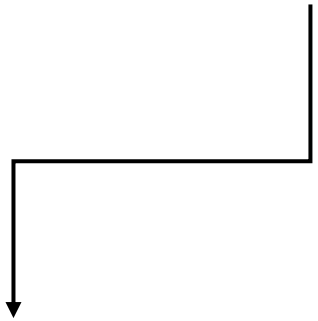
SPP-Net

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition



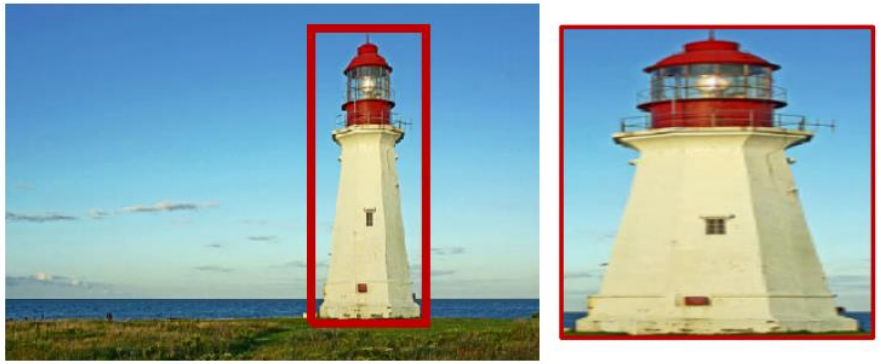
Spatial : 공간적이라는 의미로 이미지 내의 위치와 관련된 정보를 의미
Pyramid : 여러 레벨에서 이미지를 나누는 피라미드 형태의 구조를 의미
Pooling : CNN에서 특정 영역 내에서 특징을 요약하는 작업

Abstract

SPP-Net은 Spatial Pyramid Pooling을 도입한 Architecture 입니다.
이를 통해 고정된 이미지 사이즈만 학습하는 것이 아닌 여러 이미지 사이즈를 입력으로 받을 수 있습니다.
Pascal VOC 2007에서 R-CNN보다 24~102배 빠른 속도로 처리하면서도 더 나은 또는 유사한 정확도를 달성했습니다.
ImageNet 대규모 시각 인식 챌린지 (ILSVRC) 2014에서 Object Detection 부분에서 2위,
Classification 부분에서 3위를 차지했습니다.

Identified Issue - Fixed Input Image

crop



warp

- **Crop : 전체 객체를 포함하지 않을 수 있습니다.**
- **Warp : 원하지 않는 기하학적 왜곡을 초래할 수 있습니다.**

CNN의 학습 및 테스트 과정에서는 일반적으로 고정된 입력 이미지 크기를 요구합니다.
이는 입력 이미지의 가로세로 비율과 크기를 crop, warp 방법을 통해 제한합니다.

- 객체 크기가 다양한 경우 고정하게 되면 크기 관련된 문제를 간과하는 것입니다.

Why do CNN require a fixed input size ?

crop

warp

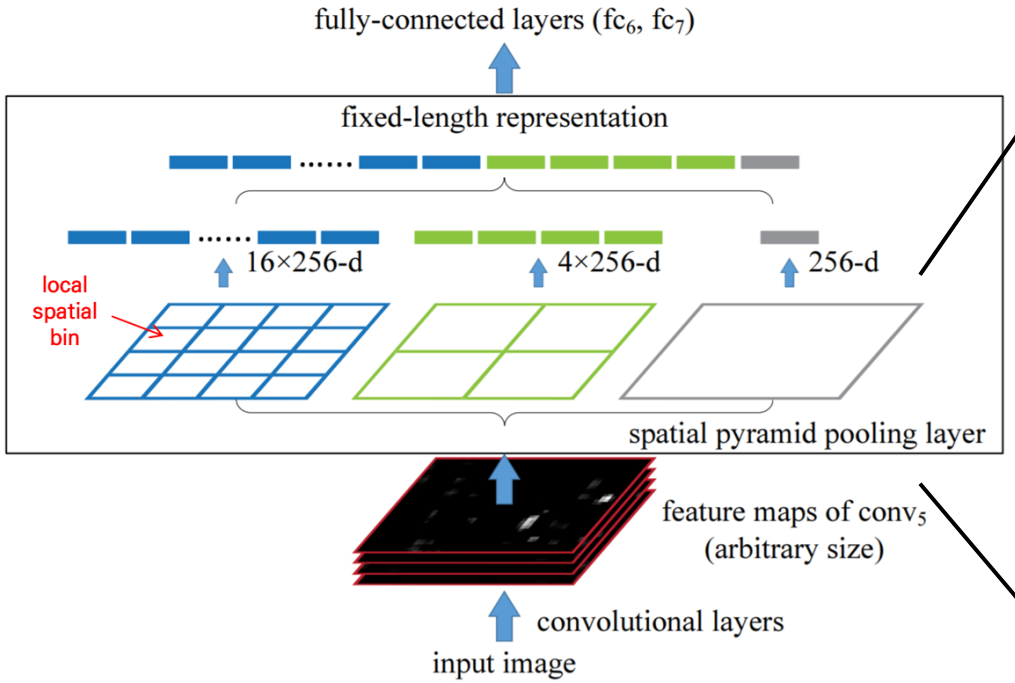
- Conv layers : 공간적 배열을 나타내는 Feature map을 출력 (fixed input size 요구 X)
- **FC layers** : fixed input size를 필요로 합니다.

