Armin Ghojehzadeh

Theory of second practical question

2.1

Normal and deformable convolution networks are two types of convolutional neural networks (CNNs) that differ in how they sample the input feature maps. Convolution is a process of applying a filter (also called a kernel) to the input feature maps to produce output feature maps. The filter slides over the input feature maps and computes the dot product between the filter weights and the input values at each location. The locations where the filter is applied are determined by a regular grid, which is defined by the filter size, stride, and padding. Normal convolution networks use a fixed and regular grid to sample the input feature maps. This means that the filter is applied to the same locations for every input feature map, regardless of the content or shape of the input. This limits the ability of normal convolution networks to model geometric transformations, such as scaling, rotation, and deformation, that may occur in the input images.

Deformable convolution networks add 2D offsets to the regular grid sampling locations in the standard convolution. This means that the filter can be applied to different locations for different input feature maps, depending on the learned offsets. The offsets are learned from the preceding feature maps, via additional convolutional layers. This enables free form deformation of the sampling grid, which enhances the transformation modeling capability of deformable convolution networks. Deformable convolution networks can adapt to the shape and content of the input images, and handle complex geometric transformations more effectively.

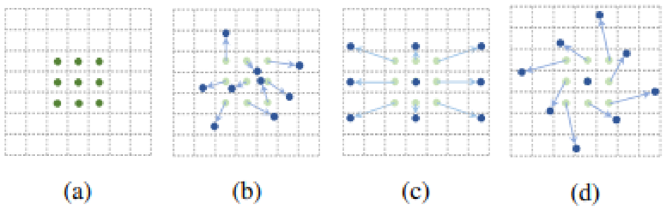


Figure 1: Illustration of the sampling locations in 3 × 3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.

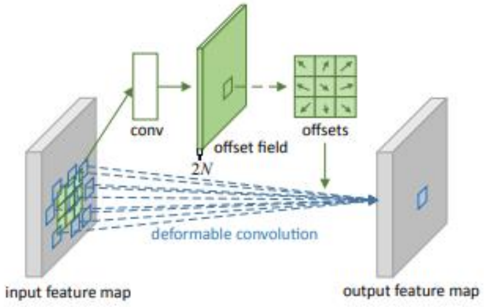


Figure 2: Illustration of 3 × 3 deformable convolution.

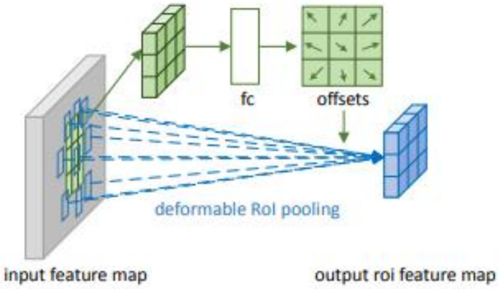


Figure 3: Illustration of 3 × 3 deformable RoI pooling.

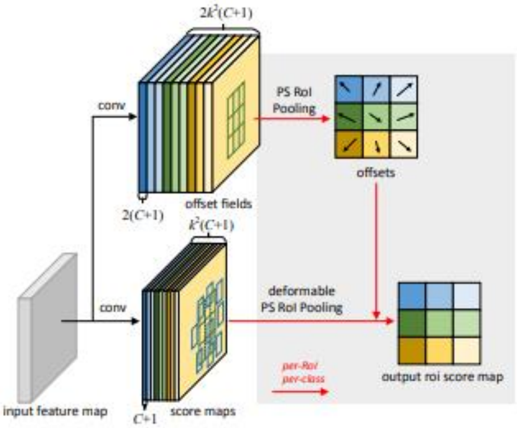
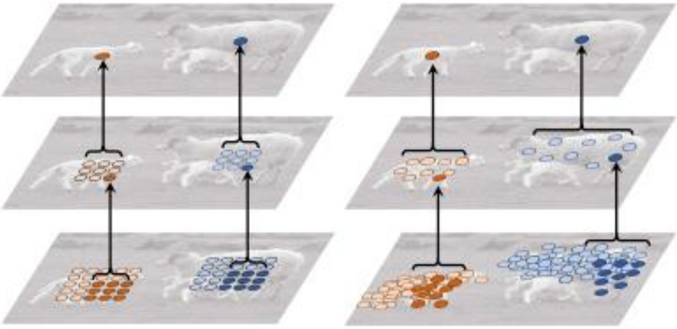


Figure 4: Illustration of 3 × 3 deformable PS RoI pooling.



(a) standard convolution (b) deformable convolution

Figure 5: Illustration of the fixed receptive field in standard convolution (a) and the adaptive receptive field in deformable convolution (b), using two layers. Top: two activation units on the top feature map, on two objects of different scales and shapes. The activation is from a 3 × 3 filter. Middle: the sampling locations of the 3 × 3 filter on the preceding feature map. Another two activation units are highlighted. Bottom: the sampling locations of two levels of 3 × 3 filters on the preceding feature map. The highlighted locations correspond to the highlighted units above.

As you can see, normal convolution applies the filter to the same locations for every input feature map, while deformable convolution applies the filter to different locations for different input feature maps, based on the learned offsets. This allows deformable convolution to capture the shape and content of the input images more accurately, and to model geometric transformations more effectively.

2.2

Deformable CNN can create flexibility in geometric transformation in images by allowing the convolution filter to adapt to the shape and content of the input images. By adding learned offsets to the regular grid sampling locations, deformable CNN can deform the sampling grid to match the input images more closely. This way, deformable CNN can handle complex geometric transformations, such as scaling, rotation, and deformation, that may occur in the input images. Deformable CNN can also learn the offsets from the preceding feature maps, which means that the offsets are not fixed but depend on the input images. This makes deformable CNN more flexible and robust to geometric variations in the input images.

2.3

Simple convolutional networks have serious problems when dealing with images where the image objects have a lot of spatial change or rotation because they use a fixed and regular grid to sample the input feature maps. This means that the convolution filter is applied to the same locations for every input image, regardless of the content or shape of the input. This limits the ability of simple convolutional networks to model geometric transformations, such as scaling, rotation, and deformation, that may occur in the input images. As a result, simple convolutional networks may fail to recognize the same object if it appears in different positions, orientations, or sizes in the image space. To overcome this problem, some techniques have been proposed to extend convolutional networks to handle nontrivial geometric transformations, such as deformable convolution networks, geometric convolutional networks, and geometric graph convolutional networks. These techniques allow the convolution filter to adapt to the shape and content of the input images, and to capture the geometric variations more effectively.

2.4

The offsets in deformable CNN are calculated by additional convolutional layers that take the preceding feature maps as input. The additional convolutional layers have the same number of output channels as the number of sampling locations in the standard convolution. For example, if the standard convolution uses a 3x3 filter, then the additional convolutional layers have 18 output channels, corresponding to the 2D offsets for the 9 sampling locations. The additional convolutional layers learn to predict the offsets from the target tasks, without additional supervision. The offsets are different for each location and each input feature map, and they can be positive or negative, fractional or integral.