CS 783A: Visual Recognition (End-Semester Examination)

Vinay P. Namboodiri

30 April 2019 (9 am - 12 noon)

Total Number of Pages: 8

Total Points 30

Instructions

1. Read these instructions carefully.
2. Write you name, section and roll number on all the pages of the answer book, **including the ROUGH pages**. You will be penalised if you fail to write the name, roll number and correct section.
3. Write the answers cleanly in the space provided. Space is given for rough work in the answer book.
4. Using pens (blue/black ink) and not pencils. Do not use red pens for answering.
5. Do not exchange question books or change the seat after obtaining question paper.
6. Even if no answers are written, the answer book has to be returned back with name and roll number written.
7. Sign the attendance sheet.
8. No clarifications will be provided. Make suitable assumptions and specify your assumption in the paper.

Question	Points	Score
1	5	
2	10	
3	3	
4	3	
5	3	
6	3	
7	3	
Total:	30	

I PLEDGE MY HONOUR THAT DURING THE EXAMINATION I HAVE NEITHER GIVEN NOR RECEIVED ASSISTANCE.

.....

Signature

Enter Name and Roll Number Above

Question 1. (5 points) **Answer the following questions very briefly in one or two sentences**

(a) How does unsupervised segmentation for proposal generation ensure diverse generation of proposals?
Answer:

> **Solution:** Selective search includes diversity in proposal generation mainly by considering complementary color spaces ((1)RGB,(2) the intensity (grey-scale image)I, (3)Lab, (4) the rgb channels of normalized RGB plus intensity denoted as rgI, (5)HSV, (6)normalized RGB denoted asrgb, (7)C which is an opponent colour space where intensity is divided out, and finally (8) the Hue Channel H from HSV), B) complementary similarity measures (color similarity, texture similarity, size similarity and fill similarity). They also use complementary similarity measures and complementary starting regions.

(b) Why are we not able to fine-tune SPP net on the Pascal VOC dataset in an end to end manner?
Answer:

> **Solution:** The SPP net based proposals are cached. In SPP net, the layers upto conv5 are obtained, the spatial pyramid net based pooling features are obtained and this is cached. Further only the FC6-7-8 layers are fine tuned for the dataset. The previous layers are just based on pre-trained imagenet weights. This is because, the SPP net had no provision to backpropagate the gradients through spatial pyramid pooling layer. This was addressed in Fast RCNN that proposed backpropagation through the ROI pooling layer.

(c) How does SLIC pixels achieve oversegmentation by not having pixels from multiple objects in the same segment?

Answer:

Solution: This is achieved by perturbing cluster centers in an nneighborhood, to the lowest gradient position.

(d) What is atrous/dilated convolution?

Answer:

Solution: Atrous/Dilated convolution is obtained by adding holes in a convolutional kernel. It increases the receptive field without increasing the number of parameters and is used in segmentation. It is possible to show a figure for the same.

(e) In unsupervised domain adaptation by backpropagation, the authors use two classifiers. In what way is the trade-off between the classifiers achieved?

Answer:

Solution: The tradeoff between the classification and domain discrimination is decided by adding classification loss as is for backpropagation and using a weighing term -λ that decides how much weight to be given to the gradients from the domain discriminator that are also reversed using the negative term while backpropagating through the feature extraction layers.

Question 2. (10 points) Answer the following questions briefly

(a) Consider that using context prediction based self-supervised learning method, you have trained your network. How would you verify that your network has

learnt a semantically meaningful representation?
Answer:

Solution: The way in which self supervised learning can be verified in various ways. One of the first ways in which it can be verified is by validating the nearest neighbors and ensuring that the feature is able to learn semantically meaningful nearest neighbours. For instance, nearest neighbors for cats should be cats and for faces should be other faces. This was used for instance to know that chromatic aberration was misleadingly able to obtain nearest neighbors just based on the location rather than the semantic closeness or feature similarity. The other ways is by using it for geometric data mining (where you consider the geometric spatial layout while retrieving and use it for discovering categories). One more way it can be checked is by transfer learning for other tasks such as object detection where you compare it with features obtained from pretrained imagenet. (explaining any one method should carry two marks)

(b) In UNET architecture how exactly does weight map contribute to improved accuracy?
Answer:

Solution: It forces the network to learn the small separation borders that we introduce between touching cells. Specifically it has costs to balance the class frequencies and costs based on distance to nearest and second nearest boundary pixels.

