

2-StageDetector(2):SPPNet

Spatial Pyramid Pooling in Deep Convolutional
Networks for Visual Recognition

Intro: Meaning

Spatial Pyramid Pooling
기법 소개

Img Classification에
서의 성능

Object Detection
에서의 성능

Conclusion

Meaning

01

02

03

04

ILSVRC-2014

© ImageNet Large-Scale Visual
Recognition Challenge

rank	team	top-5 test
1	GoogLeNet [32]	6.66
2	VGG [33]	7.32
3	<u>ours</u>	<u>8.06</u>
4	Howard	8.11
5	DeeperVision	9.50
6	NUS-BST	9.79
7	TTIC_ECP	10.22

Table 5: The competition results of ILSVRC 2014 classification [26]. The best entry of each team is listed.

rank	team	mAP
1	NUS	37.21
2	<u>ours</u>	<u>35.11</u>
3	UVA	32.02
-	(our single-model)	(31.84)
4	Southeast-CASIA	30.47
5	1-HKUST	28.86
6	CASIA_CRIPAC_2	28.61

Table 13: The competition results of ILSVRC 2014 detection (provided-data-only track) [26]. The best entry of each team is listed.

- +) architectures despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieves state-of-the-art classification results using a single full-image representation and no fine-tuning.

➤➤ SPPNet 사용한 모델이 Img Classification 부문과 Object Detection 부문에서 각각 3위, 2위에 랭크됨

01

» SPPNet 의의

02

03

04

- 고정된 인풋 크기가 아닌 임의의 크기 이미지에 대해 CNN으로 학습을 하고 conv5 layer까지 동일
- R-CNN과 다르게 전체 이미지에 대해 한 번만 수행하므로 수 백배 빠름
- 다양한 크기의 bin을 통해 고정된 길이의 결과값을 만듦과 동시에. 다양한 scale의 특징을 추출

popular idea [23], [24], [20], [5]. But SPP-net inherits the power of the deep CNN feature maps and also the flexibility of SPP on arbitrary window sizes, which

Classification과 detection 모두에서 강점을 보임

What is Spatial Pyramid Pooling(SPP)?

01

02

03

04

CV에서 CNN을 활용하기 시작했
을 당시의 문제점

When applied to images of arbitrary sizes, current methods mostly fit the input image to the fixed size,
may not be suitable when object scales vary. Fixing input sizes overlooks the issues involving scales.

➤➤ CNN을 이용해 모델을 학습시키기 위해 입력되는
이미지의 사이즈가 일정해야만 했다.

01

02

03

04

SPP Key Insight

- CNN 네트워크는 Conv와 Fc로 나뉜다.
- Convolutional layers에서는 input shape을 맞추어줄 필요가 없다.
- 왜? 아무 형태의 입력이 들어오더라도 windows를 통해 임의의 feature map만 뽑아주는 것이 conv의 역할이므로

size/length input by their definition. Hence, the fixed-size constraint comes only from the fully-connected layers, which exist at a deeper stage of the network.

SPP Key Insights

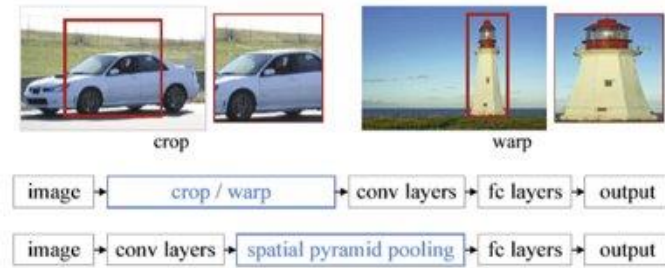


Figure 1: Top: cropping or warping to fit a fixed size. Middle: a conventional CNN. Bottom: our spatial pyramid pooling network structure.

- 컨볼루션 통과할 때까지는 input shape에 손대지 않고 마지막 conv layer 바로 다음에 SPP layer 추가
- SPP layer의 역할: 특성들 pooling함으로써 입력값에 관계없이 고정된 길이의 출력값 생성
- 즉, 모델의 처음 부분에서 입력값에 변형을 주는 것을 피하고자 conv와 fc 사이라는 네트워크의 더 깊은 차원에서 정보를 합치는 것

variable size images increases scale-invariance and reduces over-fitting. We develop a simple multi-size

