First name	Last name	Degree

Master 2 – IP Paris & Univ. Paris Saclay
Image Mining Exam – October 2021

1 – Convolution Kernels

Look at the following 2d convolution kernels, try to interpret them mathematically (i.e. what measure are they supposed to estimate when they are applied on an image), and say what is their expected effect on the image. The origin of the kernel appears in bold.

$\frac{1}{4}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{0} \end{pmatrix}$	$\begin{pmatrix} -1 & 0 & 1 \\ -4 & 0 & 4 \\ -6 & \mathbf{0} & 6 \\ -4 & 0 & 4 \\ -1 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ -2 & \mathbf{-2} & -2 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 2 & 2 & 2 & 2 & 2 \\ 2 & -1 & -1 & -1 & 2 \\ 2 & -1 & \mathbf{-24} & -1 & 2 \\ 2 & -1 & -1 & -1 & 2 \\ 2 & 2 & 2 & 2 & 2 \end{pmatrix}$
(a)	(b)	(c)	(d)

Reply:

(a) The sum of the kernel is 1, with only positive values, it is a smoothing kernel that computes the average values of the 4 pixels located at (4,4) steps at the top-left of the current pixel. The resulting image must then look both blurred and shifted to the bottom-right.

(b) The sum of the kernel is 0, it is vertically symmetric and horizontally antisymmetric. It is a composition of a (first order) horizontal derivating kernel, and a vertical smoothing kernel which is a 5x1 approximation of the Gaussian kernel based on the binomial coefficients (Pascal triangle). So the convolution by this kernel estimates the horizontal component of the gradient vector. The resulting image must display large (positive or negative) values along the straight vertical contours.

(c) The kernel is symmetric, with zero sum and negative values along the horizontal axis. It is a composition of a (second order) vertical derivating kernel and a horizontal smoothing kernel which is a 1x3 box filter. So the convolution by this kernel estimates the second derivative with respect to the *y* axis. The resulting image must produce sign changes around the horizontal contours.

(d) The kernel is symmetric, with zero sum, negative values in the centre, and positive values in the periphery. It is an approximation of the Laplacian operator at the scale of 5x5 pixels. The resulting image must produce sign changes around the contours, whatever their orientation.

2 – Image sub-sampling, Visual Aliasing and Multi-scale analysis

Define the operation of image sub-sampling, and the phenomenon of aliasing in images. Explain the rules of sub-sampling to obtain a multi-resolution representation (pyramid) of an image. What is the interest of multi-scale analysis (provide two examples)?

Reply:

Sub-sampling is the process of reducing the number of samples (pixels) in an image. If the image is not smoothed before being sub-sampled, artifact structures can appear in the regions of highest frequency, due to spectrum overlaps.
This phenomenon known as aliasing must be avoided while sub-sampling an image to obtain an image pyramid. To do so, the image must be smoothed in order to remove the highest frequencies and to respect the Shannon-Nyquist criterion (sample frequency must be at least twice the highest frequency in the image).
The interest of multi-scale analysis is to be able to estimate any local feature (contrast, orientation, curvature,...) relatively to different scales, which is essential as long as the interest objects in the image may have different sizes.
Multi-scale analysis appear in many cases in computer vision, e.g. to compute scale-invariant image representations (corner points, contours,...) or by the convolutional networks, where the intrinsic scale of a feature map, related with the size of the receptive field, increase with the depth of the layer.

3 – Interest (salient) points in images

Interest (or salient) points in images are points that are expected to be easier to track from one image to the other. How can you characterise them in terms of appearance? Cite 2 examples of algorithm to detect them. What are the expected properties of a good detector?

