Quizs

Wednesday, April 3, 2019 12:52 PM

1. Question 1

Why is the idea of having 1x1 convolutions reasonable?

- [ ] They act like L2 regularization, reducing overfitting by making weights smaller.
- [x] They act like dimensionality reduction, removing unnecessary feature maps from previous layer.
- [ ] They accelerate training by making loss function more convex.
- [ ] They accelerate inference by replacing fully-connected layers with convolutional layers.

Question 2

1 point

2. Question 2

How can one reduce computational burden suffered by the deep convolutional neural networks?

- [x] Use 1x1 convolutions to reduce number of feature maps.
- [x] Use stacked 3x3 filters to reduce the number of parameters in feature maps.
- [x] Use 3x3 filter decomposition into 1x3 and 3x1 filters to reduce the number of parameters in feature maps.
- [ ] Use Adam optimizer instead of vanilla SGD to accelerate learning.

Question 3

1 point

3. Question 3

Mark the correct statements.

- [x] Residual connections help back propagate errors in very deep networks, leading to better generalization and handling the vanishing gradient problem.
- [ ] Batch Normalization can help in CNNS, because the spatial dimensionality reduction makes covering larger parts of the input in higher layers possible.
- [ ] With stochastic depth, the network (expected) depth reduces during testing

while maintaining the full depth at training time.

- [ ] DenseNets are harder to train because of their complicated architecture.

Question 4

1 point

4. Question 4

Why do deep learning methods outperform everyone else in computer vision in most tasks?

- [x] Visual features are learned automatically and therefore focused on a specific task.
- [x] Neural networks allow us to recover the nonlinear and complex dependencies.
- [ ] Deep learning methods can be applied to any data set, as opposed to the classical ones.
- [x] Computer power has reached a level that allows you to solve optimization problems with a variety of parameters.

Question 5

1 point

5. Question 5

Check all methods of dealing with overfitting.

- [ ] Adding recurrent layers
- [x] Small random turns
- [ ] Increasing the optimization step
- [ ] Increasing resolution of images
- [x] Early learning stop
- [x] Dropouts
- [ ] Replacing the fully connected on convolutional layers
- [x] Regularization

Question 6

1 point

6. Question 6

Why can part localization be useful for fine-grained recognition problems?

- [ ] Parts are the only way to solve fine-grained classification tasks.
- [ ] It speeds up training of neural networks because they have to process little

data.

- [ ] Parts may have visual features extracted at their original resolution which helps focus on subtle appearance differences between them.
- [x] It allows focusing on differences associated with specific object parts which can be small relative to the whole image.

Question 7

1 point

7. Question 7

Which of the following are valid examples of image similarities?

- [x] Scene geometry similarity (geometrically similar scenes)
- [x] Color similarity (get objects of the same color)
- [ ] Caption similarity (get images with similar captions)
- [x] Instance similarity (get this very object)

Question 8

1 point

8. Question 8

For a local semantic hash of 10101111, which would be the closest neighbours of bit distance equal to 1?

- [ ] 10101100
- [ ] 10111110
- [x] 10101011
- [x] 00101111
- [ ] 10101111

Question 9

1 point

9. Question 9

How to combine advantages of k-means and LSH clustering into a unified indexing scheme?

- [ ] Cluster image descriptors using k-means, then quantize the very same descriptors and concatenate cluster index and LSH mask into a joint signature.
- [x] Cluster image descriptors using k-means, then compute LSH codes for the difference of original points and cluster centers using LSH

difference of original points and cluster centers using LSH.

- [ ] Compute long LSH codes for the original images, then cluster these using k-means.
- [ ] Just use k-means and LSH separately and see what works best.

Question 10

1 point

10. Question 10

Why do we need a preprocessing of the face image in the problem of face identification?

- [ ] To search for a person on the basis of photographs.
- [x] To reduce the impact of the diversity of human pose, angle, scale.
- [ ] To account for different types of camera.
- [ ] To account for the variability of the appearance of a person (make up, haircut).

Question 11

1 point

11. Question 11

What parts are used in CNN cascade for keypoints regression task?

- [ ] Generator and discriminator.
- [x] Initial (robust) and refinement models.
- [ ] Multi-task predictors for different keypoints.
- [ ] Predictors from different scales in pyramidal architecture.

Question 12

1 point

12. Question 12

Which method is the main one in the identification problem?

