INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Stamp / Signature of the Invigilator

EXAMINATION (Mid Semester)	**SEMESTER (Spring)**

Roll Number									**Section**		**Name**	

Subject Number	C	S	6	0	0	1	0	**Subject Name**	***Deep Learning***

Department / Center of the Student		**Additional sheets**	NA

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession, even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer script and do not tear off any page. **Use last page(s) of the answer script for rough work.** Report to the invigilator if the answer script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the examination hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer script to the invigilator. **In any case, you are not allowed to take away the answer script with you.** After the completion of the examination, do not leave the seat until the invigilators collect all the answer scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as '**unfair means**'. Do not adopt unfair means and do not indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be filled in by the examiner

Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Marks Obtained											

Marks obtained (in words)	**Signature of the Examiner**	**Signature of the Scrutineer**

CS60010 Deep Learning, Spring 2024–2025

18-Feb-2025, 09:00–11:00 **Mid-Semester Test** Maximum marks: 50

Instructions

- Write your answers in the respective space provided in the question paper itself. Be brief and precise. Please provide the final answer with related calculations.
- Answer all questions.
- There are no clarifications. In case of confusion, you can make a valid assumption, state that properly and proceed.

Questions start from the next page.

1. Answer the following questions.

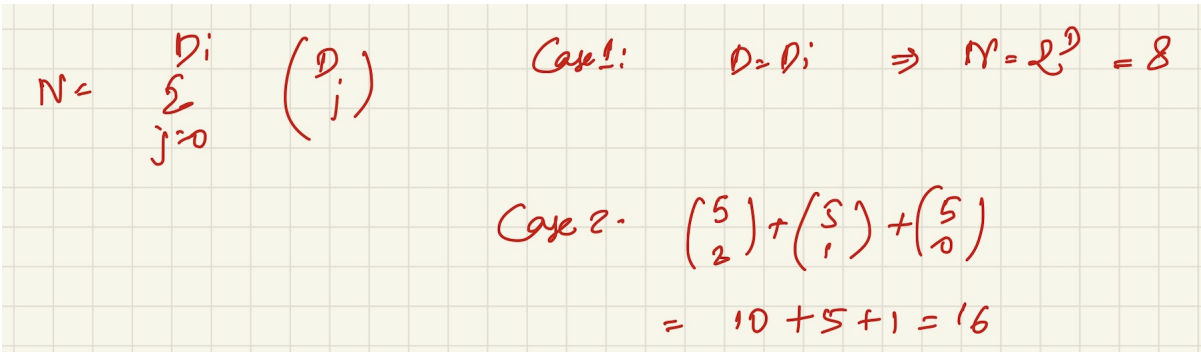
(a) Suppose you are using a shallow NN with two inputs and two outputs. Assume that there are 7 linear regions in both the outputs. What is the minimum number of neurons you would need in the hidden layer to properly approximate these outputs in the general case? **(2)**

Solution In the general case, the joins required by the 7 linear regions for the first output and 7 linear regions for the second output can be different. We also know that minimum of 3 neurons are required to get 7 linear regions with 2 inputs. Hence the answer is 3+3 = 6.

(b) What will be the maximum number of linear regions in the following cases (for a single output in shallow NN). Briefly justify your answer. **(3)**

Case 1. 3-dimensional input, 3 hidden units.

Case 2. 2-dimensional input, 5 hidden units.

