Version RC/5

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 1 May 2024 2 to 3.40

Module 4F12

COMPUTER VISION

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 A 512×512 grey scale image, $I(x, y)$, is first low-pass filtered (smoothed) by convolving with a 2-D Gaussian filter, $G_\sigma(x, y)$, before gradients are computed as part of the feature detection process.

(a) (i) Give an expression for computing the intensity of a smoothed pixel, $S_\sigma(x, y)$, using two discrete 1-D convolutions. [10%]

(ii) How is low-pass filtering at multiple scales implemented efficiently using an *image pyramid* to sample *scale-space*? Show how to construct an image pyramid of low-pass filtered images, $S_{\sigma_0} \ldots S_{8\sigma_0}$, with $s = 3$ distinct images in each octave. [20%]

(iii) How many distinct Gaussian filters are needed to build the pyramid and identify their scale parameters, σ_k? [10%]

(iv) How are the spatial first and second derivatives computed? Identify the coefficients of the filter kernels used in practice. [10%]

(b) The low-pass filtered images and their spatial derivatives can be used for image *edge*, *corner* and *blob* detection at different scales.

(i) What is an image edge? How are edges used in computer vision? How is an appropriate scale, σ, selected? [15%]

(ii) What is meant by a corner feature and how are corners detected and localised in the image? [20%]

(iii) What is an image blob and how are they used in computer vision? Show how to recover their size and position. [15%]

2 A camera is viewing a known 3-D object under perspective projection. The image co-ordinates, (u_i, v_i), are the projection onto an image plane of the 3-D world point, (X_i, Y_i, Z_i).

(a) (i) What properties of the 3-D object are desirable for calibration? [10%]

(ii) The relationship between the 3-D object coordinates and their perspective projection can be expressed with homogeneous coordinates by a 3×4 projection matrix. Derive this matrix and state any assumptions made. [20%]

(iii) How are the elements of the projection matrix estimated in practice? You should describe the approach and include the appropriate equations. [20%]

(iv) For a calibrated camera each image position, (u_i, v_i), defines a ray in 3-D space. Show how to recover the equation of the ray. [10%]

(b) We now restrict the 3-D object to be planar.

(i) Show that the relationship between image co-ordinates and plane co-ordinates can be described algebraically with homogeneous coordinates by the following 2-D projective transformation:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ 1 \end{bmatrix}$$

[10%]

(ii) How many degrees of freedom does the 2-D projective transformation have? Describe, using sketches, how a square might appear under perspective projection. Be sure to account for each degree of freedom of the 2-D projective transformation. [20%]

(iii) Consider the conic $aX^2 + bXY + cY^2 + dX + eY + f = 0$. Derive the equation of its perspective projection. What will be the image of a circle? [10%]

3 Company Y operates a social media platform where users share images. Its engineering team has designed a computer vision neural network to assist them with the task of identifying harmful content (violence, nudity etc.) that may be inappropriate for the platform. The architecture of the system is outlined in Fig. 1.

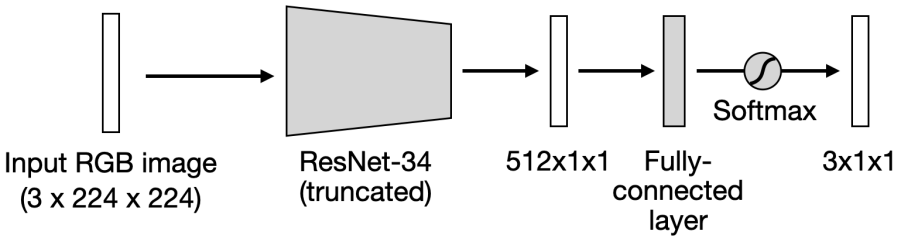

Fig. 1

The network takes in one 224×224 pixel RGB image at a time and classifies the content as belonging to one of three classes: *appropriate for everyone*, *appropriate for adults only* and *inappropriate for all users*. These classes are assigned the integer labels 0, 1 and 2, respectively.

(a) The first component of the architecture consists of a ResNet-34 network that has been pre-trained for the task of image classification on ImageNet, then truncated by removing its last fully-connected layer. What is the motivation for using a *pre-trained* neural network as part of the content moderation system? What is the motivation for *truncating* the ResNet-34? [20%]

(b) The model is trained with a cross entropy loss of the form $L(D) = \frac{1}{|D|} \sum_{d \in D} H(\mathbf{t}_d, P(c|\mathbf{x}_d))$ where D is the dataset of images, $\mathbf{x}_d$ is an individual RGB image, and $\mathbf{t}_d \in \{0, 1\}^3$ is its corresponding target label (encoded as a 3-dimensional *one-hot* vector). $P(c|\mathbf{x}_d)$ denotes the distribution over classes predicted by the model.

(i) Let s_i denote the ith output of the softmax function. Using the quotient rule or otherwise, give an expression for the derivative of the ith output of the softmax, s_i, with respect to its jth input, y_j. [10%]

(ii) We can express the cross entropy loss incurred for a single image-label pair, $(\mathbf{x}, \mathbf{t})$, as $L = H(\mathbf{t}, P(c|\mathbf{x})) = -\sum_i t_i \log s_i$, where again s_i denotes the ith output of the softmax layer. Using your answer to (b)(i), or otherwise, derive an expression for the derivative of the loss with respect to the jth input to the softmax layer, $\frac{\partial L}{\partial y_j}$. [20%]

(cont.

(c) Let $\mathbf{W}$ denote a vector containing all parameters in the truncated ResNet-34 and in the fully-connected layer. Suppose that gradients can be computed for the loss with respect to all parameters, $\frac{\partial L}{\partial \mathbf{W}}$ for a given minibatch of examples. Suppose also that company Y has gathered a collection of 100,000 labelled images. Describe in detail a recipe for training the network to perform the task and estimating its performance. Your answer should cover your choice of optimisation algorithm, hyperparameter selection, regularisation and any other strategies that could help the model achieve strong performance. [20%]

(d) Explain why a CNN is an appropriate choice for the feature extractor in this task? What inductive biases contribute to CNN statistical efficiency? What components of the ResNet architecture make it well suited to large-scale learning problems? [15%]

(e) Company Y is considering replacing the truncated ResNet-34 feature extractor with a Vision Transformer. How do the inductive biases of such a model differ from those of a CNN? What practical changes may be necessary to ensure that such a change of architecture is beneficial? [15%]

4 An object is viewed from two viewpoints with a mobile phone camera. The correspondences in the left and right images, (u, v) and (u', v'), satisfy the *epipolar constraint*:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

(a) (i) What is meant by the epipolar constraint? By considering the relative position and orientation of the camera in the two positions, derive the epipolar constraint and show how to compute the positions of the epipoles in each image. [25%]

(ii) Describe the difference in using the epipolar constraint in stereo vision compared to its use in structure from motion? [15%]

(b) The *fundamental matrix* is to be estimated from point correspondences.

(i) First a large number of *keypoints* are detected in each image and potential matches between the two images are found by comparing their descriptors. Describe two possible keypoint descriptors and give an algorithm for determining matches between the views. [20%]

(ii) How are consistent matches obtained in the presence of incorrect or outlier measurements? How is the fundamental matrix estimated when a large number of consistent matches is available? What additional constraint needs to be enforced? [20%]

(iii) Show how to recover the 3-D positions of the keypoints. What additional information is required? [20%]

END OF PAPER