Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Jian Sun(Microsoft Research),
Xiangyu Zhang(Xi'an Jiaotong University),
Shaoqing Ren(University of Science and Technology of China)

INDEX

Abstract

1. CNN에서는 완전 연결 계층 때문에 고정된 크기의 입력 사이즈가 필요.
2. 이는 입력 크기에 인식의 정확도가 크게 의존하는 결과를 가져옴.
3. 이미지를 입력 크기에 맞추다 보면 정보 손실(Cropping)이나 왜곡(Warpping)이 발생.
4. 저자들이 제안한 방법은 CNN이 임의의 사이즈로 훈련이 가능하도록 함.
5. 결과적으로 인식 성능이 더 좋아지게 됨.
6. 객체 탐지에서도 R-CNN과는 다르게 각 이미지에 대해서 CNN 연산을 한 번만 수행하게 되므로 속도를 비약적으로 감소 시킴.

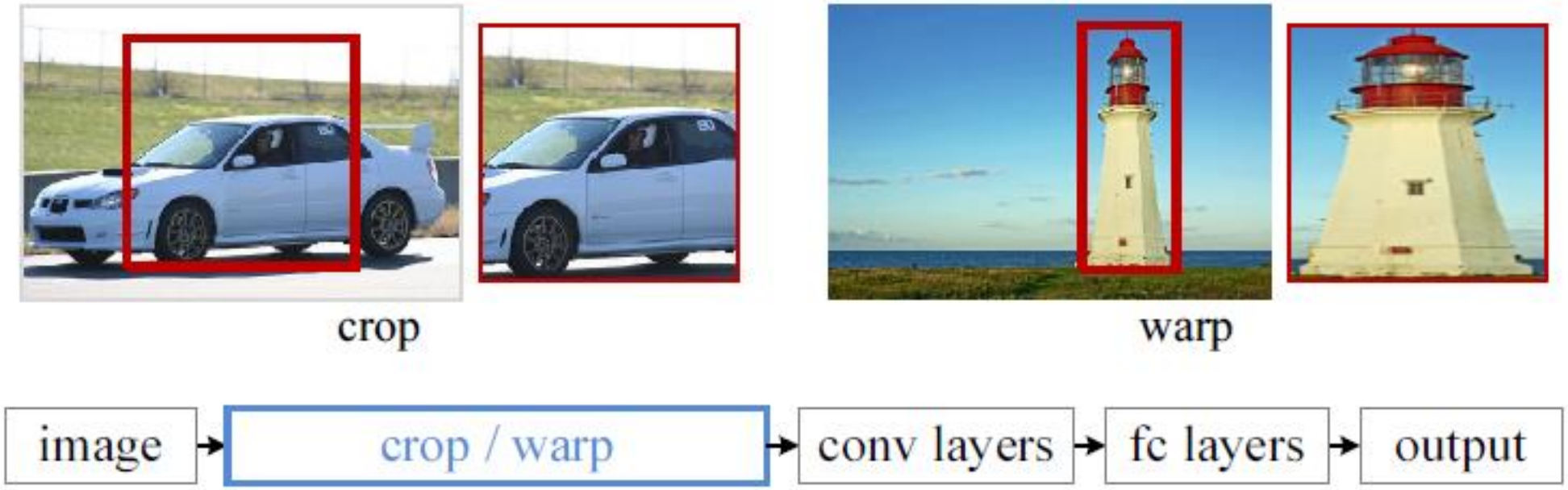
Introduction

완전 연결 계층이 고정된 입력 사이즈를 필요로 하는 이유?

Ex) 예를 들어서 Flatten된 4096차원 벡터를 8192차원 벡터로 사상시키는 선형 변환의 경우 다음의 과정을 거침.