(c) In large scale detection through adaptation, explain the adaptation approach.
Answer:

Solution: For the known classes for which we have bounding box information it learns adaptation weights δ_{Bi}. For adapting it takes the k nearest neighbors as the categories for which it has minimal distance based on normalised fc_8 parameters for classification. This it then adds by taking the average of the δ_{Bi} for these set of k classes while adapting for a target bounding box.

(d) We obtain stereo images of a chessboard. In what way would the chessboard affect the estimation of depth in the image?
Answer:

Solution: For estimating depth we need to know the patch correspondence. Using epipolar constraint we would estimate the epipolar line. However, for a chessboard image, there would be multiple patches that could correspond to each other. Therefore patch correspondence based on just correlation would suffer unless we used constraints based on neighborhood to ensure that the patches were correctly matched.

(e) We would like to use a recurrent neural network to generate image captions. If we wish to generate a single caption using two images what architecture could we use?
Answer:

Solution: If we have multiple images for which a single caption needs to be jointly genenerated, we could either feed the images through an LSTM or feed the network through a siamese network to capture the relation between the images. This could then capture the common context for the two images that could be provided to a decoder LSTM to generate the images. Alternatively, average pooling of the CNN features could also be effective to obtain the common context for the two images for captioning.

Question 3. (3 points) How is Fast RCNN able to make the SGD steps efficient?
Answer:

Solution: It uses it to obtain multiple region proposals for a few set of images. This is obtained through hierarchical sampling where it first samples images and then samples a batch of proposals in the same image that are then fed to the SGD routine instead of sampling randomly at the proposal level that would imply sampling many proposals. It also ensures that the computation between the multiple proposals is shared. This is how SGD steps are made efficient

Question 4. (3 points) How does Yolo v1 using a 7×7 grid process the output when a) there is no object present in the grid cell and b) when the centre of the object lies in the grid cell
Answer:

Solution: Each grid cell predicts 2 bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. Formally we define confidence as Pr(Object) * IOU .
1.5 Mark: No object present If no object exists in that cell, the confidence scores should be zero. So the network reduces the confidence of all the 2 bounding boxes.
1.5 Mark: Object present Otherwise we want the confidence score to equal the intersection over union (IOU) between the predicted box and the ground truth. So the network increases the confidence of the bbox having higher IoU with ground truth and reduces the confidence of the other bbox. (optional) Also regresses the bbox such that it lies closer to the ground truth.

Question 5. (3 points) GANs have been used for many applications to approximate a distribution of data such as MNIST digits. In the process of training the GAN if there are not enough real samples for one of the digits, how would the training proceed? What would be the resultant effect?

Answer:

Solution: If enough samples for some digits are not obtained for some of the digits then the discriminator would be weak in classifying those digits as being weak and therefore the generator would be able to fool the discriminator for those samples. The generator would then probably undergo mode collapse and just generate those samples as it does not have incentive to generate all samples in the vanilla GAN architecture. Alternately, the generator could also not learn to generate the sample and the modes corresponding to those digits would be lost for the samples that do not have enough number of samples. In this case the GAN would undergo mode loss. Thus the GAN would undergo either mode loss or mode collapse.

Question 6. (3 points) In the method that adopts solving jigsaw puzzles for self supervision, explain three ways in which they avoid the CNN from avoiding learning a semantic representation (i.e. three ways short-cuts for the CNN are avoided)

Answer:

Solution: To avoid shortcuts due to edge continuity and pixel intensity distribution they leave a random gap between the tiles To avoid shortcuts obtained by matching mean and standard deviation of adjacent patches, they normalize the mean and the standard deviation of each patch independently To avoid shortcuts due to chromatic aberration they jitter the color channels and use grayscale image

Question 7. (3 points) What are the steps for detecting space time interest points in a

video?

Answer:

Solution: The steps for obtaining the space time interest points discussed in class are as follows: a) Constructing the scale space using Gaussian convolution kernel in space and time b) Compute the second moment matrix using spatio-temporal gradients c) Define positions of features by local maxima The expressions for these could also be provided