ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computer Vision
Winter 2022

Prof. Marc Pollefeys *Prof. Siyu Tang* *Prof. Fisher Yu*

Final Exam

24th of January, 2023

General Remarks

- The following materials are allowed for this exam:
  - exam questionnaire & blank paper (both provided by us)
  - ruler / square & pen
- Remove all material from your desk which is not allowed by examination regulations.
- Wearing headphones is not allowed.
- Please **do not use a pencil or red color pen** to write your answers.
- Place your **student ID** in front of you.
- You have **2 hours** for this exam.
- Fill in your first and last name and your ETH number and sign the exam.
- Put your name and ETH number on top of each sheet.
- Check that your exam questionnaire is complete.
- You may provide at most one valid answer per question. Invalid solutions must be clearly crossed out.
- Write down your answers directly on the exam sheets. You can use both sides.

First Name: _____

Last Name: _____

ETH number: _____

Signature: _____

	Topic	Maximum Points	Points Achieved	Points Check
1	Deep Learning	25		
2	Object Recognition	25		
3	Image Segmentation	25		
4	Multi-View Geometry	25		
Total		100		

Grade: .....

Question 1: Deep Learning (25 pts.)

a) *CNN Arithmetics*. The following notation is used: input channels C_{in}, output channels C_{out} , kernel size K, stride S, and padding P. Below is an architecture with standard convolution and max-pooling layers.

What are the output shapes (Channels, Width, Height) after each layer if we feed in a 224×224 RGB image? We have provided the output of an intermediate layer – please fill in your answers below for the other layers. **5 pt.**

Layer	Output Shape
Conv($C_{in} = 3, C_{out} = 64, K = 7, S = 2, P = 3$)	
ReLU()	
MaxPool($K = 2, S = 2, P = 0$)	(64, 56, 56)
Conv($C_{in} = 64, C_{out} = 192, K = 3, S = 1, P = 1$)	
ReLU()	
MaxPool($K = 2, S = 2, P = 0$)	

b) *Backpropagation*. Consider a simple 1-layer neural network with input $\mathbf{x}$, output $\mathbf{y}$ and without bias:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad \mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

The target $\mathbf{t}$ and the loss function $\mathcal{L}$ are given as follows:

$$\mathbf{t} = \begin{pmatrix} 15 \\ 6 \end{pmatrix} \qquad \mathcal{L} = \frac{1}{2}\sum_{j=1}^{2}(y_j - t_j)^2$$

1) Perform a forward pass and calculate the loss. **4 pt.**

2) Perform one step of gradient descent with a learning rate of 1. Make sure to include all steps of your derivation. **8 pt.**

c) *Multiple Choice*. Determine if the following statements are true or false by marking the correct answer with a cross (X). Each correct answer yields $+1$ point. Each incorrect answer yields -1 point. If you wish, you may answer only a subset of the questions (i.e., leave some answers empty). The minimum overall score that can be attained for this task is 0. **8 pt.**

Statement	**True**	**False**
Stochastic gradient descent computes an approximation to the true gradient.		
Choosing the learning rate of SGD too low can lead to oscillations.		
Decreasing training loss but increasing validation loss indicate overfitting.		
Decreasing training loss but increasing validation loss indicate underfitting.		
Residual connections make training of deep networks more difficult.		
A convolution layer with 3×3 kernel, 2 input feature channels and 2 output feature channels has 36 parameters.		
For the max unpooling layer, one needs to remember which element was maximum during the associated pooling.		
MLP is a fully connected class of feedforward neural network.		

Question 2: Object Recognition (25 pts.)

a) Suppose we train a VGG network for image classification and we have 4 classes (class 0, 1, 2, 3). Given a training image, the VGG network predicts the score for class i as x_i. Use x_i to calculate softmax for the probabilistic score S_i. Derive the cross entropy loss for this training image given that the true label of the current image is 2. **3 pts.**

b) In VGG16, each convolutional block consists of:

1. a 2D convolutional layer, what's the kernel size? Does the convolutional layer keep the spatial resolution?
2. an activation layer, what activation function is used?
3. a layer to downsample the spatial resolution. What operation is used here?