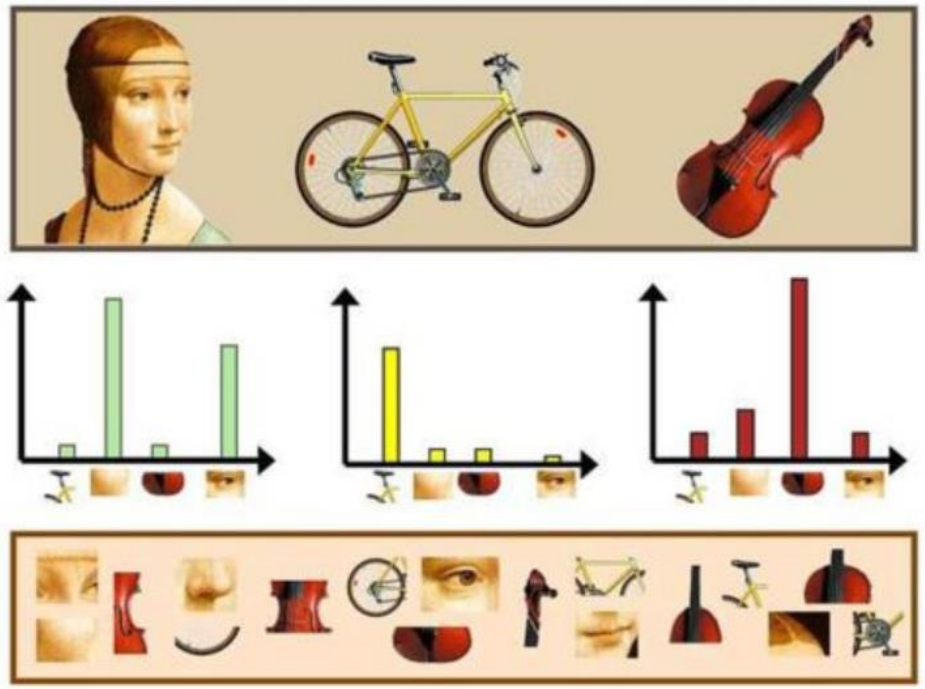
이때, 훈련 가능한 가중치들을 원소로 갖는 4096x8192 행렬은 모델 훈련 중에 동적으로 크기가 변할 수 없다.

Introduction



원본 이미지를 네트워크 입력 사이즈에 맞게 바꾸다 보면 정보 손실(Cropping)이나 왜곡(Warpping)이 발생할 수 있음.

Introduction



출처: Recognizing and Learning Object Categories(ICCV 2005 short course)

Feature Extraction
- 이미지로부터 피처를 추출(SIFT 등)

Clustering
- 피처들을 클러스터링하여 코드값을 구함

Codebook
- 코드값들이 모인 코드북을 생성

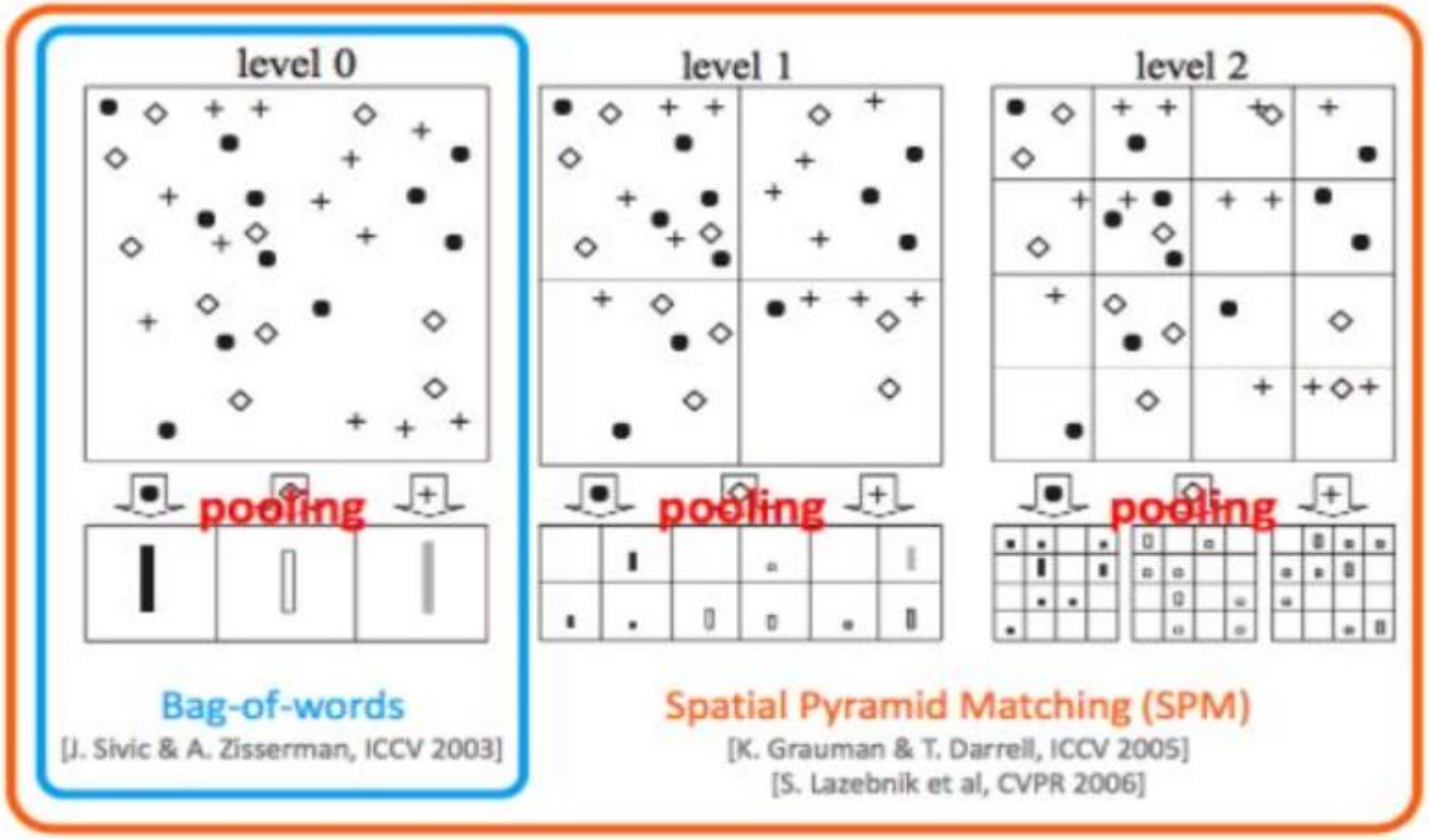
Image Representation
- 이미지를 코드값들의 히스토그램으로 표현

