Chapter6

Rachel Hu

April 29, 2019

```
In [1]: import d2l
        import mxnet as mx
        from mxnet import autograd, gluon, init, nd
        from mxnet.gluon import loss as gloss, data as gdata, nn

        import time

        import numpy as np

        from IPython.core.interactiveshell import InteractiveShell
        InteractiveShell.ast_node_interactivity = "all"
```

0.0.1 Chapter 6

6.1 Exercises

1. **Assume that the size of the convolution mask is $\Delta = 0$. Show that in this case the convolutional mask implements an MLP independently for each set of channels.**

   For any tensor k, since $\Delta = 0$, the learner are independent. i.e.

   $$h[i,j,k] = \sum_{a=-\Delta}^{\Delta} \sum_{b=-\Delta}^{\Delta} V[a,b,k] \cdot x[i+a, j+b] = V[0,0,k] \cdot x[i,j]$$

2. **Why might translation invariance not be a good idea after all? Does it make sense for pigs to fly?**

   In translation invariance, we assume that we would recognize an object wherever it is in an image. It is only reasonable to assume that the location of the object shouldn't matter too much to determine whether the object is there. For example, a face is still a face regardless of whether it is moved horizontally or vertically in an image.**

3. **What happens at the boundary of an image?**

   On the boundaries we encounter the problem that we keep on losing pixels. (Without padding)

4. **Derive an analogous convolutional layer for audio.**

   Mel spectrogram transform the input raw sequence to a 2D feature map where one dimension represents time and the other one represents frequency and the values represents amplitude. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Moving a sound event horizontally offsets its position in time and it can be argued that a sound event means the same thing regardless of when it happens. However, moving a sound vertically might influence its meaning: Moving the frequencies of a male voice upwards could change its meaning from man to child or goblin, for example.

5. **What goes wrong when you apply the above reasoning to text? Hint - what is the structure of language?**

   Language is a sequence data, hence we cannot assume a translation invariance, i.e. the location of each word matters. Sequence modeling is more effective.

6. **Prove that f⊛g=g⊛f.**

   Let $x - z = y$, then

$$[f \circledast g](x) = \int_{\mathbb{R}^d} f(z)g(x-z)dz = \int_{\mathbb{R}^d} f(x-y)g(y)dy = [g \circledast f](x)$$

6.2 Exercises

1. **Construct an image X with diagonal edges.**
   - **What happens if you apply the kernel K to it?**
   - **What happens if you transpose X?**
   - **What happens if you transpose K?**

2. **When you try to automatically find the gradient for the Conv2D class we created, what kind of error message do you see?** ??????

3. **How do you represent a cross-correlation operation as a matrix multiplication by changing the input and kernel arrays?** In the two-dimensional cross-correlation operation, the convolution window starts from the top-left of the input array, and slides in the input array from left to right and top to bottom. [See details in 6.2.1]

4. **Design some kernels manually. •What is the form of a kernel for the second derivative?**

   • What is the kernel for the Laplace operator?

   2D discrete Laplace kernel:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

   1D discrete Laplace kernel:

$$\begin{pmatrix} 1 & -2 & 1 \end{pmatrix}$$

   • What is the kernel for an integral?

   A general integral transform is defined by $g(\alpha) = \int_a^b f(t)K(\alpha, t)dt$, where $K(\alpha, t)$ is called integral kernel.

   • What is the minimum size of a kernel to obtain a derivative of degree d?

   (2d-1) x (2d-1)

6.3 Exercises

1. **For the last example in this section, use the shape calculation formula to calculate the output shape to see if it is consistent with the experimental results.**

2. **Try other padding and stride combinations on the experiments in this section.**

3. **For audio signals, what does a stride of 2 correspond to?**

   In the time dimension, stride of 2 aggregates every 2 timestamps.

   In the frequency dimension, ...

4. **What are the computational benefits of a stride larger than 1.**

   Save memory and reduce computational time.

6.4 Exercises

1. **Assume that we have two convolutional kernels of size k1 and k2 respectively (with no nonlinearity in between).**

   • Prove that the result of the operation can be expressed by a single convolution.

   Convolution associate law.

   • What is the dimensionality of the equivalent single convolution?

   k1+k2-1

   • Is the converse true?

2. **Assume an input shape of ci $\times$h$\times$w and a convolution kernel with the shape co $\times$ci $\times$kh $\times$kw, padding of (ph,pw), and stride of (sh,sw).**

   • What is the computational cost (multiplications and additions) for the forward computation?

   • What is the memory footprint?

   $O(c_i c_o k_h k_w m_h m_w)$

   • What is the memory footprint for the backward computation?

   • What is the computational cost for the backward computation?

3. **By what factor does the number of calculations increase if we double the number of input channels ci and the number of output channels co? What happens if we double the padding?**

4. **If the height and width of the convolution kernel is $k_h = k_w = 1$, what is the complexity of the forward computation?**

   $O(c_i * c_o * h * w)$

5. **Are the variables Y1 and Y2 in the last example of this section exactly the same? Why?**

   Yes. The main computation of the 1×1 convolution occurs on the channel dimension. And $k_h k_w == 1$ in both function.

6. **How would you implement convolutions using matrix multiplication when the convolution window is not 1$\times$1 ?**

6.5 Exercises

1. **Implement average pooling as a convolution.**

   6.5.1

2. **What is the computational cost of the pooling layer? Assume that the input to the pooling layer is of size c×h×w , the pooling window has a shape of ph×pw with a padding of (p_h, p_w) and a stride of (s_h, s_w).**

   $c \times [(I_h - p_h + pad_h + s_h)/s_h] \times [(l_w - p_w + pad_w + s_w)/s_w]$

3. **Why do you expect maximum pooling and average pooling to work differently?**

   - Max pooling: the strongest pattern signal in a window
   - Average pooling: The average signal strength in a window

4. **Do we need a separate minimum pooling layer? Can you replace it with another operation?**

   $argmin(Xa_{ij}) = argmax(-1 * Xa_{ij})$, hence min-pooling can be modeling through CNN and max-pooling

5. **Is there another operation between average and maximum pooling that you could consider (hint - recall the softmax)? Why might it not be so popular?**

   A pooling layer is to alleviate the excessive sensitivity of the convolutional layer to location., i.e. reduce the resolution. Softmax computation cost is too high.

6.6 Exercises

1. **Replace the average pooling with max pooling. What happens?**

   Max pooling can easily detect the edge of feature.

2. **Try to construct a more complex network based on LeNet to improve its accuracy.**

   a. Adjust the convolution window size.
   b. Adjust the number of output channels.
   c. Adjust the activation function (ReLU?).
   d. Adjust the number of convolution layers.
   e. Adjust the number of fully connected layers.
   f. Adjust the learning rates and other training details (initialization, epochs, etc.)
   g. Try out the improved network on the original MNIST dataset.

3. **Display the activations of the first and second layer of LeNet for different inputs (e.g. sweaters, coats).**