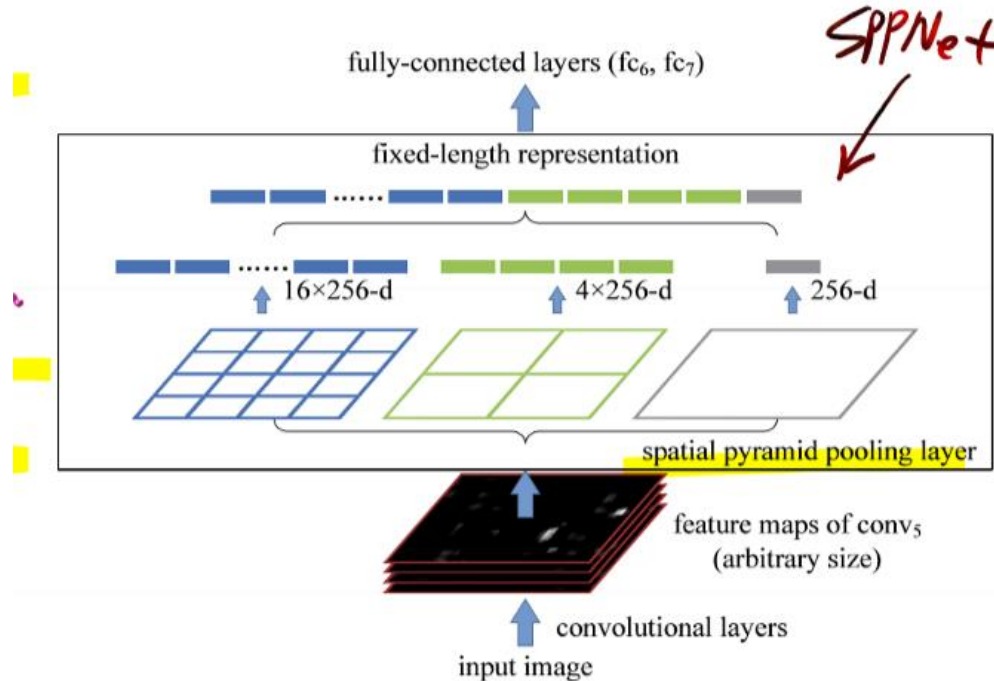
why?

→ better accuracy

Deep Network with Spatial Pyramid Pooling

SPPNet 개요

Deep Network with SPP



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

→ 2 × 2 Max-Pool →

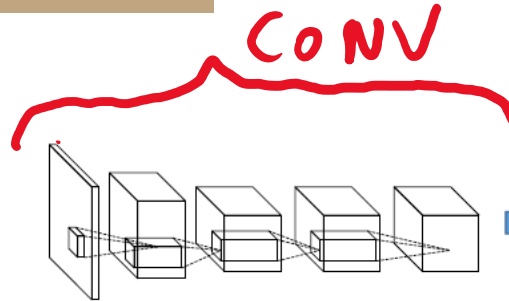
20	30
112	37

Figure 3: A network structure with a **spatial pyramid pooling layer**. Here 256 is the filter number of the conv₅ layer, and conv₅ is the last convolutional layer.

- SSP는 Multi-level spatial bins를 사용
- SPPNet의 큰 흐름: 각 Scale값을 가지고 convolution 진행한 후에, 만들어진 여러 feature maps에 대해 bins를 이용하여 pooling 수행

SPP Key Insights

SPP-net



spatial pyramid pooling

any size

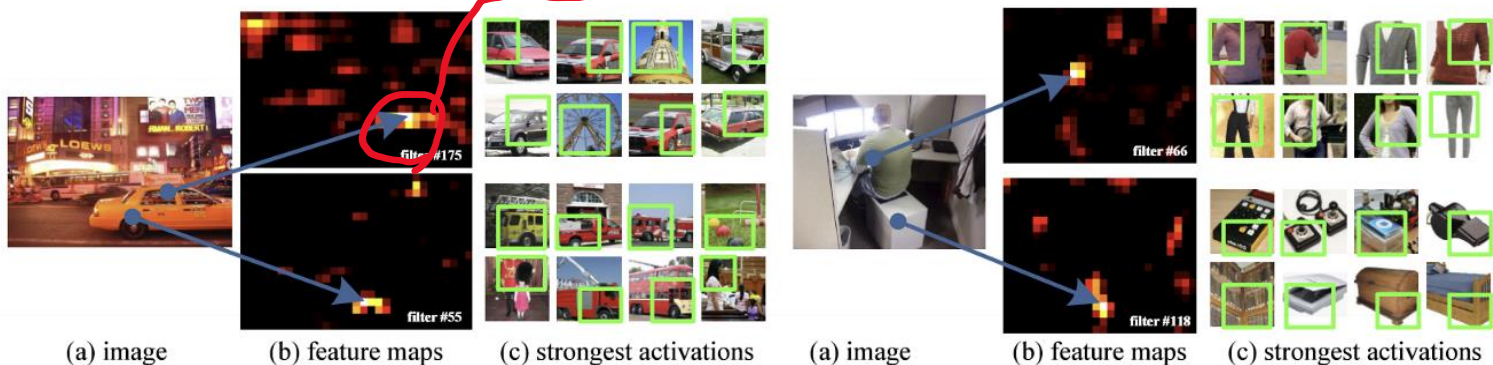
so the number of bins is fixed regardless of the image size. This is in contrast to the sliding window pooling of the previous deep networks [3], where the number of sliding windows depends on the input size.