Reply:

Salient points must be easily distinguishable from their neighbourhood. Their local geometry must then present some singularity: corner points, white or dark spots, saddle points, junctions,...
Many algorithms exist to detect them, for example:

- The determinant of the Hessian matrix (SURF, KAZE,...) can be used, as its local extrema correspond to points with two high (absolute) main curvatures K_1 and K_2. The case of $0 << K_1 < K_2$ correspond to white spots, $K_1 < K_2 << 0$ to dark spots, and $K_1 << 0 << K_2$ to saddle points.
- The FAST detector selects corner points by finding pixels whose neighbourhood, represented by the discrete circle of radius 3 centred on the pixel, is such that there exists a long run of contiguous pixels whose values are all significantly higher or all significantly lower than the centre's value.

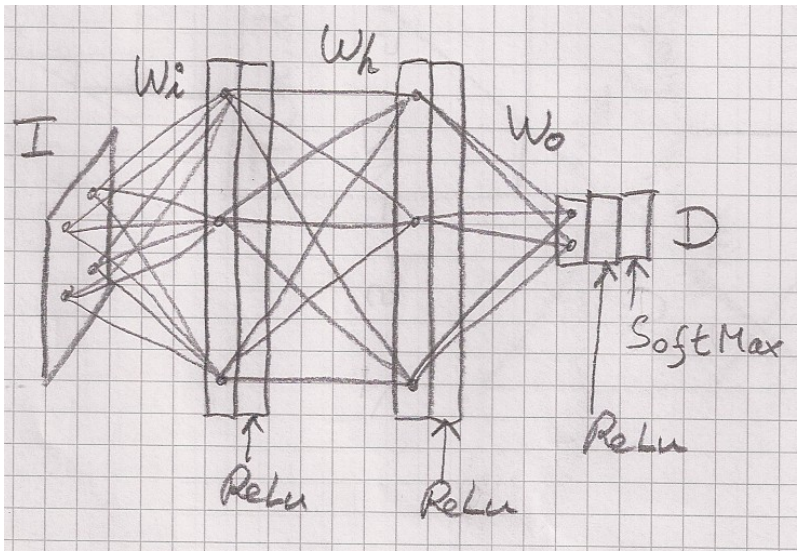
A good detector is expected to be sensitive (i.e. detect many points), repeatable (i.e. detect the same points in different images whatever their deformation), and efficient (i.e. fast to calculate).

4 – Neural networks

In the purpose to detect defaults in industrial textiles, a neural network has been trained to classify small (5×5) image patches into two classes (default / no default). The neural network is a multi-layer perceptron (MLP) with 2 hidden layers, each one composed of 10 neurons. The input is a 5×5 graylevel image patch, and the output is a vector of dimension 2. All activation functions are assumed to be ReLU. (1) Draw the network. (2) Write the function that relates the input to the output of the network. (3) Explain how this binary classifier can be used to detect defaults in large textile images.

Reply:

1-



2- Let $I \in \mathbb{R}^{25}$ the input patch, $D \in \mathbb{R}^{2}$ the binary output. W_i, W_h and W_o are the weight matrices, with sizes 10x25, 10x10 and 2x10 respectively. $b_i \in \mathbb{R}^{10}$, $b_h \in \mathbb{R}^{10}$, and $b_o \in \mathbb{R}^{2}$ are the bias vectors.
The output is given by: $D = \mathrm{SoftMax}(\mathrm{ReLu}(W_o x_2 + b_o))$,
with $x_2 = \mathrm{ReLu}(W_h x_1 + b_h)$, and $x_1 = \mathrm{ReLu}(W_i I + b_i)$.

3- On inference, one patch I is classified by $l(I) = \mathrm{Arg\ Max}\ D(I)$. For a large textile image, there are many different solutions: the image can be sliced into 5x5 tiles, each one being classified as a whole, or each pixel can be classified once according to the class of its 5x5 neighbourhood, or for more robustness, each pixel can be multiply classified according to every patch it belongs to, the label being set on a majority rule:

$$l(x) = \mathrm{Maj}\ \{\mathrm{Arg\ Max}\ D(I);\ x \in \mathrm{Supp}(I)\}$$

5 – Aesthetic appraisal

In order to measure the role of smile in the assessment of the beauty of a portrait photo, several resources are available:

- a powerful multilayer deep neural network (DNN) (for instance INCEPTION or VGG)
- a general purpose image data base with 10 Million images (for instance ImageNet), indexed with 1000 labels of every day objects
- a data base with 100 000 images dedicated to aesthetic appraisal of images, indexed with a level of beauty in the range [0,10] (for instance AVA)
- a set of 10 000 portrait photos of people, some smiling, some not, without any additional information on this set.