4 pts.

c) How does a dropout layer prevent overfitting during training? **2 pts.**

d) Judge the correctness of the following statements and **select** the corresponding box (☑). For each statement, 1 pt, 0 pt and **-1 pt** are given for a correct answer, both empty/selected boxes, and an incorrect answer, respectively. The minimum number of points is 0. **6 pts.**

		True	**False**
1)	Higher vocabulary size results in better classifier performance for Bag-of-Words (BoW) classification.	☐	☐
2)	For the same training data, repeating BoW image classification multiple times always yields the same codebook (or dictionary).	☐	☐
3)	Each classifier in AdaBoost is learned independently from each other.	☐	☐
4)	The codebook in BoW image classification is a collection of local features.	☐	☐
5)	K-means clustering is a supervised learning method.	☐	☐
6)	The sliding-window approach is robust to partial occlusion during test for samples unseen in training data.	☐	☐

e) Please explain the main steps of K-Means clustering. When used in BoW classification for code-book construction, what is the relation between vocabulary size and number of clusters K? **5 pts.**

f) Assume we have N data samples in the dataset, and each sample is a D-dimensional vector. If we operate K-means clustering with K clusters for P iterations, what is the time complexity of the algorithm? **2 pts.**

g) Given a BoW codebook and an image, how to encode the image into its associated BoW vector? **3 pts.**

Question 3: Image Segmentation (25 pts.)

We have discussed two unsupervised image segmentation algorithms in this course: (1) the K-Means algorithm and (2) the Expectation-Maximization (EM) algorithm.

a) The EM algorithm alternates between the so-called "E-step" and "M-step". Given a set of data points $\{x_i\}_{i=1}^N$, initial guess of Gaussian parameters (means and variances) $\theta = \{\mu_b, V_b\}_{b=1}^K$, and corresponding probability for selecting each of the blobs as $\{\alpha_b\}_{b=1}^K$. x's and μ's are **column vectors**, V's are **matrices**.

1. E-step: given current parameters θ and $\{\alpha_b\}_{b=1}^K$, describe the equation to compute $P(b|x, \mu_b, V_b)$, i.e. the probability that point x is in blob b. **2 pts.**
2. M-step: describe the formulas to obtain the updated blob probability, α_b^{new}, and updated parameters μ_b^{new}, V_b^{new}, for blob b. **6 pts.**

b) Judge whether the following statements about the K-Means algorithm and the EM algorithm are true or not. **2 pts.**

		True	False
1)	K-Means can only find a local minimum solution, while EM can find the global minimum solution.	☐	☐
2)	Given the same number of clusters/blobs and the same dataset, a K-Means model takes less storage than an EM model.	☐	☐

We have learned the Mean-Shift algorithm for image segmentation in the lecture.

c) Mean-Shift is a model-free approach compared to K-Means and EM. What kind of priors do K-Means and EM assume, respectively? **2 pts.**

We have learned Graph Cuts which solves Markov Random Fields (MRF) problems.

d) Optimizing MRF corresponds to maximizing the following joint probability:

$$P(x, y) = \prod_i \Phi(x_i, y_i) \prod_{i,j} \Psi(x_i, x_j)$$

where y's correspond to observations (e.g., observed pixel colors) while x's correspond to hidden states (e.g., whether the pixel belongs to foreground or background). With the above formulation we need to **maximize** the **product** of individual probabilities which can be difficult to solve in practice. What is the standard way to convert the above formulation so that we have a **minimization** problem of the **sum** of potentials? **2 pts.**

e) In *Boykov & Jolly, ICCV'01*, the energy minimization objective takes the following form:

$$E(L) = \sum_p D_p(L_p) + \sum_{pq \in \mathcal{N}} w_{pq} \cdot \delta(L_p \neq L_q)$$

where L is the set of labels assigned to pixels while L_p is the label assigned to pixel p. $D_p(L_p)$ is the so-called "regional term" while $w_{pq} \cdot \delta(L_p \neq L_q)$ is the "boundary term". $\mathcal{N}$ denotes the set of all neighboring pixels. $\delta(\cdot)$ evaluates to 1 if the input condition is true and 0 otherwise. We assume binary segmentation, i.e. $L_p \in \{0, 1\}, \forall p$.

