Cairo University
Faculty of Computers and Artificial Intelligence
Final Exam

Department: Information technology	
Course Name: Computer Vision	**Date: 14/7/2020**
Course Code: IT 444	**Duration: 1.5 hours**
Instructor(s): Prof. Dr. Reda Al Khoribi Dr. Motaz El Saban	**Total Marks: 80**

تعليمات هامة

- حيازة التليفون المحمول مفتوحا داخل لجنة الإمتحان يعتبر حالة غش تستوجب العقاب وإذا كان ضرورى الدخول بالمحمول فيوضع مغلق فى الحقائب.
- لا يسمح بدخول سماعة الأذن أو البلوتوث.
- لايسمح بدخول أي كتب أو ملازم أو أوراق داخل اللجنة والمخالفة تعتبر حالة غش .

Q1 (True OR False) [1.5 marks]

If we initialize the k-means clustering algorithm with the same number of clusters but with different starting positions for the centers, the algorithm will always converge to the same solution. F

Q2 (MCQ) [3 marks]

You are using k-means clustering in color space (RGB) to segment an image. However, you notice that although pixels of similar color are indeed clustered together into the same clusters, there are many discontiguous regions because these pixels are often not directly next to each other. A possible solution to that problem:

(a) Run K-means one more time with different cluster centers initialization
(b) Use a different color space
[✓ (c)] Concatenate the coordinates (x, y) with the color features as input to the k-means algorithm
(d) Debug your code carefully as this issue should not happen

Q3 (True OR False) [1.5 marks]

Given an image containing a cat, automatically labelling the image as ”containing cat” (without specifying where the cat exists in the image) is considered an object detection task. F

Classification not Object Detection

Q4 (True OR False) [1.5 marks]

A Gaussian pyramid is a hierarchy of images. The bottom layer is an input image. The next layer is obtained by blurring the image in the previous layer and downsampling it, and so on. They are used in many computer vision algorithms such as SIFT T

Q5 (True OR False) [1.5 marks]

One can make a visual descriptor rotationally invariant by assigning orientations to the key points and then rotating the patch to a canonical orientation. In SIFT this is done by constructing Histograms of Gradients in a neighborhood around the feature point and assigning the largest bin as the corresponding direction of the keypoint. Later, all detected features are rotated so that the corresponding orientations are vertically aligned. T

Q6 (MCQ) [3 marks]

You have designed an algorithm for classifying if an image contains a person or not. You obtain an accuracy of 98% on the training set and 33% on the validation set.

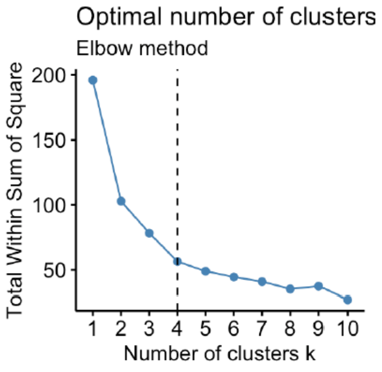
a) This is an instance of overfitting.
b) This is an instance of underfitting.
c) The training has failed.
d) The training succeeded but the training and testing examples are sampled from different distributions.
e) A & D

Q7 (MCQ) [3 marks]

Given the graph below, what is a reasonable number to use for clusters?

(a) 4
(b) 1
(c) 7
(d) 10
(e) 2

The goal is to find the "elbow point" — the point after which the WSS starts to decrease slowly.



Q8 (True OR False) [1.5 marks]

We need class labels when performing clustering F

Q9 (True OR False) [1.5 marks]

In regular shallow neural networks one needs to use a non-linear activation function between layers while in deep neural networks this is not needed F

Q10 (True OR False) [1.5 marks]

✗

In regression, the predicted output is discrete, this is in contrast to classification F

Q11 (True OR False) [1.5 marks]

In mean shift clustering algorithm we find cluster centers as the modes of the probability density function of input features T

Q12 (True OR False) [1.5 marks]

In general we need more manual work when building a labelled dataset for semantic image segmentation compared to the case of image classification T

Semantic segmentation requires pixel-level labeling, which is more labor-intensive than image classification.

