

ml-architecture-proj

1. Given the convolutional neural network block as below

Given the input feature maps $\mathbf{X} \in \mathbb{R}^{64 \times 64 \times 128}$, all convolutional layers perform zero-padding of 1 on each side of H and W dimensions.

- (a) What is the total number of parameters in the block (you can skip bias terms)?

$$K \times K \times C_{in} \times C_{out} = 3 \times 3 \times 128 \times 256 + 3 \times 3 \times 256 \times 512 = 1474560$$

- (b) What is the total number of multi-add operations in the block?

$$K \times K \times M \times N \times W' \times D'$$

$$W' = (W - K + 2P) / S + 1$$

$$W' = (64 - 3 + 2) / 2 + 1 = 32.5 = 32$$

$$W'' = (32 - 3 + 2) / 1 + 1 = 32$$

$$3 \times 3 \times 128 \times 256 \times 32 \times 32 + 3 \times 3 \times 256 \times 512 \times 32 \times 32 = 1509949440$$

- (c) What is memory requirement change to store the input and output features of this block (Use percentage)?

Input : $64 \times 64 \times 128$, final output: $32 \times 32 \times 512$, use the same amount of memory

2. Using batch normalization in neural networks requires computing the mean and variance of a tensor. Suppose a batch normalization layer takes vectors z_1, z_2, \dots, z_m as input, where m is the mini-batch size. It computes $\hat{z}_1, \hat{z}_2, \dots, \hat{z}_m$ according to

$$\hat{z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

where

$$\mu = \frac{1}{m} \sum_{i=1}^m z_i, \quad \sigma^2 = \frac{1}{m} \sum_{i=1}^m (z_i - \mu)^2.$$

It then applies a second transformation to obtain $\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_m$ using learned parameters γ and β as

$$\tilde{z}_i = \gamma \hat{z}_i + \beta.$$

In this question, you can assume that $\epsilon = 0$.

- (a) You forward-propagate a mini-batch of $m = 4$ examples in your network. Suppose you are at a batch normalization layer, where the immediately previous layer is a fully connected layer with 3 units. Therefore, the input to this batch normalization layer can be represented as the below matrix:

$$\begin{bmatrix} 12 & 14 & 14 & 12 \\ 0 & 10 & 10 & 0 \\ -5 & 5 & 5 & -5 \end{bmatrix}$$

What are \hat{z}_i ? Please express your answer in a 3×4 matrix.

$$\mu_1 = 13, \mu_2 = 5, \mu_3 = 0, \sigma_1 = 1, \sigma_2 = 5, \sigma_3 = 5,$$

$$\begin{bmatrix} -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \end{bmatrix}$$

- (b) Continue with the above setting. Suppose $\gamma = (1, 1, 1)$, and $\beta = (0, -10, 10)$. What are \tilde{z}_i ? Please express your answer in a 3×4 matrix.

$$\begin{bmatrix} -1 & 1 & 1 & -1 \\ -11 & -9 & -9 & -11 \\ 9 & 11 & 11 & 9 \end{bmatrix}$$

- (c) Describe the differences of computations required for batch normalization during training and testing.

During training time, the batch normalization should be applied as usual. During test (or inference) time, the mean and the variance are fixed. They are estimated using the previously calculated means and variances of each training batch.

- (d) Describe how the batch size during testing affect testing results.

Mini-batch can prevent over-fitting and work well in the large scale problem. If we increase the batch size, the over-fitting may occur and lead to low accuracy in the testing result.

3. We investigate the back-propagation of the convolution using a simple example. In this problem, we focus on the convolution operation without any normalization and activation function. For simplicity, we consider the convolution in 1D cases. Given 1D inputs with a spatial size of 4 and 2 channels, *i.e.*,

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \in \mathbb{R}^{2 \times 4}, \quad (1)$$

we perform a 1D convolution with a kernel size of 3 to produce output Y with 2 channels. No padding is involved. It is easy to see

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \in \mathbb{R}^{2 \times 2}, \quad (2)$$

where each row corresponds to a channel. There are 12 training parameters involved in this convolution, forming 4 different kernels of size 3:

$$W^{ij} = [w_1^{ij}, w_2^{ij}, w_3^{ij}], i = 1, 2, j = 1, 2, \quad (3)$$

where W^{ij} scans the i -th channel of inputs and contributes to the j -th channel of outputs.