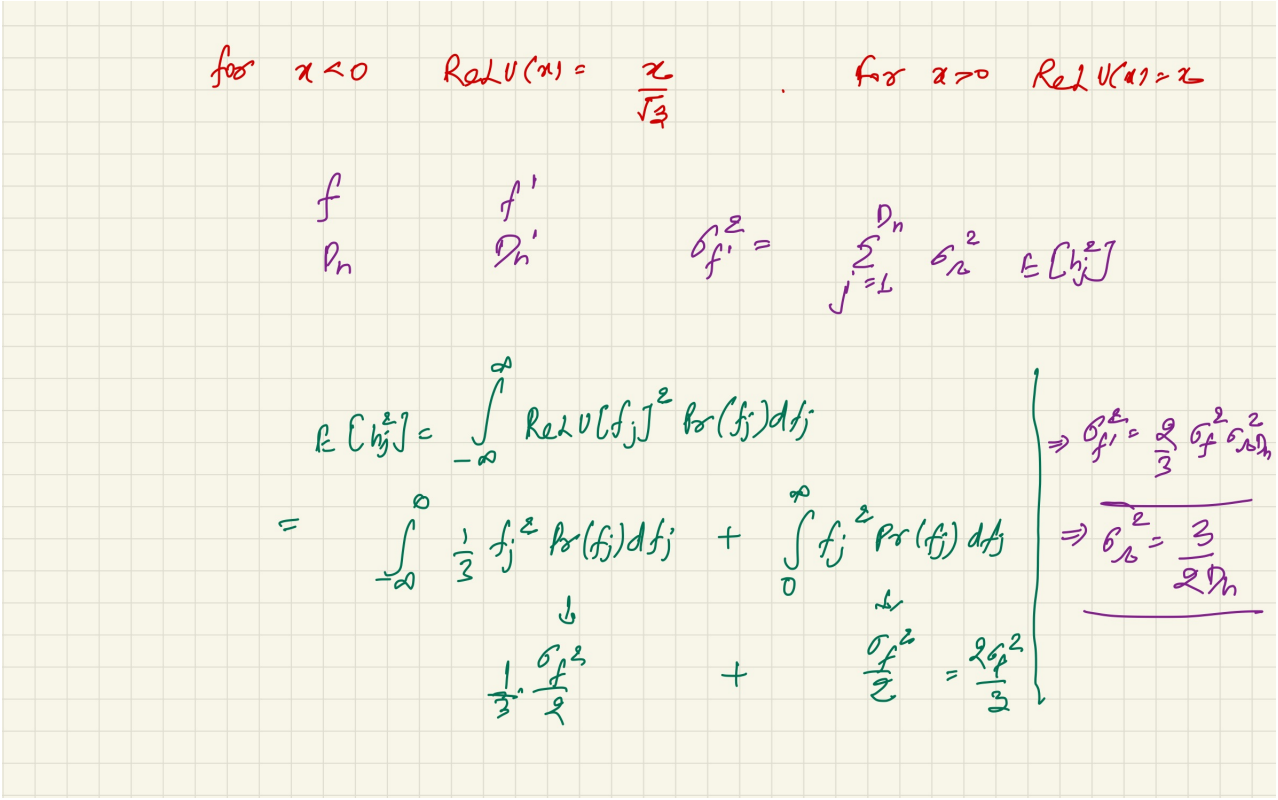


(c) In the YOLO architecture, assume that you are using a grid of size 7×7. Also, assume that for each grid, you predict three anchor boxes, and there are a total of 20 object classes. What will be the final output dimension? Suppose for an image in your training, only 5 of these grids contain objects. How many values will be "don't care"? **(3)**

Solution The final output dimension will be $7 \times 7 \times (20 + 3 * 5)$ for 20 objects and 5 predictions per anchor box (3 boxes).

Out of 49 grids, 44 grids do not contain objects. So, for each of these, $(20 + 3 * 4)$ values will be "don't care". So, $44 \times 32 = 1408$ values are don't care.

(d) Suppose we modify the activation function ReLU such that for $x < 0$, $ReLU(x) = \frac{x}{\sqrt{3}}$. What would be an appropriate initialization for variance of the weight matrix σ_Ω^2. The dimensions of the original layer to which the weight is applied is D_h. **(4)**



(e) Consider a deep network with 7 inputs, 3 outputs and 12 hidden layers, each layer containing 15 hidden units. What is the width of the network? What is the depth? What is the number of parameters (weights and biases)?

