic Mean of the two and can be used to balance both

IoU or Dice Coefficient: These measure the degree of overlap between 2 bounding boxes. Detectors can sometimes use an IoU threshold of 0.5 to classify a detection as a True Positive.

Processing Speed: Very useful in situations requiring real-time data processing. In such situations, detectors which process a higher number of frames per second are preferred over a more accurate but slower detector.

Part 2: Methods

2.1 Pre-processing

We used the LFW-Pairwise dataset for this, which is a widely used verification protocol. It consists of 2200 training and 1000 testing pairs, in which each pair could either display the same person or different people. These can then be used to classify wether a given pair consists of the same people or not. The first step is pre-processing the image and modifying the models to fit the classification task. The following steps have been performed for this:

a. **Image Resize:** Images were first re-sized to 256×256 pixels to ensure uniformity in resolution and size.

b. **Center Crop:** The resized images were cropped around the centre to fit 224×224 to ensure compatibility with VGG and AlexNet since they work best with these input resolutions

c. **Tensorization:** Each image set is then converted to a Torch-Tensor to add pytorch compatibility and GPU hardware acceleration.

d. **Normalization:** Each tensor is further normalised to set their means and standard deviations in accordance with the ImageNet weights - `[0.485, 0.456, 0.406]` for means and `[0.229, 0.224, 0.225]` for standard deviations.

2.2 AlexNet

The first model used in the analysis is the AlexNet. The pre-trained model with ImageNet weights was imported from torchvision [Figure 1]. Performance comparison was performed for the model with and without fine-tuning.

We dropped FCL-8 layer that maps the 4096 features to 1000 which correspond to each image net class probability.

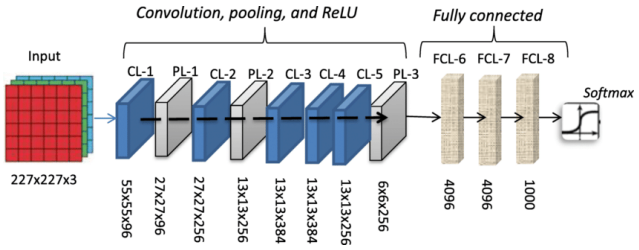

Figure 1. AlexNet Architecture

Hence, the outputs from FCL-7 are used for the distance matching via various algorithms

2.3 VGG Net

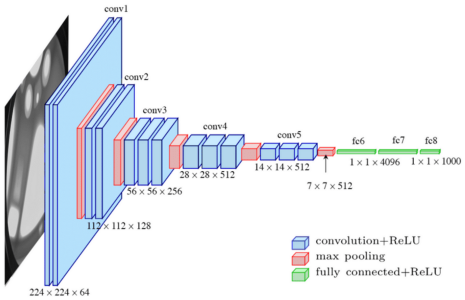

Figure 2. VGG Net Architecture

The second model was the VGG-16. The pretrained model with initial weights from ImageNet was similarly imported from torchvision and performance comparision was performed with and without finetuning. For feature extraction the last layer and softmax were again removed to output feature 4096×1 feature maps which can then be used for comparative study.

2.4 Fine-Tuning vs No-Tuning

Fine tuning was performed for both models by the following method:

- A siamese network was initialise to output feature maps from both the images in the input pair which were then compared using various distance metrics - L2 Norm and Cosine Similarity
- Training was performed for 3 epochs, with a batch size of 32 and learning rate of 5×10^{-5} on an RTX 3060 6GB.
- ROC Curves were plotted for models with and without fine-tuning for various distance metrics

Part 3: Results

3.1 Pre-trained Models

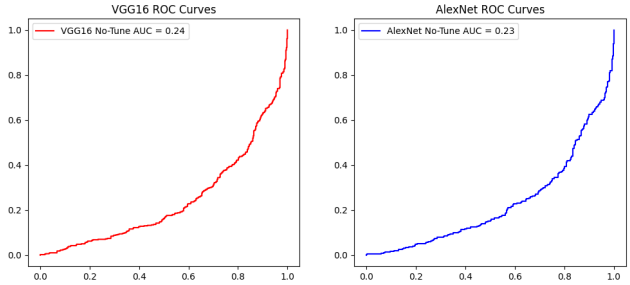

Figure 3. ROC Curves of Pre-trained Models (Euclidean Distance)

