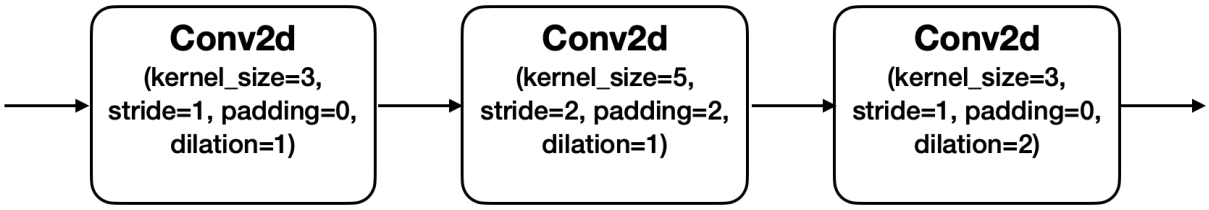
AI6126: Assignment 2

Deadline: 31 March 2023 11:59PM

Question 1: Segmentation-related questions. (7 marks)

i) What is the motivation for using downsampling and upsampling in the Fully Convolutional Network for semantic segmentation?

ii) Given the following network, calculate the receptive field.



iii) What are the advantages of dilated convolution over standard convolution for the semantic segmentation task?

iv) Given a transposed convolution kernel as follow, whose stride=1, padding=0, dilation=1,

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$$

and input,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

what's the output matrix after the transposed convolution?
(Hint: the output should be a 4x4 matrix.)

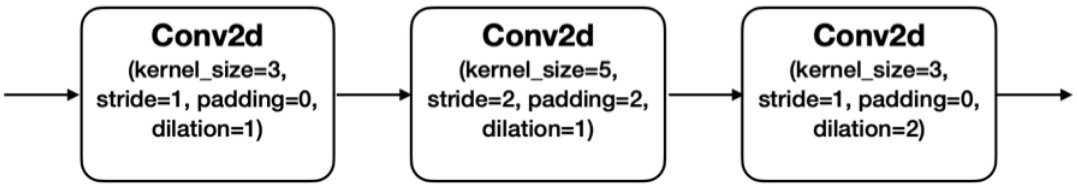
Question 2: What is the key difference between VQVAE and VAE? (2 marks)

Question 3: What is training objective of Generative Adversarial Networks? Please answer this question conceptually, i.e., do not just posing mathematical equations. (3 marks)

Question 1: Segmentation-related questions. (7 marks)

i) What is the motivation for using downsampling and upsampling in the Fully Convolutional Network for semantic segmentation?

ii) Given the following network, calculate the receptive field.



iii) What are the advantages of dilated convolution over standard convolution for the semantic segmentation task?

iv) Given a transposed convolution kernel as follow, whose stride=1, padding=0, dilation=1,

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$$

and input,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

what's the output matrix after the transposed convolution?
(Hint: the output should be a 4x4 matrix.)

i) In the Fully Convolutional Network, the main issue was the memory space and computational cost due to how it works. With n number of 3x3 convolutional layers, the receptive field size is $1+2n$. For an image size of 400 x 400 pixels, in a network with 200 3X3 convolutional layers, there needs to be 200 X 400 X 400 X 4 $\approx$ 122MB of memory space per block per image. With 100 blocks and batches of 20 images, it requires approximately 238GB of memory space. Using downsampling would reduce the number of parameters and memory space requirements, and upsampling is performed to restore the original resolution.

ii) $R_3 = 1 + \sum_{j=1}^{3} (F_j - 1) \prod_{i=0}^{j-1} S_i$

$= 1 + (3-1)(1) + (5-1)(1 \times 1) + (5-1)(1 \times 1 \times 2)$

$= 15$

$S_0 = 1$

$F_1 = 3 \quad S_1 = 1$

$F_2 = 5 \quad S_2 = 2$

$F_3 = 5 \quad S_3 = 1$

iii) Dilated convolution provides a larger field of view while maintaining the same feature map output size. By stacking several convolutions of different dilation values, it is possible to get finer segmentation feature maps as they contain information of different sizes of sub-regions of the image.

iv) $$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix} * \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & 3 & 4 & 0 \\ 3 & 4 & 5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 4 & 6 \\ 0 & 4 & 6 & 8 \\ 0 & 6 & 8 & 10 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 3 & 6 & 9 & 0 \\ 6 & 9 & 12 & 0 \\ 9 & 12 & 15 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & 8 & 12 \\ 0 & 8 & 12 & 16 \\ 0 & 12 & 16 & 20 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 4 & 7 & 6 \\ 5 & 17 & 27 & 20 \\ 9 & 27 & 37 & 26 \\ 9 & 24 & 31 & 20 \end{bmatrix}$$

Question 2: What is the key difference between VQVAE and VAE? (2 marks)

VQ VAE utilizes vector quantization, which is the use of a vector which is quantized from an input image in order to get the latent representation, whereas in VAE, the latent representation is generated from a continuous probability distribution, such as Gaussian.

Question 3: What is training objective of Generative Adversarial Networks? Please answer this question conceptually, i.e., do not just posing mathematical equations. (3 marks)

GANs consists of a generator network and a discriminator network. The generator attempts to make realistic images based on random noise, and the discriminator attempts to differentiate the realistic but fake images produced by the generator from the real sample images. As training progresses, the generator develops better images to 'fool' the discriminator while the discriminator improves its detection to determine whether the input image is real or fake. This training process becomes a minimax game where the generator attempts to minimize the discriminator's detection capability while the discriminator attempts to maximise its detection to differentiate the real sample images from the fake.