Learning and Recognition
- svm 등의 분류기로 학습하여 이미지를 분류

Cheon Wujin - Paper Review - Object Detection 3 (Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition)

Spatial pyramid pooling(Spatial pyramid matching)은 컴퓨터 비전 분야에서의 Bag-of-Words라고 볼 수 있다.

Introduction



Spatial Pyramid Pooling

이 BoW를 bins의 개념으로 위와 같이 나타낼 수 있다.

이렇게 하면 이미지 크기에 상관 없이 이미지의 각 구역에 있는 특징으로 이미지의 특징을 나타낼 수 있다.

Deep Networks with Spatial Pyramid Pooling

Convolutional Layers and Feature Maps

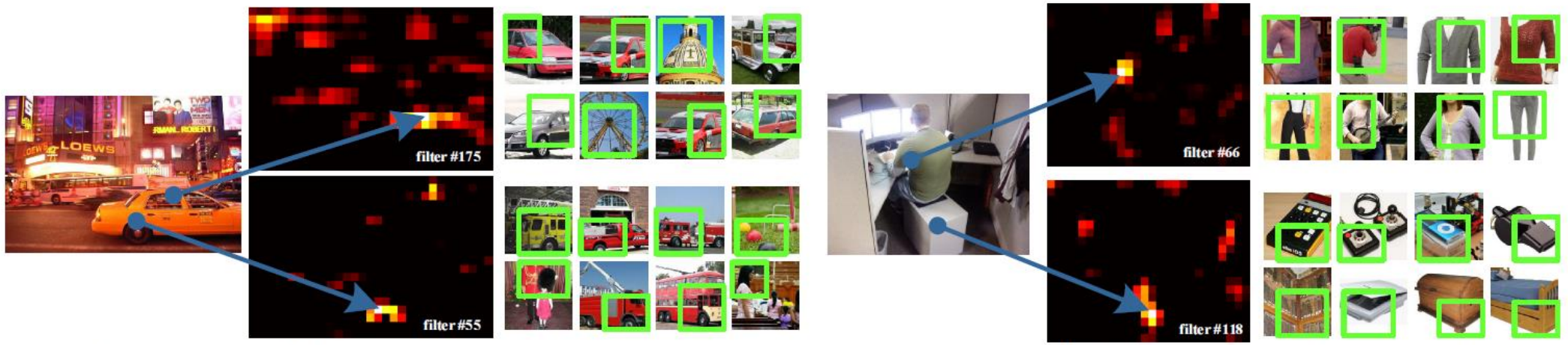


Figure 2: Visualization of the feature maps. (a) Two images in Pascal VOC 2007. (b) The feature maps of some $conv_5$ filters. The arrows indicate the strongest responses and their corresponding positions in the images. (c) The ImageNet images that have the strongest responses of the corresponding filters. The green rectangles mark the receptive fields of the strongest responses.