이 논문에서는 네트워크의 고정 크기 제약을 제거하기 위해 Spatial Pyramid Pooling(이하 SPP)을 제시합니다.

1. 합성곱 층 마지막 부분에 SPP 레이어를 추가하여 특징을 풀링합니다.
2. SPP를 통해 고정된 길이의 출력값을 생성합니다.
3. 이를 FC layer 또는 다른 분류기로 전달합니다.

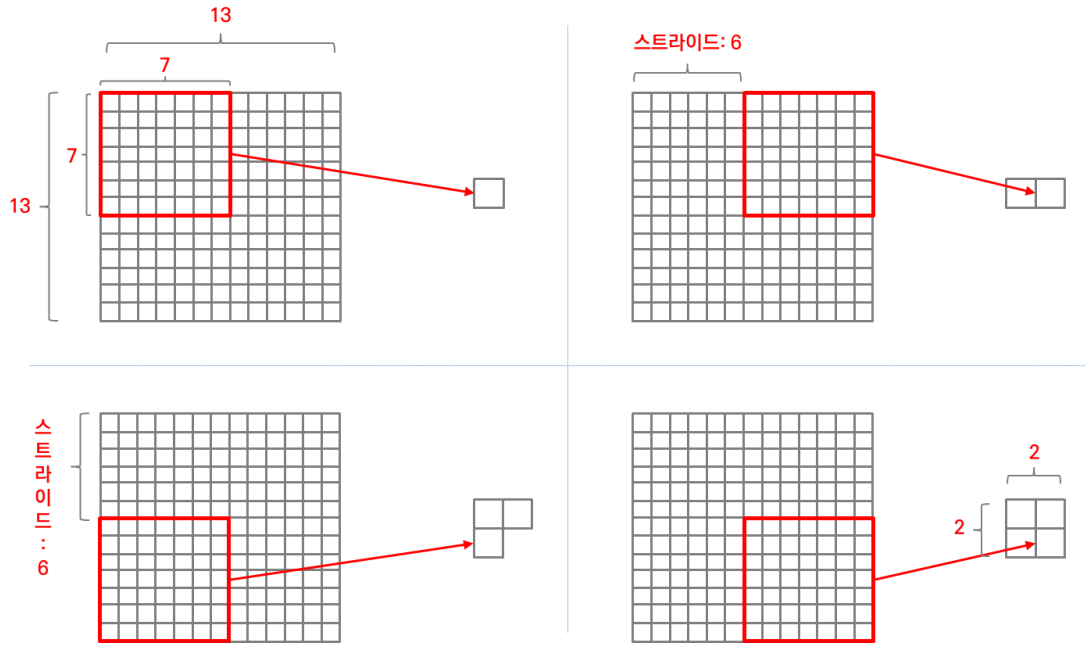
→ **네트워크 계층 구조에서 더 깊은 단계 (conv와 fcl 사이)에서 정보를 aggregation하여 초기 단계에서 crop, warp이 필요없게 만듭니다.**

Spatial Pyramid Pooling



기존 max, average pooling은 single-level pooling이고 여러 level에서 pooling을 진행해서 합치기 때문에 **multi-level pooling**이라고 부릅니다.