1. How is w_{pq} defined as a function of pixel values I_p, I_q? Use a hyperparameter σ to control the "sharpness" of w_{pq}. (Hint: σ should correspond to standard deviation. w_{pq} should be large if two pixels p and q look similar, and small if they look dissimilar) **2 pts.**
2. Similarly, define regional terms for the foreground class s and the background class t, given current pixel value I_p, expected foreground pixel value I_s, and expected background pixel value I_t. **2 pts.**
3. Judge whether the following statement about Graph Cuts is true or not. **1 pts.**

	True	**False**
The s-t Graph Cuts (min-cut/max-flow) guarantees to find the global mininum for binary image segmentation problems.	☐	☐

In the lecture, we have learned dilated convolution for deep-learning-based image segmentation

f) Consider a 3x3 convolution layer with stride 1:

1. What are the receptive fields for dilation 1, 2, and 3? **3 pts.**
2. More generally, what is the receptive field for dilation$=i$? **3 pts.**

Corner case: if the student assumes this is the first layer and argues that receptive field is always 3×3, then it is also fine.

Question 4: Multi-View Geometry (25 pts.)

a) With a set of pixel-point correspondences ($\{x_i, X_i\}_{i=1}^N$), we can estimate the projection matrix $\mathbf{P}$ of a general projective camera with Direct Linear Transform (DLT). **6 pts.**

1. What's the minimum number of N needed? Why? 3 pt
2. We usually represent the intrinsic matrix (the values are all in pixels, e.g. $p_x = 300$pixel) of a general projective camera as below. What is the meaning of α_x, α_y, s, p_x, p_y? 3pt

$$\mathbf{K} = \begin{bmatrix} \alpha_x & s & p_x \\ 0 & \alpha_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

b) We assume that $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$. For a pixel $\mathbf{m}$ with homogeneous coordinate $[x, y, 1]$, we assume that the corresponding 3D point, denoted as $\mathbf{M}$, of $\mathbf{m}$ has a Z coordinate (along principle axis) as D in the camera's coordinate frame. Compute the 3D coordinate of $\mathbf{M}$ in the world's coordinate frame. (**Hint**: $\mathbf{R}^{-1} = \mathbf{R}^T$) **4 pts.**

c) With a set of pixel correspondences ($\{x_i, x_i'\}_{i=1}^N$) between two images, we can estimate the fundamental matrix from the first image to the second $\mathbf{F}$ with the eight point algorithm. **15 pts.**

1. What are the degrees of freedom (dof) of fundamental matrix $\mathbf{F}$ and essential matrix $\mathbf{E}$ respectively? 2 pt
2. For correspondence x_i, x_i', write down the formula of corresponding epipolar lines l_i', l_i with known $x_i, x_i', \mathbf{F}$. 2 pt
3. Assume that we already have the constraint equation $\mathbf{A} \cdot \mathbf{f} = 0$, where $\mathbf{f}$ is the flattened vector of $\mathbf{F}$. How to compute $\mathbf{f}$? (**Hint**: use SVD. Description is enough.) 2 pt
4. Following previous question, it may happen that the solution of $\mathbf{F}$ has rank as 3. What should we do to modify it? (**Hint**: use SVD. Description is enough.) 2 pt
5. Normalization of pixels is usually used as pre-processing (i.e., $\hat{x}_i = Tx_i$). What is the benefit? 2 pt
6. Since there may exist outliers in the N correspondence, we use RANSAC and eight point algorithm to estimate $\mathbf{F}$ for robustness. The algorithm is stopped when $\Gamma > 0.95$ (Γ is the probability that at least one sample does not have any outliers). Fill in the following pseudo code and then write down the formula for Γ based on the definition. The following notations are used: number of samples so far S, number of potential correspondences N, number of inliers I. 5 pt

1. compute N potential correspondences;
2. do
2.1: _____

2.2: _____
2.3: determine inliers and update I
until $\Gamma > 0.95$

3. _____

4. look for additional matches and refine $\mathbf{F}$ based on all correct matches

$\Gamma =$_____