Suppose you are using dropout, that is, each hidden unit may be dropped out with a probability of 0.2. Suppose that you plan to use dropout at inference time as well. Can you see this as ensembling? How many models are you ensembling? **(4)**

Solution Width of the network = 15, depth of the network = 12 (1 mark)

Number of bias terms: 15*12+3 = 183

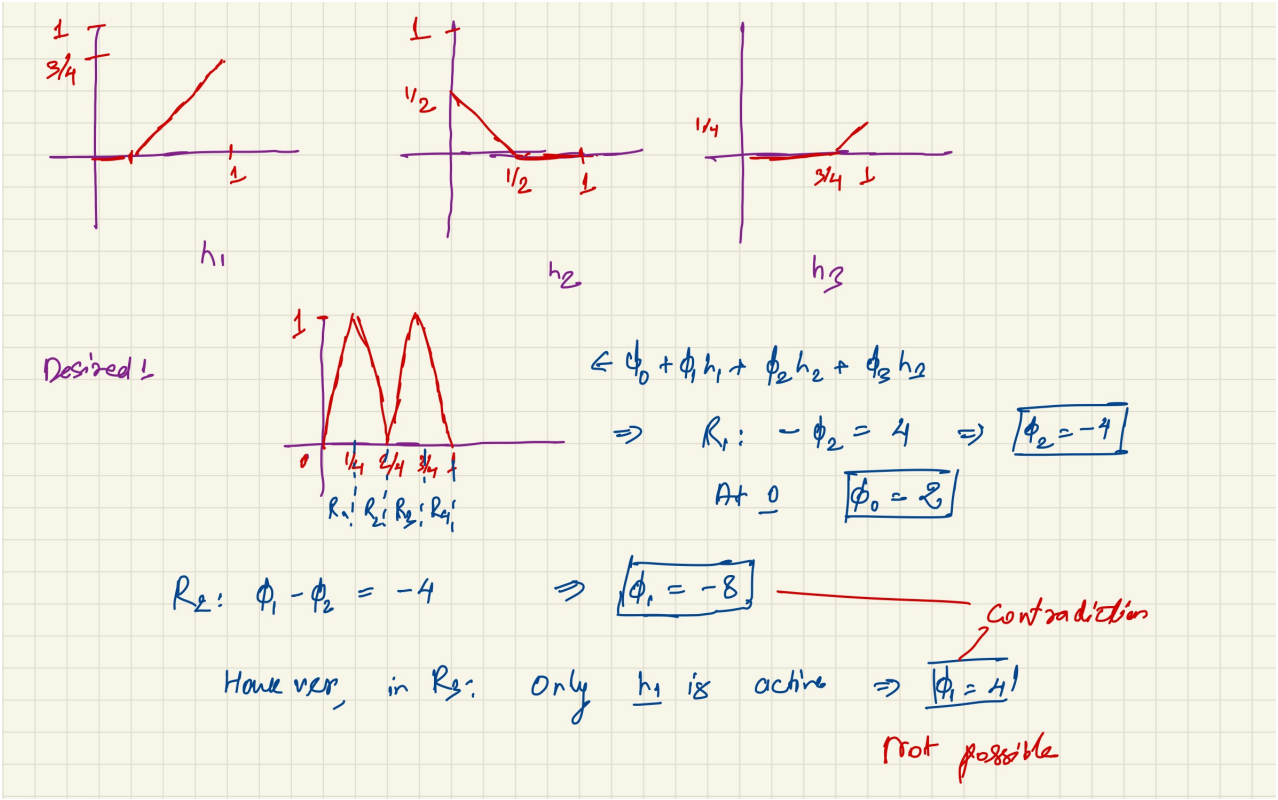
Number of weights: 7*15 + 11*15*15 + 15*3 = 2625

Total parameters = 2808 (1.5 marks)