(a) Now we flatten X and Y to vectors as

$$\tilde{X} = [x_{11}, x_{12}, x_{13}, x_{14}, x_{21}, x_{22}, x_{23}, x_{24}]^T$$

$$\tilde{Y} = [y_{11}, y_{12}, y_{21}, y_{22}]^T$$

Please write the convolution in the form of fully connected layer as $\tilde{Y} = A\tilde{X}$ using the notations above. You can assume there is no bias term.

Hint: Note that we discussed how to view convolution layers as fully connected layers in the case of single input and output feature maps. This example asks you to extend that to the case of multiple input and output feature maps.

$$\begin{aligned}
 & (a) \quad \begin{bmatrix} w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{12} & w_2^{12} & w_3^{12} & 0 \\ 0 & w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{12} & w_2^{12} & w_3^{12} \\ w_1^{21} & w_2^{21} & w_3^{21} & 0 & w_1^{22} & w_2^{22} & w_3^{22} & 0 \\ 0 & w_1^{21} & w_2^{21} & w_3^{21} & 0 & w_1^{22} & w_2^{22} & w_3^{22} \end{bmatrix} \times \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}
 \end{aligned}$$

(b) Next, for the back-propagation, assume we've already computed the gradients of loss L with respect to \tilde{Y} :

$$\frac{\partial L}{\partial \tilde{Y}} = \left[\frac{\partial L}{\partial y_{11}}, \frac{\partial L}{\partial y_{12}}, \frac{\partial L}{\partial y_{21}}, \frac{\partial L}{\partial y_{22}} \right]^T, \quad (4)$$

Please write the back-propagation step of the convolution in the form of $\frac{\partial L}{\partial \tilde{X}} = B \frac{\partial L}{\partial \tilde{Y}}$. Explain the relationship between A and B .

$$\begin{aligned}
 & (b) \quad \frac{\partial L}{\partial \tilde{X}} = B \frac{\partial L}{\partial \tilde{Y}}, \quad B = \frac{\partial \tilde{Y}}{\partial \tilde{X}}, \quad A = \frac{\partial \tilde{Y}}{\partial \tilde{X}} \quad \text{According to chain rule,} \\
 & \text{the implicit function theorem define the relationship between } \tilde{X} \& \tilde{Y}, \\
 & A \text{ should be equal to } B
 \end{aligned}$$

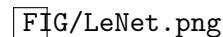
- (c) While the forward propagation of the convolution on X to Y could be written into $\tilde{Y} = A\tilde{X}$, could you figure out whether $\frac{\partial L}{\partial \tilde{X}} = B \frac{\partial L}{\partial \tilde{Y}}$ also corresponds to a convolution on $\frac{\partial L}{\partial \tilde{Y}}$ to $\frac{\partial L}{\partial \tilde{X}}$? If yes, write down the kernels for this convolution. If no, explain why.

If we know the function, then we can put the appropriate weight on it, so we can know what Y will be.

4. **LeNet for Image Recognition:** In this coding assignment, you will need to complete the implementation of LeNet (LeCun Network) using PyTorch and apply the LeNet to the image recognition task on Cifar-10 (10-classes classification). You will need to install the python packages “tqdm” and “pytorch”. Please read the installation guides of PyTorch here (<https://pytorch.org/get-started/locally/>). You are expected to implement your solution based on the given codes. The only file you need to modify is the “solution.py” file. You can test your solution by running the “main.py” file.

- (a) Complete the class **LeNet()**. In particular, define operations in function **__init__()** and use them in function **forward()**. The input of **forward()** is an image. The paper for LeNet can be found here (<http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf>)

The network architecture is shown in the figure below.



The sub-sampling is implemented by using the max pooling. And the kernel size for all the convolutional layers are 5×5 . Please use **ReLU** function to activate the outputs of convolutional layers and the first two fully-connected layers. The sequential layers are:

Inputs \rightarrow
 Convolution (6 out channels) \rightarrow Max Pooling \rightarrow
 Convolution (16 out channels) \rightarrow Max Pooling \rightarrow
 Reshape to vector \rightarrow Fully-connected (120 out units) \rightarrow
 Fully-connected (84 out units) \rightarrow Outputs (n-classes out units)

For this part, you are only allowed to use the APIs in **torch.nn**. Please refer to the PyTorch API documents below for the usage of those APIs before you use them:
<https://pytorch.org/docs/stable/nn.html>.

Run the model by “**python main.py**” and report the testing performance as well as a short analysis of the results.

- (b) Add batch normalization operations after each max pooling layer. Run the model by “**python main.py**” and report the testing performance as well as a short analysis of the results.
- (c) Based on (b), add dropout operations with drop rate of 0.3 after the first two fully-connected layers. Run the model by “**python main.py**” and report the testing performance as well as a short analysis of the results.