You are asked to provide information about the role of a smile in the user's appraisal of beauty in a portrait. Suggest a protocole to derive this information. Comment the most difficult steps.

Reply:

There is no exact solution in this question since several strategies may drive to a correct solution (until we experiment it!).
The evaluation takes into account the clarity of the explanation and the soundness of the arguments.

One of the "good solutions" could be the following :
1) fine-tune a DNN trained with ImageNet to assess beauty by training the ultimate layers with AVA data-base = Network 1 (for instance with a 10 class output layer)
2) fine tune another DNN trained with Image Net to discriminate a "level of smile" with 80 % of the images of the portrait database = Network 2 (for instance with a 4 class output : No smile, Faint Smile, Evident smile, Hilarious smile
3) process in parallel the 20 % of the unprocessed portraits images with N1 and N2
4) look for a possible statistical dependency between the 2 notes given by N1 and N2 for each image.

Another solution could be
1) to fine-tune a general purpose network to detect smiles, (for instance as a subclass of "portrait" which is a subclass of "persons")
2) then to use it to filter AVA to select the only images with portraits
3) detect some subclasses to evaluate a "level of smile" either in the first or in the second network.
4) From the filtered AVA data base and the "level of smile" note, compute the relationship between "level of smile" and "beauty" may be deduced by regressing the given notes.

6 – Remote sensing and satellite imaging

1 – For the 2024 Olympic games, the French government intend to have a permanent survey of the Stade de France, with a static camera in the visible range on a geostationary satellite.
The required resolution on the ground is 1 meter, the best available technology provides sensors with a pixel size of 1 micrometer.
The optical system being diffraction limited, what should be the diameter of the camera lens on board the satellite?

☐ 1 meter ☐ 3 meters ☒ 15 meters ☐ 75 meters ☐ 180 meters

2 – Burj Khalifa tower in Dubaï is located on the sea side. Its height is approximately 800 meters. On a SPOT 6 satellite vertical image (with resolution 1.5 meters per pixel), it appears not in the center of the image, but 24 km on the East. After a projection of the image on the reference sea level plane, the tower appears slanted. How many pixels separate the foot and the top of the tower?

☐0.3 pixels ☐3 pixels ☐10 pixels ☒16 pixels ☐32 pixels ☐164 pixels ☐340 pixels

3 – Detecting agricultural crops from texture measurements, which satellites seem likely to provide useful information for the following cultures:

	Lettuce or Carrot	Maize or Corn	Vineyards	Dense orchards of cherry trees or apple trees	Sparse orchards of olive trees
Very High Resolution (<1 meter)		X	X	X	X
High Resolution (1 to 5 meters)			X	X	X
Middle Resolution (5 to 30 meters)				X	X
Low Resolution (> 30 meters)					

Technical data to be used if needed:

- Orbiting polar orbit, 15 revolution/day, altitude = 800 km, speed wrt Earth = 3000 km/h, optical range: visible = 1 micrometer
- Geostationary : equatorial orbit, speed wrt Earth = 0, altitude = 14 000 km, optical range: visible = 1 micrometer

If you have any remark to add to your answer:

Rows of lettuces or carrots are approximately 50 cm away. They are likely to be undetected with any satellite. Corn rows are at a distance of 1 meter, vineyards at 2 meters, dense orchards have trees at a distance of 5 to 10 m, while olive orchards have trees about every 20 m.
Low resolution satellites will not detect any of theses textures. VHR will detect all of them but the smallest. A satellite will definitely detect any texture with a spatial period greater than the pixel resolution, whatever the orientation of the texture wrt the scanning direction.