Spatial Pyramid Pooling for Vehicle Detection

Raktim Dey (MDS202132),
Rishika Tibrewal (MDS202135),
Shreyansh Rastogi (MDS202144)

Project Report

September 5, 2024

Abstract

The research paper introduces a new technique known as Spatial Pyramid Pooling (SPP) for improving visual recognition in deep convolutional neural networks (CNNs). The SPP method allows for flexible input image sizes while maintaining fixed output dimensions, which is useful for recognizing objects at different scales. The paper shows that SPP-based CNNs achieve better performance than traditional CNNs on several visual recognition tasks, including image classification, object detection, and scene parsing. Additionally, the paper provides insights into the importance of the different levels of the spatial pyramid in the SPP method and suggests several possible applications of the technique in computer vision.

Contents

Work Contribution

Member	Contribution
Raktim	• Paper Research • Understanding the Paper • Code Implementation
Rishika	• Understanding the Paper • Presentation • Report
Shreyansh	• Understanding the Paper • Code Implementation • Report

1 Introduction

In the past, Convolutional Neural Networks (CNNs) were commonly applied to tasks such as image classification and object detection, where they operate by sliding a window over the input image and producing feature maps. Their drawback is that they require an input image with a fixed input size, which restricts the aspect ratio and scale of the input image. Current techniques for processing images of arbitrary sizes typically adjust the input image to a fixed size by means of cropping or warping, but this can occasionally result in information loss and produce geometric distortions. The accuracy of recognition may suffer as a result of the loss or distortion of content. Convolutional layers, in reality, do not necessitate a fixed image size and can produce feature maps of any size. However, fully-connected layers must have input of a fixed size/length by definition. Consequently, the fixed size restriction arises solely from the fully-connected layers, which are located at a deeper level of the network.

The paper we are discussing proposed the introduction of a Spatial Pyramid Pooling (SPP) layer to address the aforementioned limitation of CNNs. This layer is placed before the first fully connected layer, resulting in a network known as an SPP-net. The SPP layer aggregates the features and creates outputs of a fixed length, which are subsequently fed into the fully connected layers. By allowing for the use of images with varying scales or sizes during training, this new network helps to mitigate overfitting.

We will show that a network trained with a Spatial Pyramid Pooling layer achieves higher accuracy on both seen and unseen images as compared to a network trained without a SPP layer.

2 Main Results

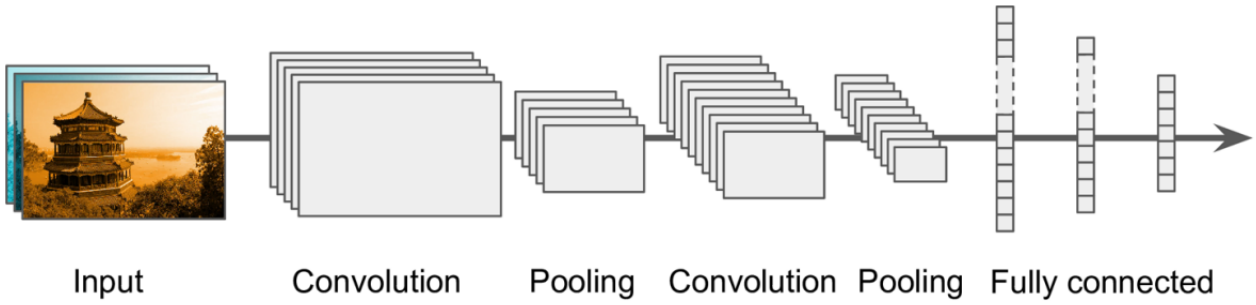
2.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of neural network that have major applications in object recognition and detection in images. They work by applying multiple iterations of convolution to extract features from the input image, with each iteration producing a feature map that captures a different aspect of the image. These feature maps are then passed on to the next layer, where they are combined and processed to generate higher-level features. Subsequent layers may include pooling layers, which reduce the size of the feature maps to make them more computationally efficient to process.

