Student Name: Mohammad Mohammad Beigi
Student ID: 99102189
Subject: Deep Learning

Deep Learning - Dr. E. Fatemizadeh
Assignment 3 - Question 2 -CNN and Vision

part 1

In a traditional CNN, convolutional layers use fixed-size filters (kernels) to scan input images. These filters slide over the input image with a fixed grid, performing convolutions at regular intervals.

Traditional CNNs use fixed grids for sampling, but DCNNs allow the network to learn spatial transformations and sample from different locations based on the content of the input image.

While traditional CNNs use fixed grids, DCNNs introduce deformable convolutions to dynamically adjust the sampling locations. This adaptability can be beneficial in tasks where objects may have varying shapes, sizes, or positions within the input images.

part 2

1. **Spatially Adaptive Sampling:**
   - The deformable convolutional layer uses the predicted offsets to adaptively sample input values from different positions in the input feature map.
   - This adaptive sampling allows the network to focus on relevant regions and adjust its receptive field for each position, providing flexibility in handling geometric variations.
2. **Enhanced Robustness:**
   - The flexibility introduced by deformable convolutions makes DCNNs more robust to variations in object shape, size, and position within the input images.
   - This adaptability helps the network generalize well to different geometric transformations, making it more effective in handling real-world scenarios.

In summary, deformable convolutions in DCNNs allow the network to dynamically adjust its receptive field based on the characteristics of the input data. This adaptability enhances the model's ability to handle geometric transformations, making it more flexible and robust in tasks where objects may undergo various spatial deformations.

part 3

(complete answer in previous parts)

Deformable Convolutional Networks (DCNNs) incorporate deformable convolutional layers, offering advantages for robustness to rotated images. These layers provide adaptive receptive fields, dynamically adjusting to focus on relevant features during convolution. Improved feature localization is achieved through learnable offsets, enhancing the network's ability to capture rotated patterns. DCNNs reduce sensitivity to rotation angles, leading to better generalization across orientations. Beyond rotation, they increase robustness to various geometric variations, learning spatial relationships more effectively and enabling the capture of richer feature representations crucial for tasks like object recognition.

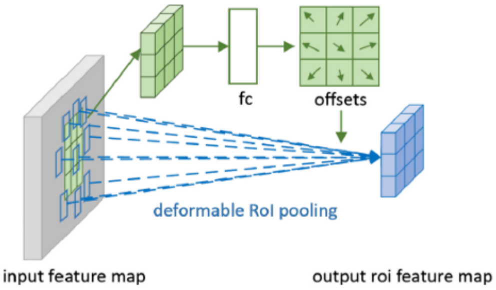
part 4

In the deformable convolution module and deformable RoI pooling module, the offsets are calculated by applying a convolutional layer over the input feature map. The convolution kernel has the same spatial resolution and dilation as the current convolutional layer. The output offset fields have the same spatial resolution as the input feature map. During training, both the convolutional kernels for generating the output features and the offsets are learned simultaneously.

In DCNNs:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n).$$

In deformable RoI pooling module:



deformable RoI Pooling

Regular RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p}) / n_{ij}$$

Deformable RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p} + \Delta\mathbf{p}_{ij}) / n_{ij}$$

where $\Delta\mathbf{p}_{ij}$ is generated by a sibling fc branch

$$\Delta\mathbf{p}_{ij} = \gamma \cdot \Delta\widehat{\mathbf{p}}_{ij} \circ (w,h)$$