SEOULTECH
Seoul National University of Science & Technology

Mid-term Solution

Date :

Code	ITM 528		Title	Deep Learning	
Time for Exam	3 hours	Questions	10	Weighting	35 %
Student's Number			Student's Name		

1. (5pts)

(a) False -> no rule-based approach (data-driven approach is suitable)

(b) True

(c) False -> no always guarantee to converge to the global optinum

(d) True

(e) True

2. (14pts)

(a) (10pts) Partial score: if appropriate formula derivation is

$$-\log P(\mathbf{y} \mid \mathbf{X}) = \sum_{i=1}^{n} \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \left(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)} - b\right)^2$$

(b) (4pts) Score when strictly describing changes in expressions and relationships with MSE in a fixed variable

- If we assume that σ is fixed, we can ignore the first term.
- In fact, as the solution does not depend on σ, minimizing the mean squared error is equivalent to the maximum likelihood estimation of a linear model under the assumption of additive Gaussian noise.

3. (5pts)

(a) (2pts) overfitting

(b) (3pts) dropout, weight decay, etc. (No score if simply referred to as regularization.)

Mid-term Solution

Date :

Code	ITM 528		Title	Deep Learning	
Time for Exam	3 hours	Questions	10	Weighting	35 %
Student’s Number			Student’s Name		

4. (20pts)

(a) (4pts) 3

(b) (4pts) 2

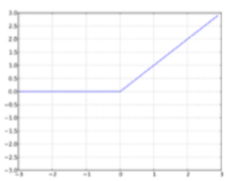
(c) (12pts) The final answer must be expressed in softmax. Description of gradient simulation must be included. Partial score: if appropriate formula derivation is

$$\partial_{o_j} l(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\exp(o_j)}{\sum_{k=1}^{q} \exp(o_k)} - y_j = \mathrm{softmax}(\mathbf{o})_j - y_j$$

The derivative is the difference between the probability assigned by our model, yhat, and the ground truth label, y. This fact makes computing gradients easy in practice.

5. (10pts)

(a) (2pts)



ReLU activation function helps to prevent the gradient vanishing problem.

(b) (5pts) Partial score: if appropriate calculation equations is

Forward pass:

$$h_1 = i_1 \times w_{11} + i_2 \times w_{21} + b_1 = 2.0 \times 1.0 - 1.0 \times 0.5 + 0.5 = 2.0$$
$$h_2 = i_1 \times w_{12} + i_2 \times w_{22} + b_2 = 2.0 \times -0.5 + -1.0 \times -1.0 - 0.5 = -0.5$$
$$h_3 = max(0, h_1) = h_1 = 2$$
$$h_4 = max(0, h_2) = 0$$
$$o_1 = h_3 \times w_{31} + h_4 \times w_{41} + b_3 = 2 \times 0.5 + 0 \times -0.5 - 1.0 = 0$$
$$o_2 = h_3 \times w_{32} + h_4 \times w_{42} + b_4 = 2 \times -1.0 + 0 \times 1.0 + 0.5 = -1.5$$

(c) (3pts)

$$MSE = \frac{1}{2} \times (t_1 - o_1)^2 + \frac{1}{2} \times (t_2 - o_2)^2 = 0.5 \times 1.0 + 0.5 \times 4.0 = 2.5$$

SEOULTECH
Seoul National University of Science & Technology

Mid-term Solution

Date :

Code	ITM 528		Title	Deep Learning	
Time for Exam	3 hours	Questions	10	Weighting	35 %
Student's Number			Student's Name		

6. (20pts, each 4pts) Processing the correct answer without considering the difference in Transpose.