- **Single-size Training** : 단일 크기의 입력 이미지로 학습
- **Multi-size Training** : (180x180), (224x224) 등의 여러 크기의 이미지로 학습



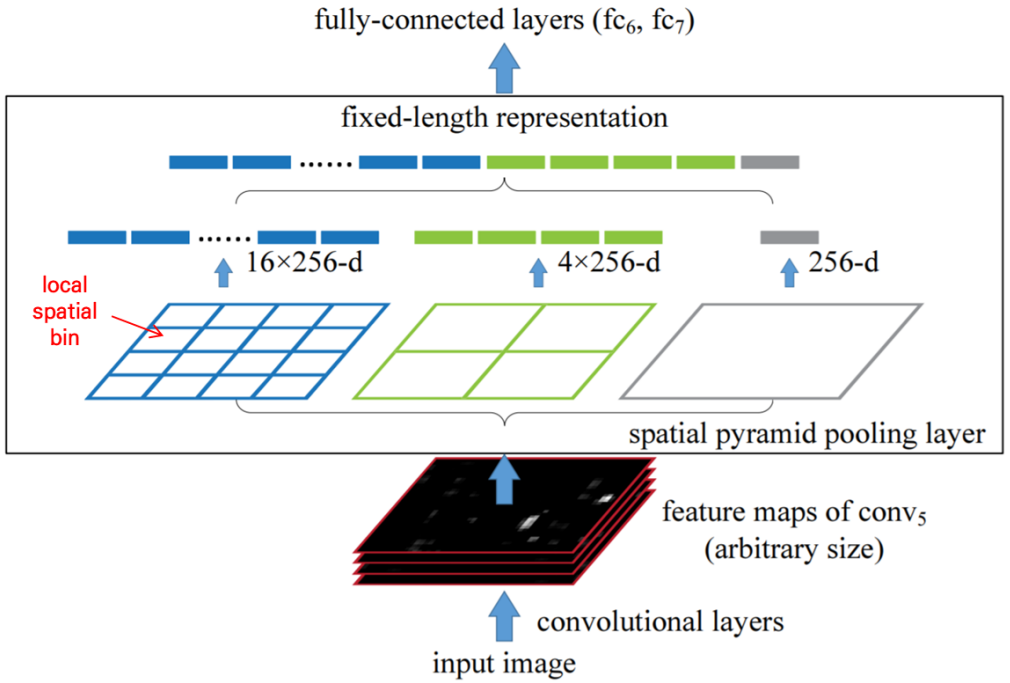
CONV5를 거친 Feature map 크기가 (13x13)일 때, spatial bin 크기를 (2x2)로 만들고 싶다고 했을 때의 연산을 예시로 설명해보겠습니다.

Conv5를 거친 feature map 크기가 (a x a)이고, 만들고자 하는 spatial bin 크기가 (n x n)일 때, window size와 stride는 아래와 같이 정의될 수 있습니다.

- Window size : ceiling(feature map size(13) / pooling size(2)) = 7
- Stride : floor(feature map size(13) / pooling size(2)) = 6

25 bin = [4x4, 2x2, 1x1]
50 bin = [6x6, 3x3, 2x2,1x1] : conv5의 feature map에 pooling을 통해 생성되는 출력 크기

SPP-Net Evaluation



기존 max, average pooling은 single-level pooling이고 여러 level에서 pooling을 진행해서 합치기 때문에 **multi-level pooling**이라고 부릅니다.

- **Single-size Training** : 단일 크기의 입력 이미지로 학습
- **Multi-size Training** : (180x180), (224x224) 등의 여러 크기의 이미지로 학습

SPP on	test view	top-1 val
ZF-5, single-size trained	1 crop	38.01
ZF-5, single-size trained	1 full	**37.55**
ZF-5, multi-size trained	1 crop	37.57
ZF-5, multi-size trained	1 full	**37.07**
Overfeat-7, single-size trained	1 crop	33.18
Overfeat-7, single-size trained	1 full	**32.72**
Overfeat-7, multi-size trained	1 crop	32.57
Overfeat-7, multi-size trained	1 full	**31.25**

Table 3: Error rates in the validation set of ImageNet 2012 using a single view. The images are resized so $\min(w, h) = 256$. The crop view is the central 224×224 of the image.