Once the final pooling layer has been reached, a fully connected network is used to classify the image. The CNN is trained using large labeled images for object detection, with the weights of the neurons in the network learned to minimize the difference between the predicted output and the true label.

CNNs automatically learn to recognize features in images, rather than relying on manual feature extraction. This makes them a powerful tool for a wide range of applications, from

image classification to object detection and segmentation.



This is a classical CNN architecture where an image is taken as an input, convolutional filters are applied to extract features in the form of feature maps, which are then pooled and the same thing repeats until the final feature map is sent to fully connected layers for the output.

The goal of this project is to identify vehicles, but it may be challenging because the input images can vary in both size and aspect ratio. Classic convolutional neural networks (CNNs) may not be the ideal solution due to their limitations in this regard.

CNNs use convolutional layers to extract features from input images by sliding a window over the image and generating feature maps. However, fully connected layers in the network require fixed input image size, which restricts the aspect ratio and scale of input images. To handle varying input image sizes, existing methods warp or crop input images to a fixed size, which can result in geometric distortions or content loss, impacting accuracy. Additionally, using a fixed input size may not be appropriate when object scales vary. Therefore, methods such as Spatial Pyramid Pooling (SPP) have been developed to allow CNNs to handle variable-sized inputs without distorting the content or losing accuracy.

2.2 Bag-of-Words Model

In the bag-of-visual-words model for image classification, images are treated as documents and their features as words. Features are extracted from images, and a codebook is created by clustering these features from many images. Each image is then represented as a "bag" of visual words, with the frequency of each visual word in the image indicating the importance of that word in the image. Finally, a classifier is trained to distinguish between object classes based on these bag-of-visual-words representations. This approach allows for efficient and effective image classification by reducing the complexity of the image to a simpler set of visual words.

However, this model also has some limitations. Firstly, it is sensitive to object scale and orientation, which means that the model may not perform well when the object appears at different scales or orientations. Secondly, the model uses a fixed vocabulary, which limits its ability to capture a wide range of variations in object appearance. Thirdly, the model creates a high-dimensional vector to represent an image, making it computationally expensive

to process. Finally, the Bag of Words model does not preserve the spatial relationships between words, which can lead to loss of information about the object's location and shape. These impact the accuracy of vehicle detection, and hence alternative approaches such as Spatial Pyramid Pooling have been developed to address these limitations.

2.3 Spatial Pyramid Pooling

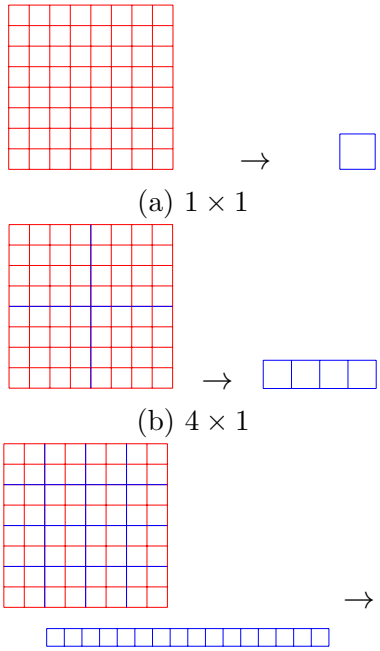
Spatial Pyramid Pooling (SPP) is a computer vision technique that removes the fixed-size constraint of a network. It was introduced by He et al. in 2014 in their paper "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". It can handle images of varying sizes and aspect ratios and still extract fixed-length feature vectors.
In contrast to the sliding window pooling technique using only a single window size, SPP uses multi-level spatial bins. It pools the features extracted from each sub-region into a fixed-length vector. This allows the network to capture information about objects at different scales and positions within the image.