- $\dfrac{\partial J}{\partial \mathbf{W}_2} = \dfrac{\partial J}{\partial \mathbf{z}_2}\dfrac{\partial \mathbf{z}_2}{\partial \mathbf{W}_2} = \delta_1 \dfrac{\partial \mathbf{z}_2}{\partial \mathbf{W}_2} = \delta_1^T \mathbf{h}_1^T \in \mathbb{R}^{n_c \times d_h}$
- $\dfrac{\partial J}{\partial \mathbf{b}_2} = \dfrac{\partial J}{\partial \mathbf{z}_2}\dfrac{\partial \mathbf{z}_2}{\partial \mathbf{b}_2} = \delta_1 \dfrac{\partial \mathbf{z}_2}{\partial \mathbf{b}_2} = \delta_1^T \in \mathbb{R}^{n_c \times 1}$
- $\dfrac{\partial J}{\partial \mathbf{W}_1} = \dfrac{\partial J}{\partial \mathbf{z}_1}\dfrac{\partial \mathbf{z}_1}{\partial \mathbf{W}_1} = \delta_2 \dfrac{\partial \mathbf{z}_1}{\partial \mathbf{W}_1} = \delta_2^T \mathbf{x}^T \in \mathbb{R}^{d_h \times d_x}$
- $\dfrac{\partial J}{\partial \mathbf{b}_1} = \dfrac{\partial J}{\partial \mathbf{z}_1}\dfrac{\partial \mathbf{z}_1}{\partial \mathbf{b}_1} = \delta_2 \dfrac{\partial \mathbf{z}_1}{\partial \mathbf{b}_1} = \delta_2^T \mathbf{I}^T = \delta_2^T \in \mathbb{R}^{d_h \times 1}$
- $\dfrac{\partial J}{\partial \mathbf{x}} = \dfrac{\partial J}{\partial \mathbf{z}_1}\dfrac{\partial \mathbf{z}_1}{\partial \mathbf{x}} = \mathbf{W}^T \delta_2^T \in \mathbb{R}^{d_x \times 1}$

7. (6pts)

(a) (2pts)

Padding helps preserve the spatial dimensions of the input after convolution, preventing the output from shrinking. It also allows for edge features to be captured better.

(b) (2pts)

Increasing the stride reduces the size of the output feature map, as the convolutional filter skips more elements in the input.

(c) (2pts)

Transposed Convolution is an operation used to reverse the downsampling effect caused by regular convolution. It takes a small input feature map and reconstructs it into a larger one, often called "upsampling."

Mid-term Solution

Date :

Code	ITM 528		**Title**	Deep Learning		
Time for Exam	**3 hours**	**Questions**		**10**	**Weighting**	**35 %**
Student's Number				**Student's Name**		

8. (6pts) You will be deducted when you write the correct answer

$$\begin{bmatrix}1 & 3\\5 & 6\end{bmatrix} \Rightarrow \max(1,3,5,6)=6$$

$$\begin{bmatrix}2 & 4\\7 & 8\end{bmatrix} \Rightarrow \max(2,4,7,8)=8$$

$$\begin{bmatrix}9 & 2\\4 & 3\end{bmatrix} \Rightarrow \max(9,2,4,3)=9$$

$$\begin{bmatrix}1 & 5\\6 & 7\end{bmatrix} \Rightarrow \max(1,5,6,7)=7$$

$$\text{Output}=\begin{bmatrix}6 & 8\\9 & 7\end{bmatrix}$$

9. (8pts)

(a) (3pts)

32x3x3x64 = 18,432

(b) (3pts)

32x1x1x5 = 160 5x3x3x64 = 2,880 Total = 3,040

(c) (2pts)

Reduce number of parameters, 18,432 >>>2,880.

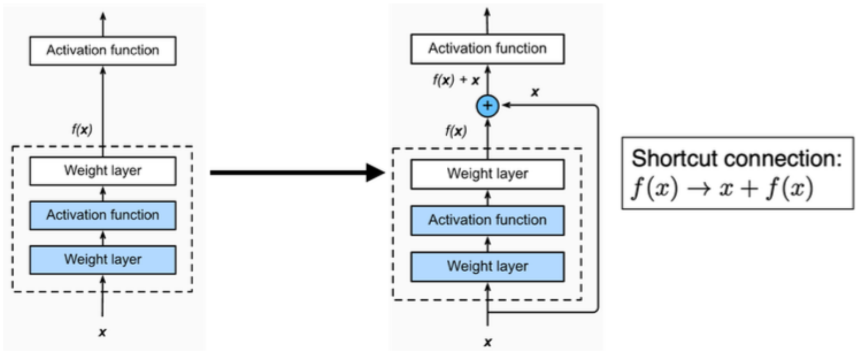
Mid-term Solution

Date :

Code	ITM 528		Title	Deep Learning	
Time for Exam	3 hours	Questions	10	Weighting	35 %
Student's Number			Student's Name		

10. (6pts)

(a) (3pts)



(b) (3pts)

In a residual block, we add what's called an identity map, also known as a shortcut connection. Instead of learning a new function completely from scratch, the shortcut allows the model to pass the input, identity x directly through to the output and only learn the difference, f(x). This makes training easier and prevents problems like degradation in very deep networks.