- [ ] Training of the classifier, compare the classification results.
- [x] Training descriptor, the comparison of distances between descriptors.
- [ ] Applications of finding similar individuals, a comparison of intersection results are similar.
- [ ] The prediction of attributes, comparison of the predicted attributes.

Object Detection

1. Question 1

Two rectangles R1 and R2 have left-up and right-down points A and B, C and D accordingly. Coordinates of points: A (0, 77), B (23, 26), C (15, 51), D (41, 0). Compute IoU of these rectangles in percents. Round answer to the nearest integer in percents.

Intersection = (23-15)*(51-26)=200

Union = 23*(77-26) + 51*(41-15) - 200 = 2,299

IoU = 200/2299 = 0.087

Answer: 9

2. Question 2

Consider face detector have Miss rate 0.40 for FPPI = 10^-1. We are working with dataset that has 5 faces on each image in average. What Precision and Recall corresponds to this parameters? Enter precision and recall values in percentages with space:

FPPI = FPR per image = FP/number of images(let N denotes) = 0.1

Then FP = 0.1N

Ground truth = Cond. P = 5N

Miss rate = FN/(TP + FN) = 1 - recall, so recall = 0.6

TP = 0.6*5N = 3N

Precision = TP/(TP + FP) = 3N/(3N+0.1N) = 3/3.1 = 0.9677

Answer: 97 60

3. Question 3

Consider constructing pyramid for sliding window object detection. We'd like to use window with size 20x20 pixels and find objects with size from 20 to 50 pixels. Images in pyramid are upscaled with factor 1.1 (i.e. by 10% each time). How many images (including original, not scaled image) are needed for this pyramid?

scaled, image) are needed for this pyramid?

Number of images scanned = number of scales * number of aspect ratios
20*(1.1)^10 = 51.8748
Answer: 10+1 = 11

4. Question 4

What data augmentation methods are useful for training HOG+SVM or VJ face detector?

Flip around vertical axis	☑
Flip around horizontal axis	☐
Rotate on big angle > 90 degrees	☐
Rotate on small angle < 20 degrees	☑
Random crop	☐
Small shifts: 1-3 pixels	☑

5. Question 5

We work with video of size 1024x768 pixels and 25 fps. We use sliding window object detector with window size 20x20 pixels and stride 2, 1 image scale (i.e. without pyramid). What should be false positive rate of the detector s.t. detector output false positive less frequently than 1 time per second? Round answer with 1e-07 precision.

(1024-20)/2 + 1 = 503
(768-20)/2 + 1 = 375
503 * 375 * 25 = 4,715,625
1/4,715,625 = 0.000000212

6. Question 6

Select correct sentences for R-CNN object detection architecture:

- ☐ Uses sliding window to obtain object position proposals
- ☑ Use selective search to obtain object position proposals
- ☐ Uses neural network to obtain object position proposals
- ☐ Uses HOG features
- ☐ Uses ROI pooling layer to compute features effectively for every sliding window
- ☑ Uses SVM for object classification
- ☐ Uses dense+softmax layers for object classification
- ☐ Has neural network with multitask loss

Has neural network with multitask loss

7. Question 7

1. Select correct sentences for Faster R-CNN object detection architecture:

Uses sliding window to obtain object position proposals	☐
Use selective search to obtain object position proposals	☐
Uses neural network to obtain object position proposals	☑
Uses HOG features	☐
Uses ROI pooling layer to compute features effectively for every sliding window	☑
Uses SVM for object classification	☐
Uses dense+softmax layers for object classification	☑
Has neural network with multitask loss	☑

8. Question 8

How many numbers will YOLO detector output per image if the model has 6 x 10 grid, every cell has 3 position hypotheses and there are 25 object classes in the training sample?

6*10 = 60 cells

3*(4+1)+25 = 40 each cell output

60*40 = 2,400

1. Question 1

Which of the following is an operation, not a task in computer vision?

Object detection	☐				
Perspective projection	☑				
Gradient computation	☑				
Instance segmentation	☐				

segmentation					
Image convolution	☑				
Max-pooling	☑				

★ **2. Question 2**

For a 3-class semantic segmentation problem, how many numbers must an algorithm output for a 640x480 image?
// 640*480*3=921,600

1
point

★ **3. Question 3**

Why is SLIC algorithm better suited to the image oversegmentation task than k-means method?