2.3.1 Spatial Pyramid Pooling Layer

The SPP-net (CNN network with SPP layer) is able to process images with different sizes and aspect ratios, resulting in outputs of varying sizes. To create fixed-length vectors required for fully connected layers, Spatial Pyramid Pooling (SPP) layer is used. SPP preserves spatial information by dividing the image into local spatial bins and pooling the information within each bin. The size of the spatial bins is proportional to the size of the image, but the number of bins is fixed regardless of image size. The last pooling layer following the last convolutional layer is replaced with an SPP layer. The output of the SPP layer is a vector with kM dimensions, where k represents the number of filters in the last convolutional layer, and M is the number of bins.

2.3.2 Working of SPP Layer

To feed into an SPP layer in a CNN, the input is a feature map, which is the output of the previous convolutional layer. The feature map is then divided into a set of rectangular sub-regions at different scales. Within each sub-region, max-pooling is applied to the features to produce a fixed-length feature vector. The resulting feature vectors from all sub-regions are then concatenated together to form the final output. This concatenated feature vector is then passed onto a fully connected layer for further processing.



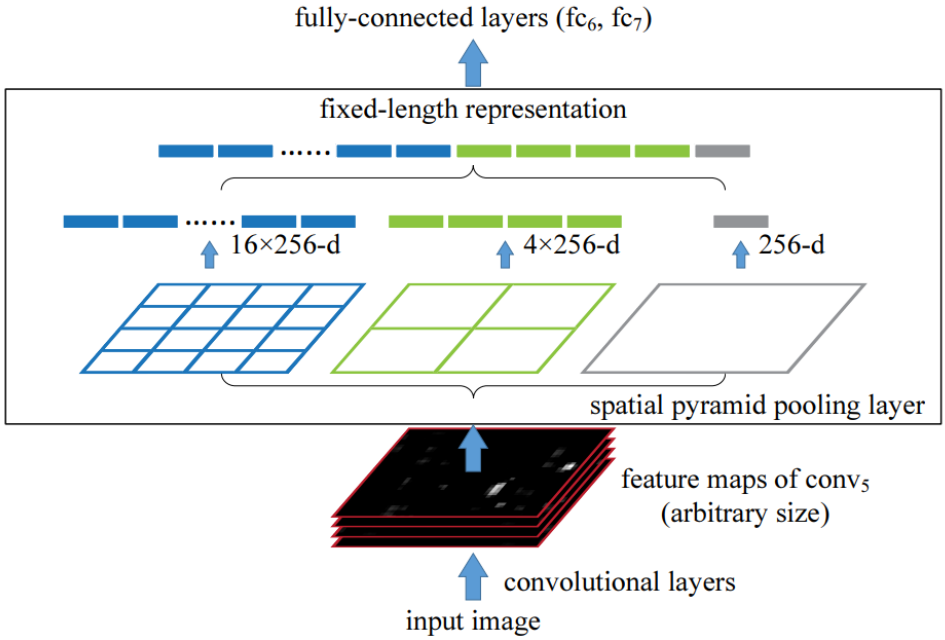
(a) 1×1

(b) 4×1

(c) 16×1

Feature map processing

Hence, the final output would be a concatenated vector of dimension 21×1 Below is an example of the SPP-net architecture.



2.3.3 Single-Size Training the SPP-Net

It is possible to pre-compute the bin sizes required for spatial pyramid pooling for an image of a specific size. For a feature map of size $a \times a$ after the last convolutional layer and a pyramid level of $n \times n$ bins, window size,$win = \lceil a/n \rceil$ and stride, $str = \lfloor a/n \rfloor$. l such layers are implemented for an $l-$level pyramid, the outputs of which are concatenated by the next fully connected layer.