컨볼루션 계층의 연산 결과인 특징 맵은 각 커널이 강하게 반응하는 특징 정보나 그 특징 정보가 이미지 안에서 어디에 위치하고 있는지에 대한 위치 정보도 담고 있다.

또, 특징 맵의 크기는 동적으로 변화시키는 것이 가능하다(Padding, Stride, Pooling, Upsampling 등).

따라서 입력 이미지의 크기에 독립적이다.

Deep Networks with Spatial Pyramid Pooling

The Spatial Pyramid Pooling Layer

완전 연결 계층에 Pooling 연산을 거치고 나서 Flatten 시켜서 입력으로 넣는 것 대신에 오른쪽과 같이 Pooling 연산 계층을 SPP 계층으로 바꾼다.

SPP 계층에서는 BoW와 Bins 개념을 통해서 각 Bin 크기 별 Max Pooling 연산을 진행한다.

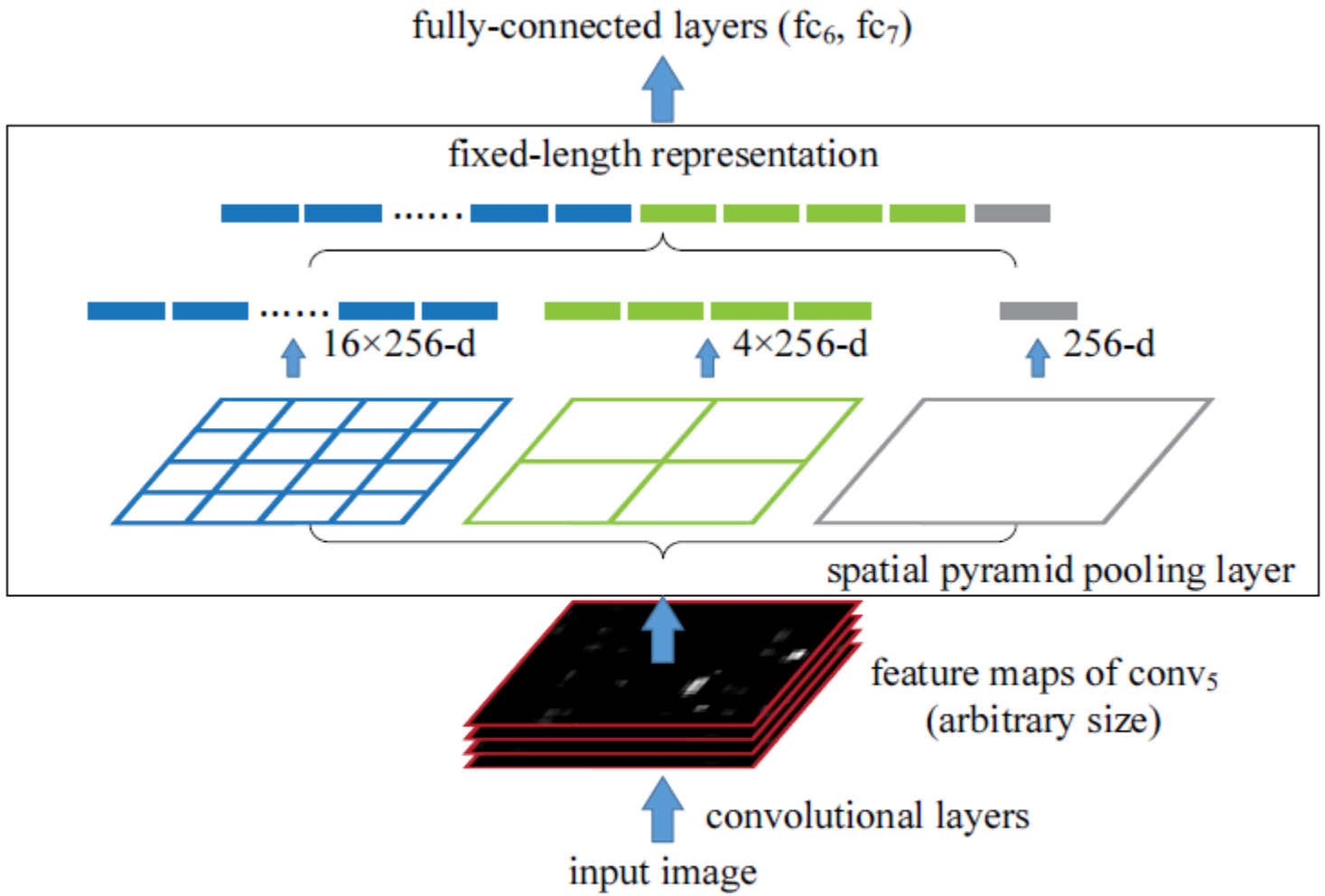


Figure 3: A network structure with a **spatial pyramid pooling layer**. Here 256 is the filter number of the $conv_5$ layer, and $conv_5$ is the last convolutional layer.

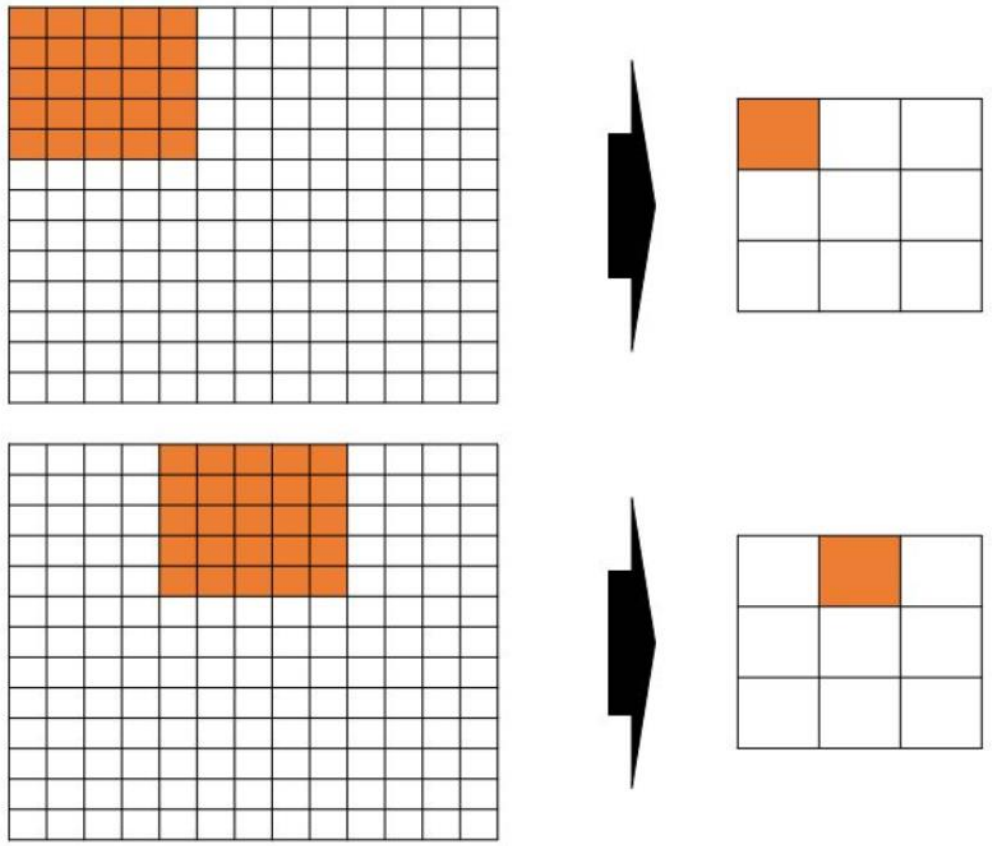
Deep Networks with Spatial Pyramid Pooling

The Spatial Pyramid Pooling Layer

이때 Max Pooling 연산에서의 커널 크기와 Stride 값은 아래의 공식에 의해서 정의된다.