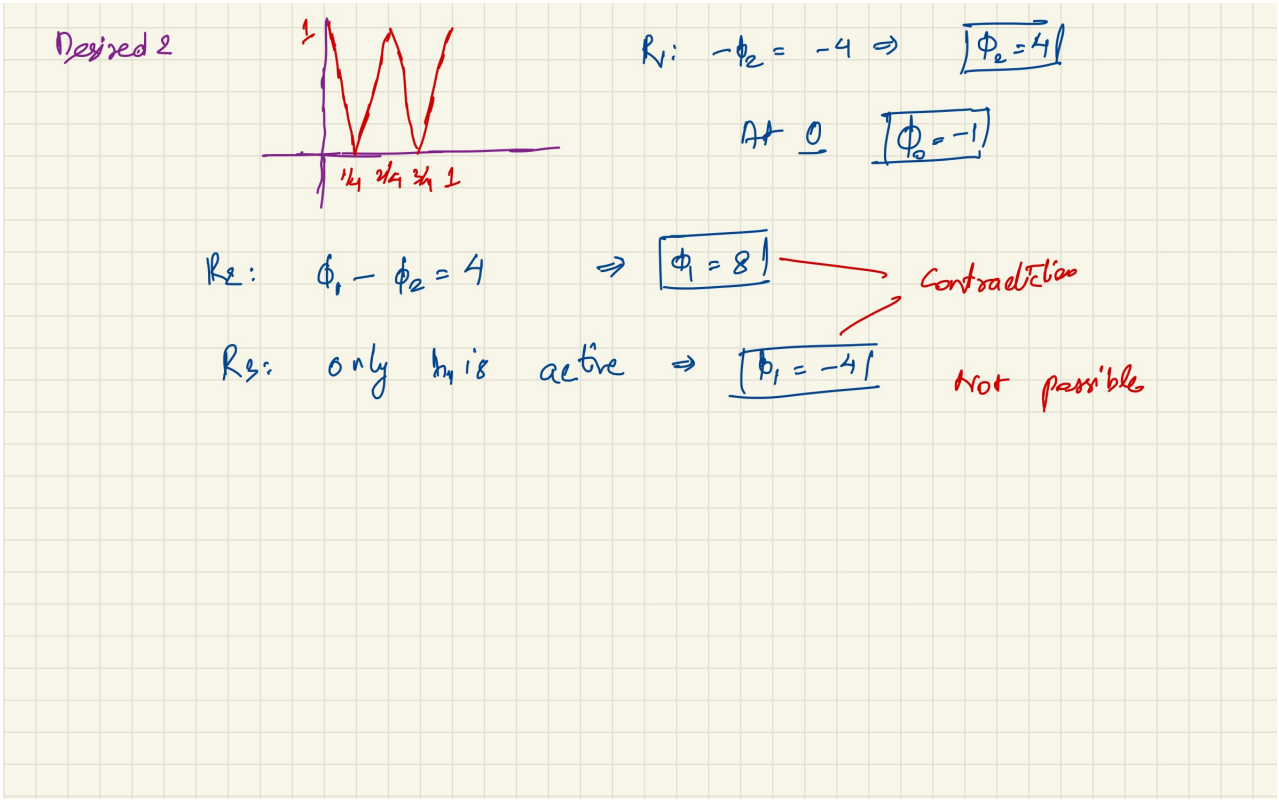
Each hidden unit (out of 12*15 = 180) may be dropped with a certain probability. So, this is an ensembling of 2^{180} models (1.5 mark)

2. Assume that a shallow network uses three hidden units with slopes as 1.0, -1.0, and 1.0, respectively, and the joints in the hidden units are at positions 1/4, 2/4, and 3/4, respectively. Find values of ϕ_0, ϕ_1, ϕ_2, and ϕ_3 that will combine the hidden unit activations as $\phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$ to create a function with four linear regions that oscillate between output values of zero and one. The slope of the leftmost region should be positive, the next one negative, and so on. If such a construction is not possible, prove that.