Q13 (True OR False) [1.5 marks]

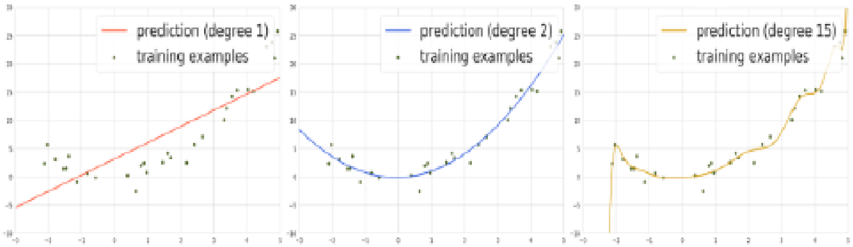
Object localization is different than object detection because in object localization we could have multiple instances of the target object F

Object localization typically identifies one instance; detection handles multiple instances.

Q14 (MCQ) [3 marks]

In the following figure

(a) All plots show a case of underfitting
(b) Left plot is a case of underfitting, middle plot is a case of good fitting and right plot is a case of overfitting
(c) Left plot is a case of overfitting, middle plot is a case of good fitting and right plot is a case of overfitting
(d) Left plot is a case of underfitting, middle plot is a case of underfitting and right plot is a case of overfitting
(e) None of the above



Q15 (True OR False) [1.5 marks]

Viola Jones detector is originally developed to detect frontal faces but the same algorithm can be also used for profile detection and other objects F

designed for frontal faces; adapting it for profiles requires significant changes.

Q16 (True OR False) [1.5 marks]

Some of the reasons we use convnets instead of fully connected layers in images are computational efficiency (sharing of weights) and preserving spatial structure T

ConvNets preserve spatial structure and reduce parameters via weight sharing.

Q17 (True OR False) [1.5 marks]

We can solve the object detection problem as a regression problem to find all boxes of objects in images T

Methods like YOLO frame detection as regression for bounding boxes.

Q18 (True OR False) [1.5 marks]

The main idea behind resnet blocks is to add shortcuts in the network blocks so that we can learn identity mapping in the worst case scenario. This helps learning even when we increase number of layers T

Q19 (True OR False) [1.5 marks]

In the case of L2 regularization we have two terms in the loss function: a data term and a regularization term. The regularization term is multiplied by a parameter, let’s call it “lambda”. If we increase lambda we are risking more overfitting. F

Increasing lambda strengthens regularization, reducing overfitting.

Q20 (True OR False) [1.5 marks]

Single shot detection (SSD) and You-only-look-once (YOLO) detection algorithms operate on regions extracted from a separate region proposal network (separate from the network that performs detection) F

SSD and YOLO do not use separate region proposal networks (unlike Faster R-CNN).

Q21 (True OR False) [1.5 marks]

A condition for a point to be a corner feature in an image is that there is a large intensity gradient along the X direction only F

Corners require large gradients in multiple directions (e.g., X and Y).

Q22 (True OR False) [1.5 marks]

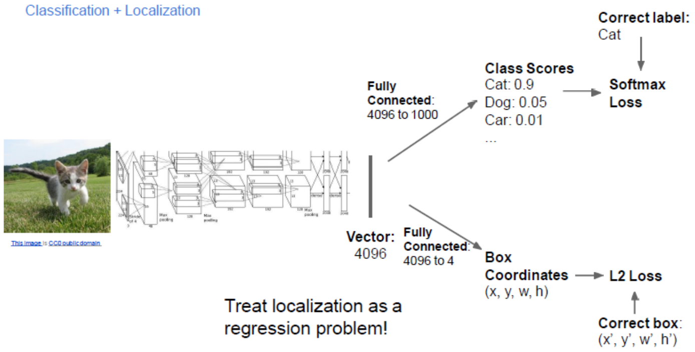
Using gradient-based feature descriptors does not guarantee invariance against linear illumination changes F

Q23 (True OR False) [1.5 marks]

The goal of the L2 loss in the following architecture is to predict the location of the object of interest T

L2 Loss is appropriate here because localization is treated as a regression problem, not a classification one.