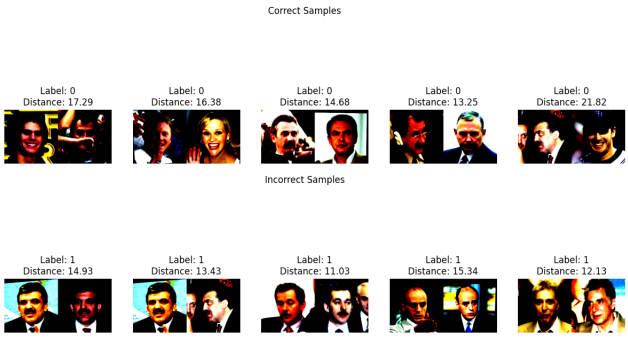

Figure 4. Euclidean distance of True Positive and False Positive

Figure 3 presents the ROC curves of both VGG 16 and AlexNet without fine-tuning LFW. For this, the distance metric used was the L2-Norm (Euclidean Distance). We can see that the ROC is fairly poor and lies below the threshold of FPR = TPR with an $AUC < 0.5$. The **respective accuracies were, however, 0.5.** Further analysis revealed some interesting observations. From **??** we can see that the distance of samples for well-classified and mis-classified samples is the same. Hence, we can probably say that all-data points are being classified as 0 and since the data is balanced, the accuracy is 50% even though the AUC is less.

3.2 Fine-Tuned Models

Fine tuning was performed for both models for 3 epochs and the following ROC curves were plotted

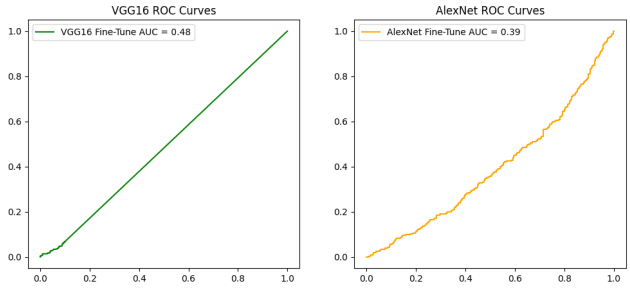

Figure 5. ROC Curves of Fine-tuned Models (Euclidean Distance)

The ROC curve for the fine-tuned models was mich better than the pre-trained models with a higher AUC. The AUC,

however, was still fairly less. The accuracy was a bit better at 52%. We can also see that the AUC for VGG is higher than AlexNet, which could be an outcome of the higher model size and complexity of VGG.
We then compared the AUC when a different distance metric was used. Cosine Similarity was used as the comparison algorithm for fine-tuning and ROC curves were plotted:

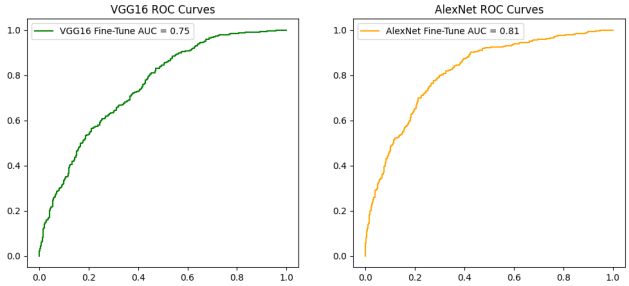

Figure 6. ROC Curves of Fine-tuned Models (Cosine Distance)

We can see that the ROC for Cosine Similarity is much better than Euclidean distance with a higher AUC. This time, however, the AUC for AlexNet was higher than that of the VGG16. This shows that cosine similarity is a better distance metric than Euclidean distance since the ROC curves for Euclidean distance were straight lines with a much lower AUC.

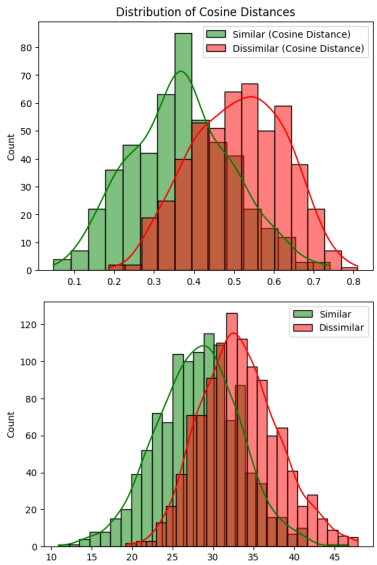

Figure 7. Cosine and Euclidean Distances for fine-tuned features

To further analyse this, we plotted the distribution of the cosine and euclidean distances for features extracted from the fine-tuned models. We can see that both are fairly similar but since the cosine distance is a bit more seperated than euclidean distance, it provides a better seperation and model performance.