Assignment 2
Zheng Weixiang
G2103278G

Question 1

1) In Semantic Segmentation, every pixel in the image is labelled with a category label. It does not differentiate instances, only care about pixels. Background is also labelled.

   In Instance Segmentation, there is a difference of things which can be separated into object instances, and stuff which cannot be separated into instances. Instances of the same category are differentiated.

2) Rk = 1 + sigma(Fj -1) Pi(Si) (formula from the lecture)
   = 1 + 4*1 + 4*2 + 2*1
   = 1 + 4 + 8 + 2
   = 15
   Note: the third Conv2d has a kernel size of 3 but dilation of 2, to make thing convenient, we use 5-1 = 4 instead.
   The receptive field is 15

3)
- Dilated Convolution
- Markov Random Field
- Pyramid Scene Parsing Net
- Using multiple neighbourhoods around the pixel of interest and aggregating different hypotheses about the pixel's label [1].

4)Transposed convolution can be seen as regular convolution with padding
With matrix =
1 2 3
2 3 4
3 4 5
And input =
1 2
3 4
The result is the same as doing convolution with

1	2
3	4

On matrix

0	0	0	0	0
0	1	2	3	0
0	2	3	4	0
0	3	4	5	0
0	0	0	0	0

The result is therefore

1	4	7	6
5	17	27	20
9	27	37	26
9	24	31	20

Question 2

The reparameterization trick allows us to restructure the way we take the derivative of the loss function so that we can optimize our approximate distribution.

Question 3

GAN is a generative model that learns p(x).
Conditional GAN is a generative model that learns p(x|y).

Question 4

1) Component A: Positional Encoding.
   Component B: Multi-Head Attention.
   Component C: Feed Forward Neural Network.
   Component D: Masked Multi-Head Attention.
   Component E: Multi-Head Attention.
   Component F: Feed Forward Neural Network.

2) Positional Encoding is used to provide information of every token’s position. Because all tokens enter the network simultaneously, we need a way to keep track of the positions of these tokens.

3) The input of Multi-Head Attention is Query, Key, and Value.

4) Because during the process of attention, we don’t want the token to see the whole sequence of data, i.e., if we are to predict the third token, only the first token and the second token should be seen, if the third token is also seen, that’s cheating. Therefore, we need to force the model to use the left data to do the attention.
In practice, we could substitute the value that needs masking to a very small value, e.g., 1e-09, when doing the SoftMax, because the value is small, it gets ignored.

Reference:
[1] R. Mesbah, B. McCane, S. Mills and A. Robins, "Improving Spatial Context in CNNs for Semantic Medical Image Segmentation," 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), 2017, pp. 25-30, doi: 10.1109/ACPR.2017.15.