Day30: MidTerm

2022. 02. 17.
이창석

1. [20 points] For each of the following statements, indicate if it is *true* or *false*. A correct answer will get 2 points, but a wrong answer will get -2 points. No answer will get 0 point.

   (a) A softmax loss can become completely zero. (F)

   (b) When using dropout, which randomly removes nodes, it is applied in both the training and the testing times. (F)

   (c) A batch normalization has trainable parameters. (T)

   (d) A 1×1 convolution considers (local) spatial patterns. (F) dim reduction

   (e) A bias in a linear layer is a trainable parameter. (T)

   (f) Both saddle points and local optima have zero gradients. (?) True

   (g) An L1-norm regularized model gives a sparser solution than an L2-norm regularized model. (T)

   (h) During backpropagation, when the gradient flows backwards through the sigmoid (0~1) or tanh (-1~1) nonlinear units, it cannot change the sign of the gradient. (F) True

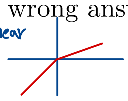
   (i) Dropout leads to sparsity in the trained weights. (F)

   (j) A tanh activation function has zero-centered outputs. (T)

2. [4 points] Which of the following are valid activation functions you could use in a neural network? (That is, which functions could be effective when training a neural network in practice?) A correct answer will get 1 points, but a wrong answer will get -1 points. No answer will get 0 point.

   (a, b, ~~d~~)

   V (a) $f(x) = max(0.25x, 0.75x)$ → non-Linear

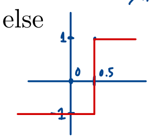
   V (b) $f(x) = min(0, x)$ ⇒ non-Linear

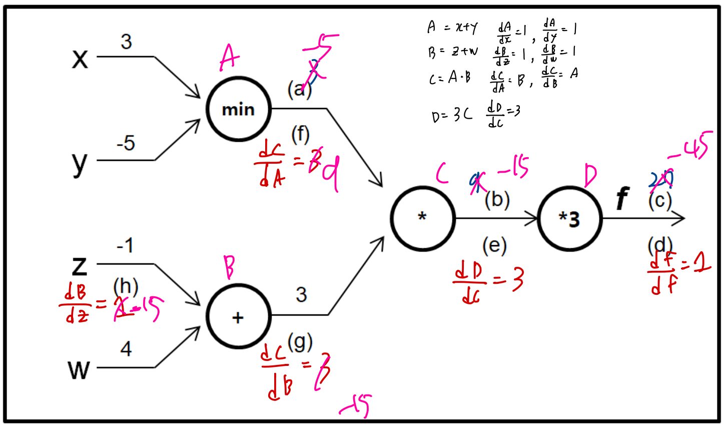
   (c) $f(x) = 0.7x$ ⇒ Linear

   (d) $f(x) = \begin{cases} 1 & \text{if } x > 0.5 \\ -1 & \text{else} \end{cases}$ ⇒ non-Linear (그러나 미분이 되지 않음)

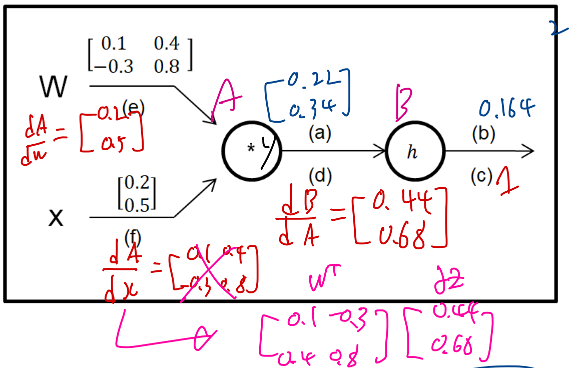
   

3. [4 points] (backpropagation) Fill in (a)-(h) in the figure below.



4. [6 points] (backpropagation) Fill in (a)-(f) in the figure below.
Please note that $h(\mathbf{y}) = \sum_{i=1}^{n} \mathbf{y}_i^2$, where $\mathbf{y} \in \mathcal{R}^{\mathbf{n}}$. Each answer can be a matrix, vector or scalar.