Below is an example:

[pool 3×3]	[pool 2×2]	[pool 1×1]
type=pool	type=pool	type=pool
pool=max	pool=max	pool=max
inputs=conv5	inputs=conv5	inputs=conv5
sizeX=5	sizeX=7	sizeX=13
stride=4	stride=6	stride=13

2.3.4 Multi-Size Training the SPP-Net

The main purpose of multi-size training is to replicate the effect of different input sizes, while still taking advantage of the optimized fixed-size implementations. Multiple sizes can be considered, say 180×180 and 224×224. Instead of classical cropping or warping the image, a 224×224 image is resized to 180×180 to create input images that differ only in resolution. Another neural network can be implemented that takes in fixed-size inputs of 180×180. The size of the feature map is denoted by $a \times a$, and the window size win is calculated as the ceiling of a/n, where n is a predetermined integer. The stride str is calculated as the floor of a/n.The output of the Spatial Pyramid Pooling layer in the 180×180 network has the same fixed length as the output of the 224×224 network, meaning that both networks share the same parameters. One epoch can be trained on one network and then transferred to the other, while keeping all of the weights unchanged.

2.4 Details of implementation

The dataset obtained from the Stanford website contained images and the co-ordinates of the opposite diagonal of the bounding box. Selective search algorithm was used to obtain the Intersection over Union for the bounding boxes obtained and calculated using the said algorithm. The images for which the IoU value was greater than 0.5 was labelled as "car" and "not car" otherwise.

The architecture of the CNN without SPP layer consisted of 3 convolutional layers of 512, 512 and 256 neurons respectively, with a fixed kernel size of 5×5. A dense layer of 128 neurons followed by a 2-neuron dense layer was attached at the end with a softmax activation. Adam with a learning rate of 0.0001 was used as the optimization algorithm. The model had 88 million trainable parameters. The network was trained for 50 epochs with 10 steps per epoch.

The architecture of the CNN with SPP layer was the same as the previous one, with an added Spatial Pyramid Pooling layer of dimensions (1,2,4) attached before the first dense layer.

3 Results and Observations

Architecture	Training Accuracy	Validation Accuracy
Classical CNN	87%	85%
CNN with SPP	90%	92%

3.1 Data Collection

The dataset used for training the networks were open source datasets, Stanford Cars Dataset, that was downloaded from Kaggle that can be found here:
`https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset`

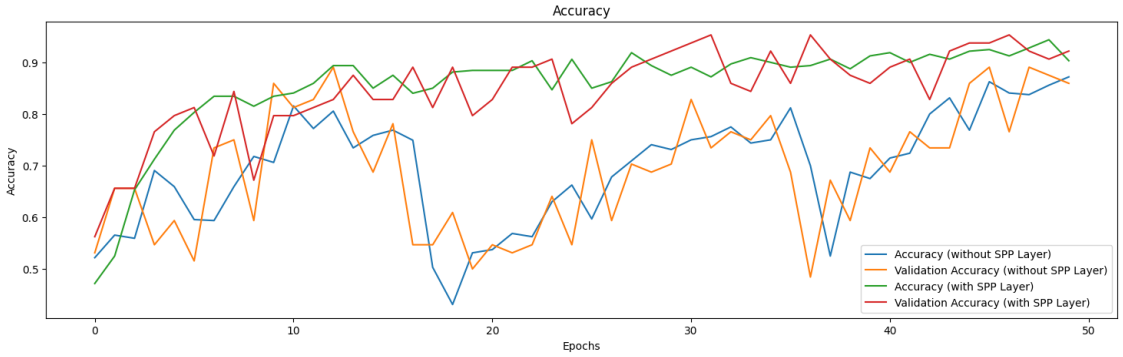
3.2 Data Visualization and Examples

(a) Without Bounding boxes (b) With Bounding boxes

Input Image

(a) Without Bounding boxes (b) With Bounding boxes

Input Image



(a) (b) (c)

Model Predictions

Link to the google colab notebook

4 References

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition".
- Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. "Beyond Bag off Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories"
- Tao Qu, Quanyuan Zhang, Shilei Sun. "Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based Deep Convolutional Neural Networks."