-> **Crop 하는 것 보다 SPP-Net을 통해 full image를 학습하는게 성능이 더 좋다.**

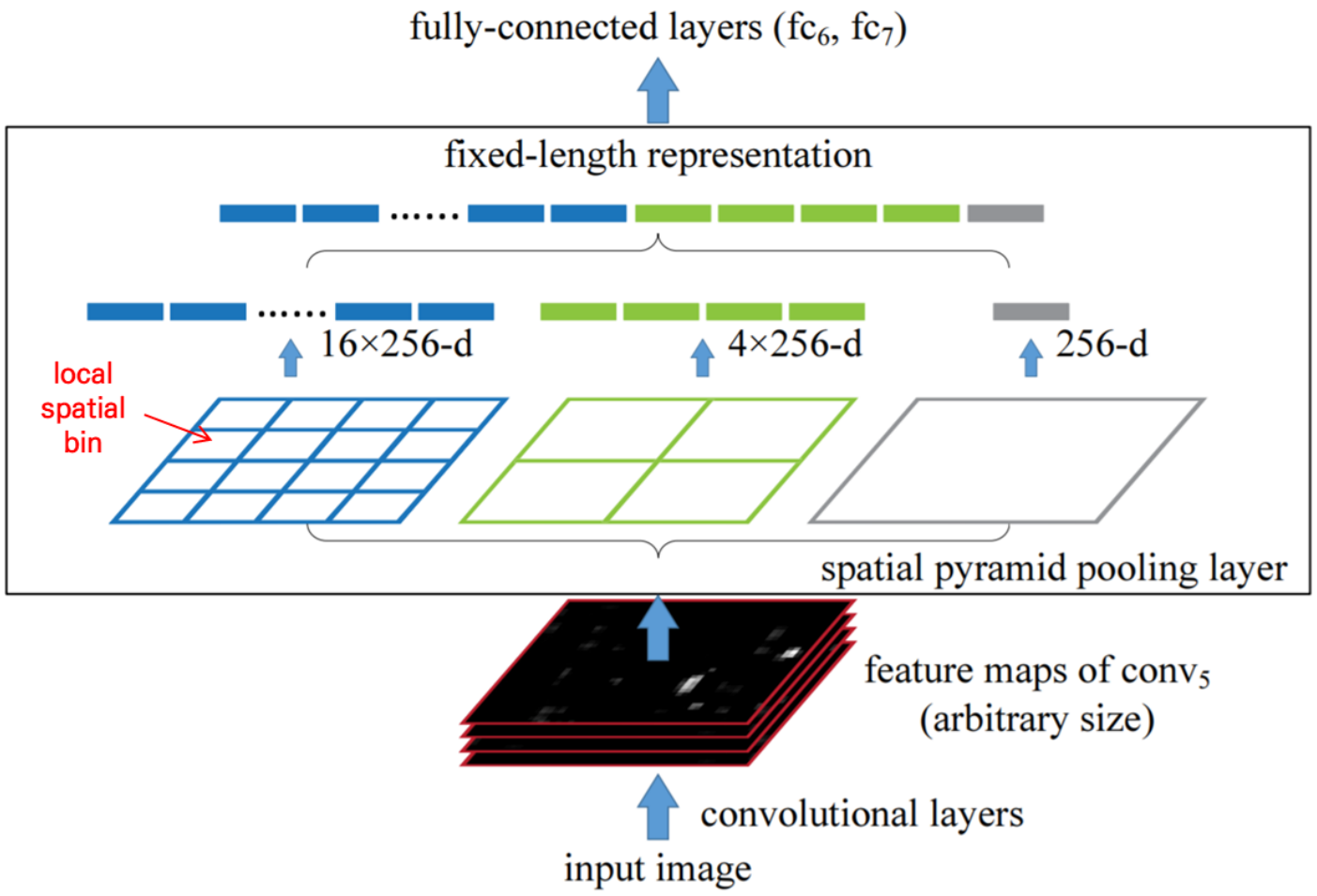
		top-1 error (%)			
		ZF-5	Convnet*-5	Overfeat-5	Overfeat-7
(a)	no SPP	35.99	34.93	34.13	32.01
(b)	SPP single-size trained	34.98 (1.01)	34.38 (0.55)	32.87 (1.26)	30.36 (1.65)
(c)	SPP multi-size trained	34.60 (1.39)	33.94 (0.99)	32.26 (1.87)	29.68 (2.33)

		top-5 error (%)			
		ZF-5	Convnet*-5	Overfeat-5	Overfeat-7
(a)	no SPP	14.76	13.92	13.52	11.97
(b)	SPP single-size trained	14.14 (0.62)	13.54 (0.38)	12.80 (0.72)	11.12 (0.85)
(c)	SPP multi-size trained	13.64 (1.12)	13.33 (0.59)	12.33 (1.19)	10.95 (1.02)

Table 2: Error rates in the validation set of ImageNet 2012. All the results are obtained using standard 10-view testing. In the brackets are the gains over the "no SPP" baselines.

Standard 10-view란 10개의 cropping 이미지를 multi-size training에 사용한다는 뜻.