In the above example, what if you want the slope of the leftmost region to be negative, next one positive, and so on? **(6)**



Contd..



3. Consider a multivariate regression problem where we predict five outputs, so $y \in R^5$, and model each with an independent normal distribution where the means μ_d are predicted by the network, and variances σ^2 are constant. Define the loss function. Is the final form still similar to the mean squared error loss if we do not estimate the variance? **(3)**

$$Pr(y|f[x,\phi], \sigma^2) = \prod_{d=1}^{5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_d - f_d[x,\phi])^2}{2\sigma^2}\right]$$

$$L = -\sum_{i=1}^{I} \log\left[Pr(y|f[x,\phi], \sigma^2)\right]$$

$$= -\sum_{i=1}^{I}\sum_{d=1}^{5} \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left[\frac{-(y_{id} - f_d(x_i,\phi))^2}{2\sigma^2}\right]\right]$$

$$\Rightarrow \hat{\phi} = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^{I}\sum_{d=1}^{5} (y_{id} - f_d[x_i,\phi])^2$$

→ Not required during argmin

Final form is similar to MSE

4. We use the normalized gradients as follows:

$$\phi_{t+1} = \phi_t - \alpha \frac{m_{t+1}}{\sqrt{v_{t+1}} + \varepsilon}$$

i. What should be the choice of ε?

ii. How do AdaGrad and RMSProp modify the above equations? **(3)**

Solution ε is chosen as a very small positive constant.

Normalized gradients $\phi_{t+1} = \phi_t - \alpha \frac{m_{t+1}}{\sqrt{v_{t+1}} + \epsilon}$

Adagrad $v_{t+1} = v_t + \left(\frac{\partial L}{\partial \phi_t}\right)^2$

RMS Prop $v_{t+1} = \gamma v_t + (1-\gamma)\left(\frac{\partial L}{\partial \phi_t}\right)^2$

5. What can be done for the (non-stochastic) gradient descent to escape the local minima? Mention atmost 2 points very briefly. No need to explain. **(2)**

Solution One can perform gradient descent by starting from various different initializations.

6. Suppose, we run the stochastic gradient descent algorithm for 5,000 iterations (steps) on a dataset of size 10,000 with a batch size of 50. How many epochs are we running the algorithm for? **(2)**

Solution Number of epochs = $\frac{5000 \times 50}{10000} = 25$

7. Consider a loss function $l[f]$, where $f = \beta + \Omega h$. Show that

$$\frac{dl}{d\Omega} = \frac{dl}{df} h^T$$

Note: h, β and f are vectors, while Ω is a matrix.

(5)

