

Introduction to Machine Learning: Homework V

Due on Dec 20th, 2022 at 11:59pm

Professor Ziping Zhao

Bingnan Li
2020533092

1. [Deep Learning Models]

- (a) Consider a 3D convolution layer. Suppose the input size is $32 \times 32 \times 3$ (width, height, depth) and we use ten 5×5 (width, height) kernels to convolve with it. Set stride = 1 and pad = 2. What is the output size? Let the bias for each kernel be a scalar, how many parameters do we have in this layer?

Solution:

The input data shape is

$$[C_{in}, D_{in}, H_{in}, W_{in}] = [1, 3, 32, 32]$$

and the kernel size is

$$[C_{out}, D_{kernel}, H_{kernel}, W_{kernel}] = [10, 3, 5, 5]$$

Then by the formula, we have output data with a shape

$$[C_{out}, D_{out}, H_{out}, W_{out}]$$

and

$$\begin{aligned} D_{out} &= \frac{D_{in} + 2 \times pad - (D_{kernel} - 1) - 1}{stride} + 1 = 5 \\ H_{out} &= \frac{H_{in} + 2 \times pad - (H_{kernel} - 1) - 1}{stride} + 1 = 32 \\ W_{out} &= \frac{W_{in} + 2 \times pad - (W_{kernel} - 1) - 1}{stride} + 1 = 32 \end{aligned}$$

Thus, the output size is $[C_{out}, D_{out}, H_{out}, W_{out}] = [10, 5, 32, 32]$

Moreover, the total number of parameters is

$$\#\{parameters\} = \#\{parameters\ in\ kernel\} + \#\{biases\} = 5 \times 5 \times 3 \times 10 + 10 = 760$$

- (b) The convolution layer is followed by a max pooling layer with 2×2 (width, height) filter and stride = 2. What is the output size of the pooling layer? How many parameters do we have in the pooling layer?

Solution:

The input data shape is

$$[C_{in}, D_{in}, H_{in}, W_{in}] = [10, 5, 32, 32]$$

and the kernel shape is

$$[C_{out}, D_{kernel}, H_{kernel}, W_{kernel}] = [10, 5, 2, 2]$$

Then by the formula, we have output data with a shape

$$[C_{out}, D_{out}, H_{out}, W_{out}]$$

and

$$\begin{aligned} D_{out} &= \frac{D_{in} + 2 \times pad - (D_{kernel} - 1) - 1}{stride} + 1 = 1 \\ H_{out} &= \frac{H_{in} + 2 \times pad - (H_{kernel} - 1) - 1}{stride} + 1 = 16 \\ W_{out} &= \frac{W_{in} + 2 \times pad - (W_{kernel} - 1) - 1}{stride} + 1 = 16 \end{aligned}$$

Thus, the output size is $[C_{out}, D_{out}, H_{out}, W_{out}] = [10, 1, 16, 16]$.

Moreover, given that the max pooling layer only performs maximizing, there are no parameters in the pooling layer. Thus, the total number of parameters of the pooling layer is 0.

2. [Deep Learning Models]

- (a) (i.) When $r = 1$, \mathbf{W} is exactly the eigenvector of $\mathbf{X}\mathbf{X}^T$ corresponding to its largest eigenvalue.

Proof:

When $r = 1$, we know that $\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}$ is a scalar, which means

$$\text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) = \mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}$$

Hence, the optimization problem can be rewritten as follows:

$$\underset{\mathbf{W}}{\text{maximize}} \quad \mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W} \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{W} = 1$$

Then, the Lagrangian is:

$$\mathcal{L}(\mathbf{W}, \lambda) = \mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W} - \lambda \mathbf{W}^T \mathbf{W}$$

Taking the derivate of \mathcal{L} w.r.t \mathbf{W} and setting it to 0, we get

$$\begin{aligned} 2\mathbf{X}\mathbf{X}^T\mathbf{W}^* - 2\lambda^*\mathbf{W}^* &= 0 \\ \Rightarrow \\ \mathbf{X}\mathbf{X}^T\mathbf{W}^* &= \lambda^*\mathbf{W}^* \end{aligned}$$

Thus, λ^* is eigenvalue of $\mathbf{X}\mathbf{X}^T$ and \mathbf{W}^* is the corresponding eigenvector. Moreover, since

$$\mathbf{W}^{*T} \mathbf{X} \mathbf{X}^T \mathbf{W}^* = \lambda^*$$

Therefore, we need to get the largest eigenvalue λ^* to maximize $\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}$, which means \mathbf{W}^* is the eigenvector corresponding to the largest eigenvalue.

- (ii.) \mathbf{W} is also the solution to

$$\underset{\mathbf{W}}{\text{minimize}} \quad \|\mathbf{X} - \mathbf{W}\mathbf{W}^T \mathbf{X}\|_F^2 \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}_r$$

Proof:

Given that $\|\mathbf{P}\|_F = \sqrt{\text{Tr}(\mathbf{P}^T \mathbf{P})}$, we get

$$\begin{aligned} \|\mathbf{X} - \mathbf{W}\mathbf{W}^T \mathbf{X}\|_F^2 &= \text{Tr}[(\mathbf{X} - \mathbf{W}\mathbf{W}^T \mathbf{X})^T (\mathbf{X} - \mathbf{W}\mathbf{W}^T \mathbf{X})] \\ &= \text{Tr}[\mathbf{X}^T \mathbf{X} - \mathbf{X}^T \mathbf{W}\mathbf{W}^T \mathbf{X} - \mathbf{X}^T \mathbf{W}\mathbf{W}^T \mathbf{X} + \mathbf{X}^T \mathbf{W}\mathbf{W}^T \mathbf{W}\mathbf{W}^T \mathbf{X}] \\ &= \text{Tr}[\mathbf{X}^T \mathbf{X} - \mathbf{X}^T \mathbf{W}\mathbf{W}^T \mathbf{X}] \\ &= \text{Tr}(\mathbf{X}^T \mathbf{X}) - \text{Tr}(\mathbf{X}^T \mathbf{W}\mathbf{W}^T \mathbf{X}) \\ &= \text{Tr}(\mathbf{X}^T \mathbf{X}) - \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \end{aligned}$$

Since the first term of the equation has nothing to do with \mathbf{W} , the optimization problem can be rewritten as follows:

$$\underset{\mathbf{W}}{\text{minimize}} \quad -\text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}_r$$

which is equivalent to the PCA problem.

(b) (i.) \mathbf{A}_2^* can be solved from:

$$\underset{\mathbf{A}_2}{\text{minimize}} \quad \|\mathbf{X} - \mathbf{A}_2 \mathbf{A}_2^\dagger \mathbf{X}\|_F^2$$

(ii.) The solution \mathbf{W} from (a) can be taken as the same as \mathbf{A}_2^* .

3. [Ensemble Learning] Suppose there are L independent two-class classifiers used for simple voting and the output of classifier j ($j = 1 \cdots L$) is denoted as d_j . From the point of view that the mean squared error of an estimator can be decomposed into the bias part and the variance part, explain why increasing L can lead to an increase in classification accuracy.
4. [Model Assessment and Selection] Suppose we carry out a K -fold cross-validation on a dataset and obtain the classification error rates $\{p_i\}_{i=1}^K$, describe the steps of a one-sided t test on testing the null hypothesis H_0 that the classifier has error percentage p_0 or less at a significance level α .