Computer Vision 1 - Master AI
Tutorial Deep Video

Pascal Mettes and Theo Gevers

1 3D convolutions

Consider 5 gray-scale videos, each with 16 frames of size 224x224. You want to generate a feature map with 64 output channels using 3x3x3 convolutional filters. Proper padding is applied.

1.a What is the dimension of the output weight tensor the 3D layer? How many parameters are there in total?

1.b What is the dimensionality of the output feature maps if stride = 1 for all 3 dimensions?

1.c What is the dimensionality of the output feature maps if stride = 2 for the temporal dimension and stride=1 for the spatial dimensions?

1.d Suppose you have the following 3x3x3 kernel (no bias):

$$K_1 = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} K_2 = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} K_3 = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix}$$

Compute the output feature maps for the following 4x4x4 input:

$$I_1 = \begin{bmatrix} 4 & 4 & 3 & 0 \\ 3 & 4 & 2 & 3 \\ 2 & 3 & 1 & 1 \\ 1 & 4 & 3 & 1 \end{bmatrix} I_2 = \begin{bmatrix} 1 & 3 & 4 & 3 \\ 1 & 4 & 1 & 4 \\ 4 & 1 & 4 & 4 \\ 4 & 0 & 1 & 2 \end{bmatrix} I_3 = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 2 & 4 & 1 & 3 \\ 3 & 3 & 1 & 4 \\ 1 & 2 & 3 & 0 \end{bmatrix} I_4 = \begin{bmatrix} 2 & 1 & 0 & 4 \\ 3 & 1 & 1 & 0 \\ 3 & 1 & 1 & 2 \\ 4 & 1 & 3 & 4 \end{bmatrix}$$

1.e Suppose that a standard 3D convolution layer takes an input feature matrix F of shape (l_F, w_F, h_F, c_F) and yields a feature matrix G of size (l_G, w_G, h_G, c_G), where c denotes the number of channels. Also suppose the kernel size is $k \times k \times k$. What is the computational cost of a 3D convolution here?

1.f For which kinds of motions are 3D convolutions suitable and for which ones not?

1.g What is an advantage of (2+1)D convolutions over 3D convolutions?

2 Standard RNN

Consider the same 5 videos as for the first question. Now, we want to learn temporal dependencies through RNNs. We have an RNN with one layer, no bias, and 20 hidden neurons. We make each frame a vector before feeding it to the RNN.

2.a What is the function for the hidden layer of a standard RNN?

2.b What is the input dimensionality of the RNN for all 5 videos?

2.c How many parameters does the RNN layer have?

2.d What is the output dimensionality for all 5 videos?

2.e For this example, the frames are vectorized. What are disadvantages of this approach and can you name a solution?

3 Open questions

The following questions do not have strict correct/incorrect answers. What cover research topics that are not fully explored in research. Keep an open mind and use your intuition to come to interesting answers to the questions.

3.a We have seen that for actions with a larger inter-class separation (e.g. *skateboarding* versus *diving*), objects can be used to classify videos into the correct action category. Can you come with a blueprint for a solution to classify unseen actions when the inter-class separation is small (e.g. *throwing* versus *catching* a ball in the same video)?

3.b Current deep networks obtain high classification accuracies but seem to do so by focusing on context, rather than on the actors performing the action. Can you come with a blueprint for a solution to make deep networks focus on actors and their spatio-temporal patterns, rather than the context in which actions occur?

3.c With action recognition making strides every new conference, we are getting closer to real-world applications. Where do you think that society can benefit greatly from action recognition? What open problems need to be addressed to make that happen? And what concerns may arrise from the application?