$f_i = \beta_i + \sum_{ij} \Omega_{ij} h_j$

$\Rightarrow \frac{\partial f_i}{\partial \Omega_{ij}} = h_j$

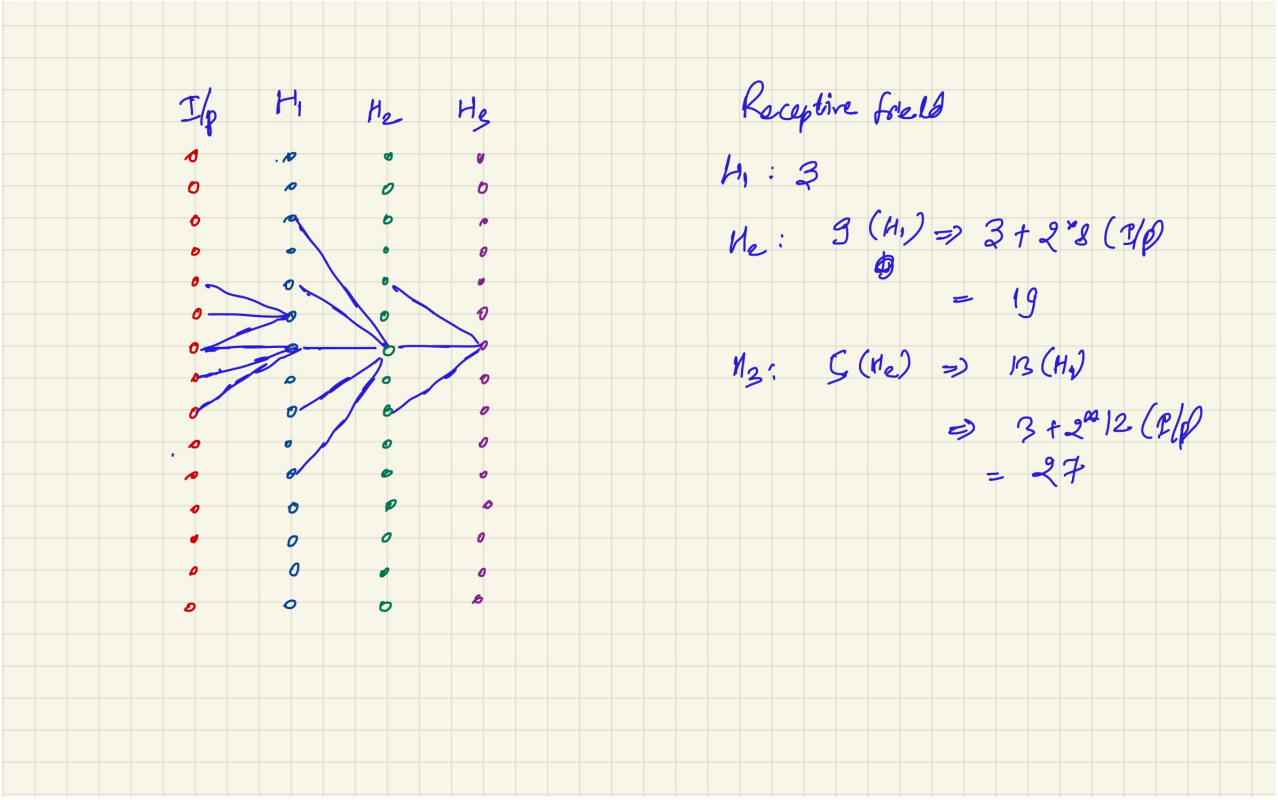
Chain Rule: $\Rightarrow \frac{\partial l}{\partial \Omega_{ij}} = \frac{\partial l}{\partial f_i} \frac{\partial f_i}{\partial \Omega_{ij}} = \frac{\partial l}{\partial f_i} \cdot h_j$

$\Rightarrow \frac{\partial l}{\partial \Omega} = \frac{\partial l}{\partial f} \underset{3\times 1}{\left[h^T \right]_{1\times 3}}$

for a 3×3 Ω

$l \rightarrow$ [$\frac{\partial l}{\partial f_i} h_j$] (j)