$2\begin{bmatrix} 0.1 & 0.4 \\ -0.3 & 0.8 \end{bmatrix}\begin{bmatrix} 0.2 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.02+0.2 \\ -0.06+0.4 \end{bmatrix} = \begin{bmatrix} 0.22 \\ 0.34 \end{bmatrix}$



$A = W \cdot x$ $\quad \frac{dA}{dW} = x, \quad \frac{dA}{dx} = W$

$B = y^2 A$ $\quad \frac{dB}{dA} = 2y$

↳ 0 $\quad W^T \begin{bmatrix} 0.1 & -0.3 \\ 0.4 & 0.8 \end{bmatrix} \quad \partial z \begin{bmatrix} 0.44 \\ 0.68 \end{bmatrix}$

5. [3 points] In a softmax classifier with ten classes, when randomly initializing the parameters, what would be approximately the initial loss function value? ? $-\log 10$

6. [3 points] In hyperparameter search, which candidate set are better to try? (b)

(a) 0.1, 0.2, 0.3, 0.4, ...) ⇒ Term이 큰 편.

√(b) 0.01, 0.03, 0.1, 0.3, ...

↳ log Scale

7. [3 points] Which of the two does Adam combines together? (b, c)

√(a) Momentum

√(b) Adagrad

√(c) RMSProp

(d) AdaDelta

$\frac{W-F+2P}{S}+1$

8. [8 points] Consider the following CNN architecture, and fill in the blank. Note that when the stride size is 2 in the convolution layer, ignore the last remaining row or column (this is how the first Max Pooling layer has $16 \times 16 \times 16$ input size, but not $17 \times 17 \times 16$).

Layers	Input Size	Filter Information	Number of Parameters
Convolution	$32 \times 32 \times 3$	16 5×5 filters with stride 2, padding 2	(a) $16 \times (5\times5\times3+1)$
Max Pooling	$16 \times 16 \times 16$	2×2 filters with stride 2, padding 0	(b) 0
Convolution	(c) $8 \times 8 \times 16$	32 5×5 filters with stride 2, padding 2	(d) $32 \times (5\times5\times16+1)$
Max Pooling	(e) $4 \times 4 \times 32$	3×3 filters with stride 2, padding 1	(f) 0
Convolution	(g) $2 \times 2 \times 32$	64 2×2 filters with stride 1, padding 0	(h) $64 \times (2\times2\times32+1)$
Fully connected	$1 \times 1 \times 64$	10D fully connected, softmax	$64 \times 10 + 10$

? $\frac{5-3+0}{1}+1$ conv $= 3$ $+2$ ⟹ conv ⟹ 1×1

$\frac{3-d}{1}+1$ -2 1 3 5 7 9 11

9. [6 points] Answer the following:

(a) Consider (1) the stack of five 3×3 convolution layers. To have the same receptive field size as this case, what would be the corresponding filter size of (2) a single convolution layer? ~~(15×15)~~ (11×11)

(b) Assume that the number of channels of the input and the output at each layer remains a constant c. Compute the number of parameters of each of the two cases (1) and (2). Ignore the bias term in your answer. (1) $5 \times (C \times C \times 3 \times 3)$ (2) $C \times C \times 11 \times 11$ ⟹ $121C^2$

$45C^2$

(c) Describe the two (or more) advantages of (1) compared to (2). 1) 신경망은 깊이가 깊을수록 성능이 좋음
2) Parameter의 수가 (2) 보다 적으므로, 메모리 효율적이고, overfitting 문제도 상대적으로 적게 발생함.

10. [3 points] List the following CNN architectures in a chronological order from the oldest to the newest: VGGNet, AlexNet, ResNet and GoogleNet

oldest ⟹ newest
(AlexNet → VGGNet → GoogleNet → ResNet)

?

11. [0 points] Given a binary activation map of the size $16 \times 16 \times 1$, where each element is either zero or one, consider the first convolutional layer composed of four convolutional filters of the size $3 \times 3 \times 1$ as

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Now, suppose you want to add the second convolutional layer with a single convolutional filter that can detect the input-domain pattern of

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Give such a single filter f of the size $3 \times 3 \times d$, where each element is either zero or one. That is, you first have to solve for the value of d (4), and give d number of 3×3 binary-valued matrices as the answer, i.e., $f(:,:,1) \in \{0,1\}^{3\times3}$, $f(:,:,2) \in \{0,1\}^{3\times3}$, ..., $f(:,:,d) \in \{0,1\}^{3\times3}$.