SPP-Net Architecture



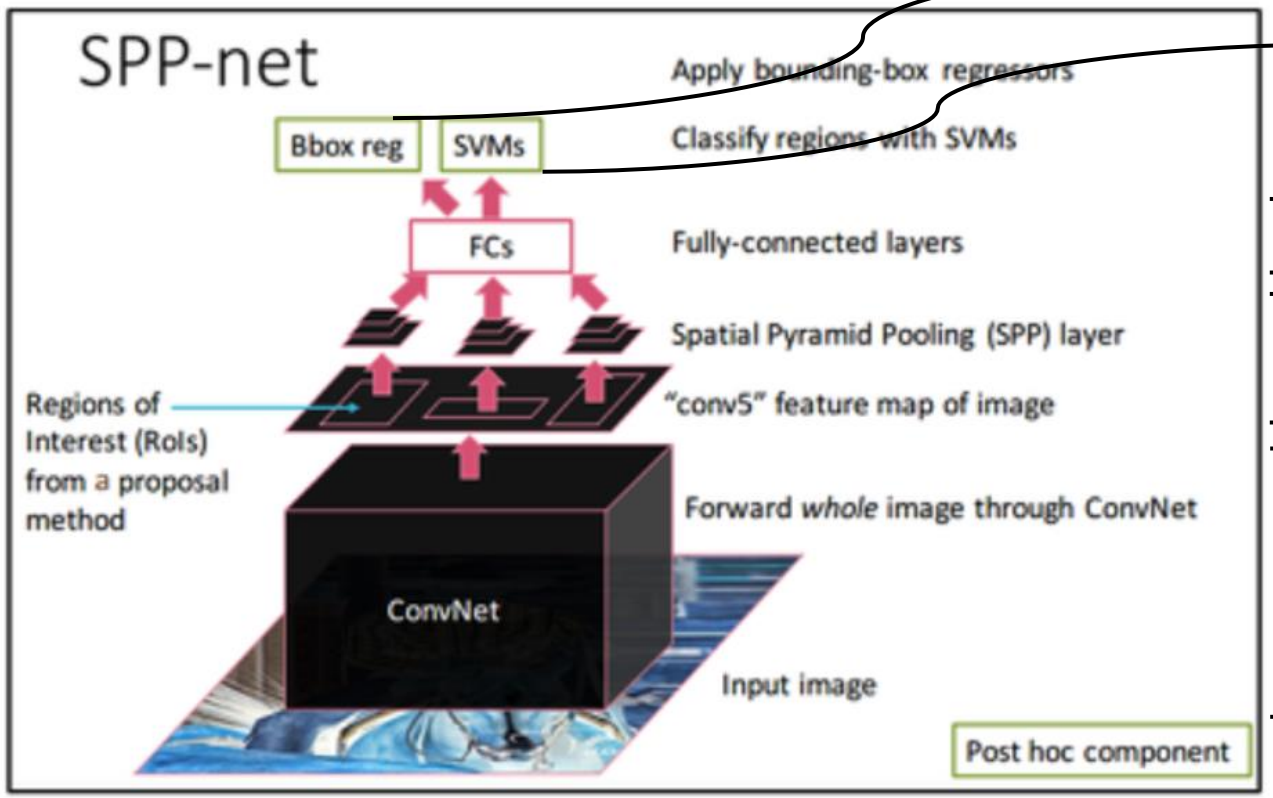
3. fully-connected layers의 fixed input size에 맞게 입력

2. 마지막 conv 레이어인 $conv_5$에서 얻어진 feature map을 SPP를 통해 Pooling 진행

- 입력 크기와 관계없이 고정된 크기의 출력을 생성할 수 있습니다.
- Multi-level spatial bins를 사용하여 다양한 위치와 크기에 대해 여러 해상도에서 정보를 추출하기에 객체 변형에 강인한 특성을 갖습니다.
- 다양한 크기의 이미지에 대해 특징을 pooling 할 수 있어 다양한 입력 크기에 대해 유연성을 보입니다.
- 다양한 크기의 이미지를 사용하여 학습하면 크기에 대한 불변성을 증가시키고 과적합을 줄일 수 있습니다.

1. R-CNN에서와 같이 2000장의 region proposals 전부 Conv를 거치는 것이 아니라 Input Image 한 장만 Conv를 거친다.

Object Detection SPP-Net



6. 점수를 매긴 window에 대해 Non-maximum suppression(30% 임계값)을 적용

5. Bounding Box Regression을 사용하여 predict window를 후처리 합니다.

4. 각 category에 대한 feature vector를 Linear SVM에 학습합니다.

- Negative sample : IoU <= 0.3
- Negative sample들끼리 IoU >= 0.7만큼 겹치는 Negative sample은 제거
- Hard negative mining을 사용하여 학습

3. Pooling을 통해 얻은 feature vector를 fully-connected layers에 넣습니다.

2. 4-level spatial pyramid (1x1, 2x2, 3x3, 6x6, 총 50 bins)를 사용해 특징을 Pooling 합니다.

1. Input Image를 ZF-5 (Single-size trained) ConvNet에 입력으로 줍니다.