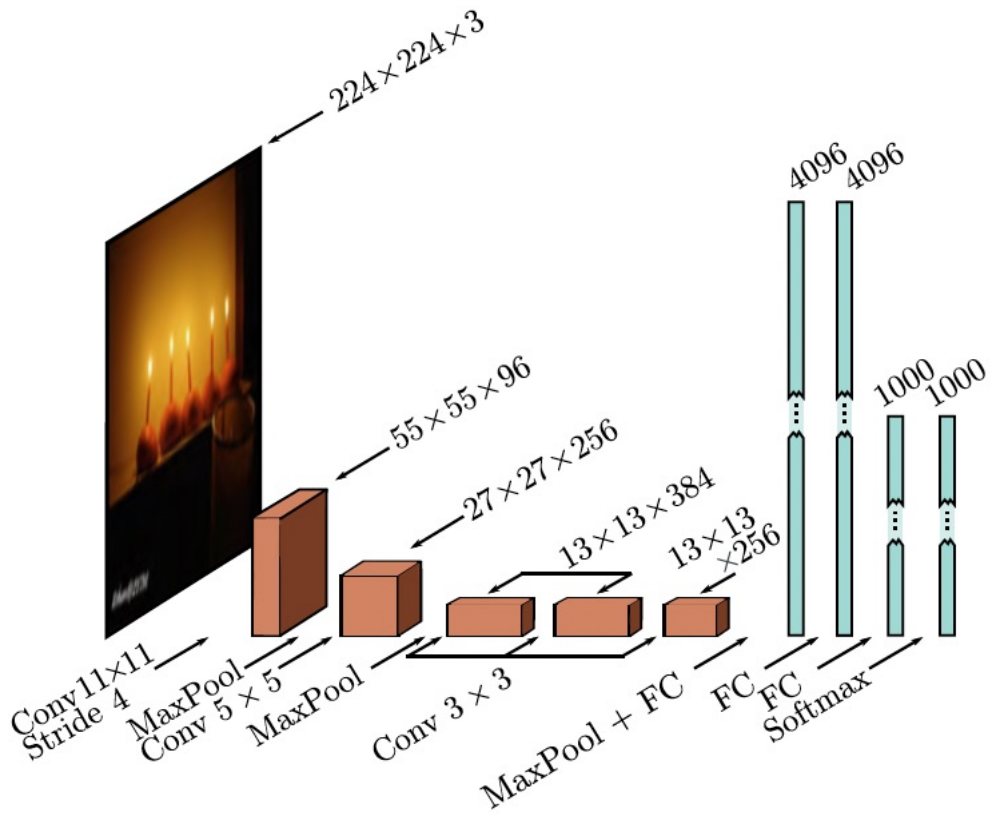
8. Consider a convolutional network with 1D input. The first hidden layer H_1 is computed using a convolution with kernel size three, stride two, and a dilation rate of one. The second hidden layer H_2 is computed using a convolution with kernel size five, stride one, and a dilation rate of two. The third hidden layer H_3 is computed using a convolution with kernel size three, stride two, and a dilation rate of two. What are the receptive field sizes at each hidden Layer? **(4)**



9. Suppose you are using the final feature map ($13 \times 13 \times 256$) from AlexNet to build region proposal network. You are using a sliding window of 3×3. Suppose that you are using 12 anchor boxes. How many additional parameters will be required for region proposal network on top of the last feature map? Assume that the fixed-dimension representation has 256 dimensions. **(3)**

Solution Number of parameters = $256 \times 3 \times 3 \times 256 + 256 \times 6 \times 12$

10. Consider the AlexNet network. How many parameters are used in each convolutional and fully connected layer? What is the total number of parameters? [All MaxPool layers are 3×3 with a stride of 2] **(6)**



- Between the image and first layer, there are $3 \times 96 \times 11 \times 11 = 34,848$ weights and 96 biases.
- Between the first layer and second layer, there are $96 \times 256 \times 5 \times 5 = 614,400$ weights 256 biases
- Between the second layer and third layer, there are $256 \times 384 \times 3 \times 3 = 884,736$ weights and 384 biases.
- Between the third layer and fourth layer, there are $384 \times 384 \times 3 \times 3 = 1,327,104$ weights and 384 biases
- Between the fourth layer and fifth layer, there are $384 \times 256 \times 3 \times 3 = 884,736$ weights and 256 biases
- At the end of the last convolutional layer, the representation is halved in size by the maxpool operation. Rounding, down, it now has size $6 \times 76 \times 256 = 9,216$. So there are $9,216 \times 4096 = 37,748,736$ weights and 4096 biases
- Between the next two fully connected layers, there are $4096 \times 4096 = 16,777,216$ weights and 4096 biases.
- Between the last two fully connected layers, there are $4096 \times 1000 = 4,096,000$ weights and 1000 biases.