Spatial Pyramid Pooling Network (SPPNet)

Mayank Nagar
MDS202334

1 Introduction

1.1 Technical Issue in CNN Training and Testing

There is a fundamental technical issue in the training and testing of Convolutional Neural Networks (CNNs):

- **Fixed Input Size Requirement:** Most prevalent CNN architectures require a fixed input image size (e.g., 224×224), which limits both the aspect ratio and scale of the input image.
- **Handling Arbitrary Image Sizes:** When applied to images of arbitrary sizes, current methods typically fit the input image to the fixed size using:
  - **Cropping:** A region is cropped from the image, but this may exclude important object features.
  - **Warping:** The image is resized to fit the fixed dimensions, leading to geometric distortions.
- **Accuracy Issues:** Recognition accuracy can be compromised due to:
  - **Content Loss:** Cropping may remove essential parts of the object.
  - **Distortion:** Warping alters the object's proportions, leading to misleading features.
- **Scale Variability:** A pre-defined input size may not be optimal when object sizes vary significantly across images.

In the next sub-section let's deep divide into why do we need to resize the inputs.

1.2 Why do we resize the inputs

A Convolutional Neural Network (CNN) mainly consists of two parts:

- **Convolutional Layers:**
  - Operate in a sliding-window manner.
  - Output feature maps that represent the spatial arrangement of activations.
  - Do not require a fixed image size and can generate feature maps of varying dimensions.
- **Fully Connected Layers:**
  - Require a fixed-size input due to their definition.
  - Are responsible for the fixed-size constraint in CNNs.
  - Exist at a deeper stage of the network.

This motivated the researcher of this paper (Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun) to propose a solution which gives a fixed size feature map to Dense layers.

1.3 Spatial Pyramid Pooling Layer

The Spatial Pyramid Pooling (SPP) layer is introduced to address the issue of fixed-size input constraints in CNNs. It enables the network to process images of arbitrary sizes without requiring resizing or cropping.

- **Multi-level Pooling Mechanism:** Uses multiple pooling windows at different scales (e.g., 1×1, 2×2, 4×4).
- Each level captures spatial information at different resolutions.
- The pooled features are concatenated into a fixed-length vector.

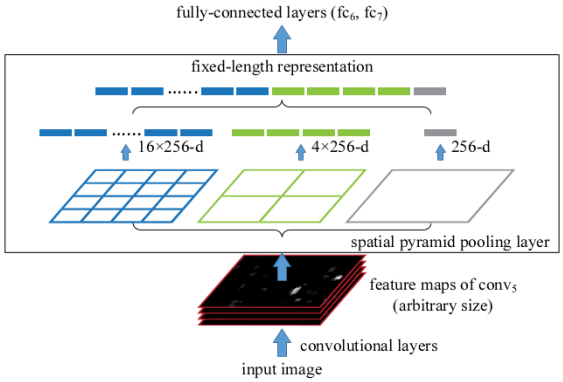

Figure 1: A network structure with a spatial pyramid pooling layer.

2 Architecture of SPPNet

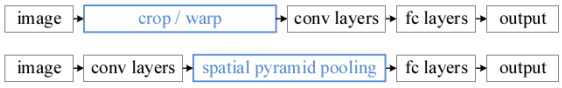

Figure 2: Top: A conventional CNN. Bottom: Spatial Pyramid Pooling network structure.

SPPNet utilizes a CNN backbone, such as AlexNet or VGG, to process images of varying sizes. The SPP layer receives feature maps from the convolutional layers and converts them into a fixed-size representation. This fixed-length feature map is then passed to the dense layers, which generate the final output.

3 Improvements Over Previous Work

SPPNet introduces several improvements compared to traditional CNN architectures such as AlexNet and R-CNN:

- **Processes Images of Any Size:** This preserves spatial information without the need for resizing.
- **Enhanced Feature Representation:** Multi-scale pooling allows SPPNet to capture more spatial context. Leads to more robust and informative feature extraction.
- **Improved Object Detection Performance:** The spatial pyramid pooling approach enhances classification accuracy. Achieves superior performance on benchmark datasets like PASCAL VOC.