CS 698O: Visual Recognition (Mid-Semester Examination)

Vinay P. Namboodiri

23 September 2017 (1 pm - 3 pm)

Total Number of Pages: 4 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.

Question	Points	Score
1	5	
2	5	
3	5	
4	5	
5	5	
6	5	
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) In instance retrieval, the challenge is to avoid matching of background during retrieval. The background is characterized by common visual words. Explain two ways in which common visual word matching is penalized in instance retrieval

Answer:

Solution: In instance retrieval, the visual words are grouped into clusters. On grouping into clusters, they go through a stop word removal procedure. This is the first method to penalize common visual words. This is similar to removal of the common words in language that occur too frequently such as 'the', 'a' etc. The visual words are typically those words that occur too frequently in background and could be mistakenly matched while matching a query window.
The second technique used is that of using TF-IDF. In TF-IDF, the common visual words are inversely scored and while matching, they are less commonly considered.

Question 2. (5 points) In local feature representation using SIFT, scale invariance is obtained using difference of Gaussian scale space. Explain how we obtain rotation invariance in the feature representation.

Answer:

Solution: In this question, we have already obtained the feature with respect to the appropriate scale space. We want a rotation invariant representation of the feature. To obtain this, we identify the dominant direction of the gradient in the patch. We then rotate the patch so that the dominant direction of gradient is consistently aligned along a fixed direction such as horizontal direction or at 180 degree angle. We then describe the feature using a the four histogram of gradients for each window.

Question 3. (5 points) Consider that we have a specific set of 3 categories, 'bike', 'motorcar' and 'bus' that we would like to classify. Each image in the train and test set is guaranteed to have only one of the 3 classes present. The object in the class could be present in any location. If you were required to use a pyramidal bag of words representation, which representation would you choose? Explain the representation used and why it would allow for categorising the object irrespective of its position in the image.

Answer:

Solution: So far in the course we have considered feature based pyramidal representations and spatial pyramid representations. In this case, we would like to allow for an object to be matched irrespective of its location. This can be obtained through a feature based pyramid as it neglects spatial location. However, if an object is present in a location, we want the entities belonging to the object (wheel, body of car) to be grouped spatially. This can be obtained through a spatial pyramidal representation. Therefore, ideally we would like to have a spatio-feature pyramid representation which could be trained such that the classifier enhances only the spatio-feature representation that contains the object.

Question 4. (5 points) In ResNet architecture, there are identity functions that are used to skip the layers. Instead of identity function while skipping the layers, if we were to add a sigmoid layer in the path instead of identity, how would the network architecture change and what would it result in?

Answer:

Solution:
In this architecture, instead of identity functions, we would have sigmoidal layers that are learned end-to-end. A sigmoidal layer selectively switches on and off a path. This would imply that based on the representation, either the by-pass layer would switch on or off. In this case, the network would learn either to learn the residue of the function or to learn the whole function. This can be useful when for certain features, the function cannot be expressed as a residue of the original function. This has recently been proposed as a new architecture SE-Net

(

Question 5. (5 points) In Faster RCNN there could be overlap between proposals. How would backpropagation occur in the ROI pooling layer where the proposals could overlap? What would be the effect across scales in the ROI pooling layer?

Answer:

Solution: In backpropagation through ROI pooling, the approach is similar to max-pooling. However, there may be multiple proposals that have an IoU over 0.5 and are therefore positive. Others having an IoU less than 0.5 are negative. The error for the different proposals have to be backpropagated. However, this has to be done with respect to overlap. Now, in case of overlap, the overlap could be across scales of the ROI that are used in a spatial pyramid fashion. Appropriate care has to be taken that the appropriate scale of the ROI receives gradients corresponding to the overlap. It may be possible that the ROIs overlap only slightly. Only a few cells in the finest layer would overlap, but the appropriate gradient has to be propagated across the various scales.

Question 6. (5 points) While training an SSD object detection routine you are provided with the information that a random number of images have been rotated by +/- 90 degrees. However, you do not know which images have been rotated. What changes would you make in the SSD training pipeline in order to account for such a rotation?

Answer:

Solution: One solution is to modify the data-augmentation step to allow for rotated variants of the ground-truth bounding boxes for all instances. This would enable the object to be invariant to rotation.
A better option is to initially train the network with the supervision provided. Then to retrain the network by by rotating and obtaining the max-scoring inference that scores maximally. Same procedure can be used at test time. This would ensure that at test time only the appropriate rotated boxes are used.