It utilizes a more robust distance metric, rather than simple Euclidean distance used in k-means method	☐	
It is more computationally efficient because segment sizes are bounded, limiting the number of pixels examined at each iteration	☑	
It limits distance between pixels by a certain threshold, utilizing the notion of hard spatial neighbourhood	☑	
Unlike k-means, SLIC is a supervised learning method and thus can use labels to improve segmentation	☐	

1
point

★ **4. Question 4**

What is the goal of the unpooling operation?

☐ To undo channel concatenation by decreasing the number of convolutional feature maps

☐ To undo convolution by applying the transposed convolution

☐ To help backpropagate errors by introducing sparse convolutions

☑ To undo pooling by outputting an image with larger resolution (i.e., pixels in spatial directions)

1
point

5. Question 5

In unpooling, how do we approximate the inverse of the non-invertible max-pooling operation?

- [x] We output maximal values at their respective indexes (called max location switches) and place zeroes elsewhere
- [ ] We do bilinear interpolation to compute the output
- [ ] We use ‘bed of nails’: output the maximal values in the top left corner and zeros elsewhere

1
point

6. Question 6

What is a Gram matrix in linear algebra?

- [x] A matrix produced by computing dot product between two sets of vectors
- [ ] A matrix of feature activations in a CNN
- [ ] A confusion matrix of CNN
- [ ] A positive-semidefinite matrix used to generate random numbers from a Gaussian distribution

1
point

7. Question 7

What makes a good generator for a GAN model?

- [x] It produces data that is hard to distinguish from real
- [ ] It achieves superior performance in generating Gaussian mixtures
- [ ] It produces nicely looking images

1
point

1. Question 1

Calculate the number of 25 fps FullHD RGB video channels that can be

simultaneously streamed through the 1 Gbit Ethernet LAN with 10x video compression ratio. Round the answer down to nearest integer

1920 * 1080 * 25 *3 * 8 = 1244160000 / 10 = 124416000

1e9/124416000 = 8.0376

8

2. Question 2

Which of these metrics may serve as performance measures for optical flow estimation?

- [x] Endpoint Error
- [ ] Average Precision
- [ ] Correlation between two vectors
- [x] Angular Error
- [ ] Detection Error Tradeoff curve

3. Question 3

Calculate Endpoint Error for two motion vector: Ground Truth = [1,1], Estimated = [2,0]. Specify 3 digits after comma.

2^0.5 = 1.4142

1.414

4. Question 4

In visual object tracking task, what does Equivalent Filter Operations metric measure?

- [ ] The number of convolutions required to achieve a specified tracking quality
- [ ] The number of feature maps required to produce an appropriate robustness for the tracker
- [x] The time required for tracking algorithm to run compared to the time

required for image filtering operation to run

5. Question 5

Which of these are types of errors that a multiple object tracker can suffer?

- [ ] False coverage error
- [ ] False acceptance error
- [ ] Mean absolute error
- [x] ID switch
- [x] False positive error
- [x] False negative error

6. Question 6

Compute MOTA score for a multiple object tracking method, which produces 530 detections, 50 false positive errors, 20 false negative errors, 30 ID switches on a dataset with 200 frames and 500 ground truth detections and 300 trajectories? Use at most one decimal precision places.

1 - (20 + 50 + 30)/(500)=0.8

7. Question 7

What is the effect of using re-identification on the tracking errors in multiple-object tracking methods?

False positives are decreased	[x]	[ ]
ID switches are reduced	[x]	[x]
Number of Mostly Tracked is increased	[ ]	[ ]
Number of Mostly Lost is increased	[ ]	[ ]
False negatives are decreased	[x] 0.6	[x] 0.8

Question 8

8. Question 8

Select correct statements regarding action classification.

To localize actions in videos we usually detect and track relevant objects first, and then apply action classification in a temporal window along the track.	[ ]	[x]

By explicit consideration of motion information in form of optical flow maps, point and keypoint trajectories, we can currently improve the performance of action recognition.	☑	☑
In dense trajectories with CNN features, point neighbourhoods are cropped from frames along the trajectory, concatenated into space-time volume along the trajectory, and then supplied to CNN for feature computation.	☐	☑
It is easy for convolutional neural network to extract and use motion information automatically, when applied to whole video volume.	☑ 0.5	☑ 0.5