Problem Set 3

Please submit a typed PDF addressing all problems below. This problem set contains 3 questions and is worth 25 points. **All responses MUST be in *your* own words.** Justification must be provided for **all** written answers. Statements made without any supporting explanation/justification will receive **no credit**. For mathematical derivations and plots, you may insert pictures of handwritten work if you find this easier. The required weekly readings and lecture slides should be helpful in completing the assignment. You can find these on our course website.

1. **Computer Vision Problems [6 points]:**

   (a) Define image classification, object detection, instance segmentation, and semantic segmentation. Then, describe the similarities and differences between these tasks. Your response should include discussion of the inputs and outputs of neural networks used for these tasks.

   (b) Provide one example real world use-case for each of the four tasks discussed in part (a).

2. **Model Size [9 points]:** You are designing a neural network to classify inputs of shape $3 \times 224 \times 224$ into 100 classes, and need to investigate the differences between a fully connected and convolutional approach.

   (a) Compute the number of model parameters needed for a fully connected architecture. This model first "flattens" the input, then feeds forward through a single hidden layer with half as many neurons. In other words, the hidden layer down-samples the flattened input by a factor of 2. Then, a second fully connected layer transforms the intermediate representation to the 100 dimensional output required for the task.

      i. Report the number of parameters per layer.

      ii. Compute the number of total model parameters (weights and biases).

   (b) Compute the number of model parameters needed for a convolutional neural network architecture. This model uses 2-D convolution layers with 7×7 filters, no padding, and 1×1 stride. The layers do not change the number of channels, i.e., the output from the convolutions should also have 3 channels, same as the input. This type of convolutional layer is repeatedly applied until the features are of shape $3 \times 8 \times 8$. Then, a flatten layer is used and the result is fed into a fully connected layer with an output size of 100.

i. Find the number of parameters for a single 7×7 convolutional layer which takes an input with 3 channels and returns an output with 3 channels.

ii. Using convolutional layers of the type described in (i), compute how many of these layers are required to reduce the input to the shape $3 \times 8 \times 8$.

iii. Compute the number of parameters required for a fully connected layer which takes an input of shape $3 \times 8 \times 8$, flattens it, and then produces an output of size 100.

iv. Compute the total number of parameters for the entire architecture.

3. **Convolutional Operations [10 points]:** This is **NOT** a programming exercise. Parts (a) and (b) both refer to the following data:

Data for Question 3

Input

2	0	-1	-1
0	1	0	2
-2	0	1	1
1	1	0	-1

Filter

1	-1	1
0	1	0
1	-1	1

(a) Compute and report the output after performing a convolution of the input with the filter. Use a 1×1 stride and no padding. For full-credit, you **MUST** show your mathematical derivations.

(b) Compute and report the output after performing a *transposed* convolution of the input with the filter. Use a 1×1 stride and no padding. For full-credit, you **MUST** show your mathematical derivations.

4. **Extra Credit [2.5 points]:** You are designing a fully convolutional architecture for an image segmentation task. The model consists of two stages: a series of 2-D convolution layers for down-sampling followed by a series of 2- D transposed convolution layers for up-sampling. The inputs are of shape $3 \times 224 \times 224$, the desired smallest *latent* representations are of shape $256 \times 16 \times 16$, and the final outputs are of shape $3 \times 224 \times 224$. The down-sampling side uses filters of size 3×3, and the number of channels is given by:

$$[3, 16, 32, 64, 128, 256, 256, 256...] \tag{1}$$

where the channels are kept at 256 until the desired latent representation shape is achieved by continual application of 2-D convolution layers. After the features reach shape 256 × 16 × 16, transposed convolution layers with 9 × 9 filters are used to up-sample the features back to their original spatial resolution. The number of channels for this stage is given by:

$$[256, 256, ..., 256, 128, 64, 32, 16, 3] \tag{2}$$

In other words, the channels are kept at 256 until the final 5 layers, which follow the same pattern as the down-sampling stage in reserve. For all layers, a 1 × 1 stride is used with no padding. Compute the number of parameters required for: (1) the down-sampling stage, (2) the up-sampling stage, and (3) the entire model.

Collaboration versus Academic Misconduct: Collaboration with other students (or AI) is permitted, but the work you submit must be your own. Copying/plagiarizing work from another student (or AI) is not permitted and is considered academic misconduct. For more information about University of Colorado Boulder's Honor Code and academic misconduct, please visit the course syllabus.