- Fix bin numbers
- **DO NOT** fix bin size

in several concurrent works. In [31], [32] a global average pooling is used to reduce the model size and also reduce overfitting; in [33], a global average

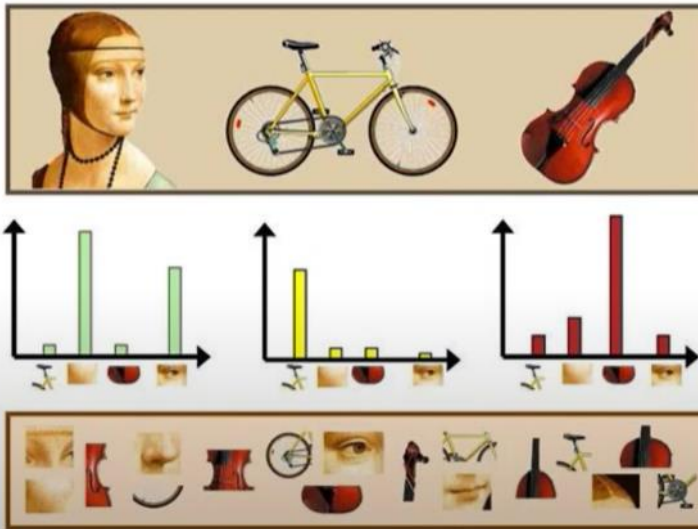
- 원도우의 개수는 이미 정해져 있다. 윈도우의 크기를 조절.

- Conv 결과로 얻어진 feature map은 특징의 강도를 나타낼 뿐 아니라 해당 특징의 공간적인 정보 또한 담고 있다.



2-2

+) 전체 CNN에서 Pooling 역할 간단 설명



Feature Extraction

- 이미지로부터 피처를 추출(SIFT 등)

Clustering

- 피처들을 클러스터링하여 코드값을 구함

Codebook

- 코드값들이 모인 코드북을 생성

Image Representation

- 이미지를 코드값들의 히스토그램으로 표현

Learning and Recognition

- svm 등의 분류기로 학습하여 이미지를 분류

= feature map
: pooling
영역
(이미지
feature를
요약)

01

02

03

04

Training 방법1: Single size

- Multi level pooling behavior하기 위한 초석
- 224*224로 input이미지 crop한 후 필요한 bins개수 계산
- 여러 크기 조합의 빈을 통해 다중 수준의 pooling을 가능하게 하여 정확도를 높이하고자 함

01

02

03

04

Training 방법1: Multi size

- 여러 사이즈의 사진을 가지고 훈련하는 것
- 여기서는 180*180 과 244*244의 두 가지 사이즈를 가지고 학습 진행
- 224를 180사이즈로 resize해서 내용과 배치를 동일하게 만든다.
- 이렇게 하여 학습한 두 네트워크는 결과적으로 같은 파라미터를 가지게 됨
- 임의의 크기의 사이즈를 가지고도 학습 진행해 보았으나 두 사이즈로 한 정확도가 더 높음.

Img Classification

01

02

03

04

Basic Network Structure

* Imagenet 2012 데이터로 학습

- ImageNet 2012 의 1000-category train set으로 학습
- 분류할 이미지를 256x256의 사이즈로 크기 조절
- 조절한 이미지에 대해 224x224의 사이즈로 crop(코너4개와센터 및 좌우대칭)하여 augmentation
- SPPNet을 거친 피쳐벡터에 softmax 스코어를 계산하여 분류에 사용
- 더 높은 정확성이 나온 이유가 단지 파라미터가 늘어서가 아니라 SPP를 썼기 때문임을 보여주려 함

conv - pooling - FC - softmax

Basic Network Structure

It is worth noticing that the gain of multi-level pooling is not simply due to more parameters; rather, it is because the multi-level pooling is robust to the variance in object deformations and spatial layout [15]. To show this, we train another ZF-5 network with a different 4-level pyramid: $\{4 \times 4, 3 \times 3, 2 \times 2, 1 \times 1\}$ (totally 30 bins). This network has fewer parameters than its no-SPP counterpart, because its fc_6 layer has 30×256 -d inputs instead of 36×256 -d. The top-1/top-5 errors of this network are 35.06/14.04. This result is similar to the 50-bin pyramid above (34.98/14.14), but considerably better than the no-SPP counterpart (35.99/14.76).

3x2x1x1 3x3x1 4x4x1x1 5x5x1x1 SPP가 좋다!