Classification + Localization


Correct label:
Cat
Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...
Softmax Loss
Fully Connected: 4096 to 1000
Vector: 4096
Fully Connected: 4096 to 4
Box Coordinates (x, y, w, h)
L2 Loss
Correct box: (x', y', w', h')
This image is CC0 public domain

Treat localization as a regression problem!

Q24 (True OR False) [1.5 marks]

Being a template-based method, Viola Jones can be also very suitable for deformable objects F

Viola-Jones uses rigid templates and struggles with deformable objects.

Q25 (True OR False) [1.5 marks]

Bag of words feature matching does not take into account spatial feature arrangement when matching images T

Q26 (MCQ) [4 marks]

In this problem we are using an indexing approach to match a query image to a database of images.

- Suppose we already computed a set of visual words by K-means quantization, we have a total of 10,000 visual words (w_1 through w_{10000})
- Suppose the query image has the following visual words: W_{40}, W_{50}
- Suppose we have 1 million images in the database, the table below shows the index containing visual words (only a portion of the table is shown), each visual word has an associated "inverse document frequency" (IDF). For each database image each image has a "term frequency" (TF) for the visual word also (given after the DB image name)

For the given query image, the top 3 matching database images using TF*IDF as score are

a) 10, 11, 50
b) 12, 212, 142
c) 212, 142, 132
d) 122, 345, 765
e) None of the above

1. Identify IDF values for W40 or W50 (30,50)
2. List all images that contain either W40 or W50 and get TF*IDF for each
3. Sum the scores per image across both visual words:

Image	Total TF×IDF
Image_12	90 + 1500 = **1590**
Image_212	300
Image_132	60
Image_142	150

Visual Word index	IDF	Entry 1	Entry 2	Entry 4	Entry 5	Entry 6
40	30	Image_12, TF = 3	Image_212, TF = 10	Image_132, TF = 2		
50	50	Image_142, TF = 3	Image_12, TF = 30			
55	34	Image_122, TF = 6	Image_345, TF = 23	Image_98, TF = 9	Image_100, TF = 1	Image_1432, TF = 1
60	23	Image_341, TF = 5	Image_12, TF = 43	Image_42, TF = 98		
63	45	Image_12, TF = 3	Image_132, TF = 12	Image_42, TF = 32	Image_52, TF = 21	
2450	239	Image_345, TF = 3	Image_90, TF = 4	Image_89, TF = 53		
4539	321	Image_321, TF = 9				

Q27 (True OR False) [1.5 marks]

In designing cascaded object detection systems it is usually the case that the first detector in the cascade has high recall and possibly high false positive rate T

Cascade detectors prioritize high recall initially, tolerating higher false positives.

Q28 (MCQ) [3 marks]

Select all valid solutions for improving data annotation quality

a) Consensus / Multiple Annotation / “Wisdom of the Crowds”
b) Using a gold Standard
c) Using a second tier of workers who grade others
d) All of the above
e) A & B

Q29 (True OR False) [1.5 marks]

You are building an image classifier for a new set of classes for which there is no existing classifier. You want to start from an already trained classifier, you select the Resnet trained on the imagenet dataset. If you have a small number of labelled images in your dataset a good option would be to fine tune many layers in the already trained Resnet classifier. F

With limited data, fine-tuning many layers risks overfitting; only later layers should be adjusted.

Q30 (True OR False) [1.5 marks]
Random forest classifier is an example of an ensemble classifier methods T

Random Forest combines multiple decision trees (ensemble method).

Q31 (True OR False) [1.5 marks]

Running a detector through an image results in many possible overlapping boxes with different detection scores, a popular method to select the best detection window(s) is called Non-maxima suppression T Non-maximum suppression removes overlapping detections, keeping the highest-scoring ones.

Q32 (MCQ) [3 marks]

The main purpose of 1x1 convolution is to:

1x1 convolutions reduce channel depth efficiently.

a) Perform dimensionality reduction
b) This is not a valid operation, doing a convolution with a kernel size of 1 does not have any effect
c) I don’t think we studied this in computer vision
d) 1x1 convolution is an approximate way for performing 3x3 convolution
e) A & D