Basic Network Structure

		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)

3.1.3 Multi-size Training Improves Accuracy

Table 2 (c) shows our results using multi-size training. The training sizes are 224 and 180, while the testing size is still 224. We still use the standard 10-view prediction. The top-1/top-5 errors of all architectures further drop. The top-1 error of SPP-net (Overfeat-7) drops to 29.68%, which is 2.33% better than its no-SPP counterpart and 0.68% better than its single-size trained counterpart.

Besides using the two discrete sizes of 180 and 224, we have also evaluated using a random size uniformly sampled from [180, 224]. The top-1/5 errors of SPP-net (Overfeat-7) is 30.06%/10.96%. The top-1 error is slightly worse than the two-size version, possibly because the size of 224 (which is used for testing) is visited less. But the results are still better than the single-size version.

Basic Network Structure

		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-1 error (%)	
		Overfeat-5	Overfeat-7
(a)	no SPP	35.52	32.01
(b)	SPP single-size trained	30.06 (0.72)	29.68 (0.85)
(c)	SPP multi-size trained	29.68 (1.19)	29.68 (1.02)

3.1.3 Multi-size Training Improves Accuracy

Table 2 (c) shows our results using multi-size training. The training sizes are 224 and 180, while the testing size is still 224. We still use the standard 10-view prediction. The top-1/top-5 errors of all architectures further drop. The top-1 error of SPP-net (Overfeat-7) drops to 29.68%, which is 2.33% better than its no-SPP counterpart and 0.68% better than its single-size trained counterpart.

Besides using the two discrete sizes of 180 and 224, we have also evaluated using a random size uniformly sampled from [160, 224]. The top-1/5 errors of SPP-net (Overfeat-7) is 30.06%/10.96%. The top-1 error is slightly worse than the two-size version, possibly because the size of 224 (which is used for testing) is visited less. But the results are still better than the single-size version.

Object Detection

Multi-view Test

mer, driven by our adaptation framework, we find that multi-view testing on feature maps with flexibly located/sized windows (Sec. 3.1.5) can increase the classification accuracy. This manuscript also provides

>> ECCV 2014 초기 버전 발행 후 . ILSVRC 2014에 출전하여 좋은 성적을 보이면서 피쳐맵에 대한 multi-view test가 classification task에 좋은 성능을 가져옴을 알게 됨.


01

02

03

04

>> Layers Configuration

- Selective Search의 fast 모드를 사용해서 이미지당 2천여개의 후보 윈도우를 생성.
- 전체 이미지에 대해 피처맵을 뽑고, 4-level의 spp layer를 적용시킴 (총 50개의 bin * 256개의 필터)
- 2천여개의 윈도우마다 1280이라는 고정된 길이의 피처벡터를 추출해내서 svm classifier에 학습
- positive, negative sample를 생성해서 이진분류의 개념으로 학습
(negative sample은 positive 윈도우와 IoU값 기준으로 0.3 이하인 부분)
- 여러 스케일들에서 따로 피처맵들을 계산하고, 224에 가장 가까운 값을 가지는 스케일을 선택하여 그 피처맵을 이용 



Layers Configuration

01

02

03

04

	SPP (1-sc)	SPP (5-sc)	R-CNN
	(ZF-5)	(ZF-5)	(Alex-5)
pool ₅	43.0	<u>44.9</u>	44.2
fc ₆	42.5	44.8	<u>46.2</u>
ftfc ₆	52.3	<u>53.7</u>	53.1
ftfc ₇	54.5	<u>55.2</u>	54.2
ftfc ₇ bb	58.0	<u>59.2</u>	58.5
conv time (GPU)	0.053s	0.293s	8.96s
fc time (GPU)	0.089s	0.089s	0.07s
total time (GPU)	0.142s	0.382s	9.03s
speedup (vs. RCNN)	64×	24×	-

rank	team	mAP
1	NUS	37.21
2	<u>ours</u>	<u>35.11</u>
3	UvA	32.02
-	(our single-model)	(31.84)
4	Southeast-CASIA	30.47
5	1-HKUST	28.86
6	CASIA_CRIPAC_2	28.61

- Pascal VOC 2007에 대한 mAP 결과값으로, scale을 변화시키며 SPP를 적용한 것과, RCNN의 결과를 비교
- Ft(fine tuning)과 bb(bounding box regression)을 이용하면서 더 좋은 성능을 보임
- R-CNN보다 좋은 성능을 보이는 SPP layer 적용 실험도 보임

Reference

- <https://www.youtube.com/watch?v=i0lkmULXwe0>
- <https://www.youtube.com/watch?v=m3anASWelsc>

CrePAS 6th the first session

아주 상식적인 프레젠테이션 기

번

Thank you

크레파스 1기 | 물고기