특징 맵의 사이즈를 aXa라고 하고(예를 들어 13x13), bin의 개수가 nXn이라고 할 때(예를 들어 4x4 Max pooling) 윈도우 사이즈는 ceiling of a/n이고 stride는 floor of a/n.

13x13 특징 맵에서 3x3의 9개의 bin을 생성할 때, 커널의 크기가 5x5, stride 4



[Cheon Wujin - Paper Review - Object Detection 3 (Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition)](https://wujincheon.github.io/wujincheon.github.io/deep%20learning/2019/02/24/sppnet.html)

Deep Networks with Spatial Pyramid Pooling

Training the Network – Single size Training

단일 사이즈 훈련 시에는 224x224의 이미지로 훈련.

Training the Network – Multi size Training

예를 들어서 180x180, 224x224의 이미지로 훈련을 시킬 때, 동일 모델로 한 에폭은 180을 다음 에폭은 224 이미지로 훈련.

훈련 시간은 단일 사이즈 훈련과 비슷한 것을 저자들인 확인함.

테스트 시에는 훈련시킨 사이즈 범위 내에서 임의의 사이즈로 테스트 함.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Baseline Network Architectures

저자들이 SPP-Net의 효과를 검증하기 위해서 4개의 네트워크 아키텍처에 대해서 ImageNet 2012 Classification 과제로 실험을 진행했다. 이때 사용한 4개의 네트워크는 다음과 같다.

model	conv_1	conv_2	conv_3	conv_4	conv_5	conv_6	conv_7
ZF-5	96×7^2, str 2 LRN, pool 3^2, str 2 map size 55×55	256×5^2, str 2 LRN, pool 3^2, str 2 27×27	384×3^2 13×13	384×3^2 13×13	256×3^2 13×13	-	-
Convnet*-5	96×11^2, str 4 LRN, map size 55×55	256×5^2 LRN, pool 3^2, str 2 27×27	384×3^2 pool 3^2, 2 13×13	384×3^2 13×13	256×3^2 13×13	-	-
Overfeat-5/7	96×7^2, str 2 pool 3^2, str 3, LRN map size 36×36	256×5^2 pool 2^2, str 2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18

Table 1: Network architectures: filter number×filter size (*e.g.*, 96×7^2), filter stride (*e.g.*, str 2), pooling window size (*e.g.*, pool 3^2), and the output feature map size (*e.g.*, map size 55×55). LRN represents Local Response Normalization. The padding is adjusted to produce the expected output feature map size.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Multi-level Pooling Improves Accuracy

		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.

{6x6, 3x3, 2,x2, 1x1}의 총 50개의 bins를 사용 + 원본 이미지에 대해서 Standard 10 view - 센터, 4개의 코너, 위아래, 양 옆을 크롭 하여 사용.

Multi-level의 Pooling이 정확도 향상에 도움이 된다고 하는데 이는 단순히 모델 파라미터가 더 많기 때문이 아니고 이미지 내의 객체의 변형이나 위치 변경에 영향을 덜 받기 때문.

실제로 30 bins No-SPP보다 파라미터 수가 적은데도 불구하고 성능이 더 좋은 것을 확인함.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Multi-size Training Improves Accuracy

		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.

Table 2에 의하면 Single-size로 훈련시킬 때보다 Multi-size로 훈련시킬 때가 성능이 더 좋다. Standard 10 view 에 대한 예측 값을 뽑아냈고, 180과 224 사이즈에서 균등하게 샘플을 뽑아서 실험을 진행했다.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Full-Image Representations Improves Accuracy

Table2와 Table3를 보면 Multi-size로 훈련시킨 모델이 Single-size로 훈련시킨 모델보다 성능이 좋긴 하나 Single-size로 훈련시킨 모델도 나름의 장점을 가지고 있다.

1. 여러 View (이미지를 크롭한)와 상관 없이, 전체 이미지와 이를 Flipping시킨 이미지를, 모델을 훈련시킬 때 추가하면 성능이 더 좋아진다.

2. 전체 이미지를 사용하는 것은 마찬가지로 전체 이미지를 사용하는 전통적인 방법(SIFT 등)과 어느정도 부합한다.

3. 이미지 검색 같은 다른 Application에서는 전체 이미지에 대한 특징이 필요할 수 있다.

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.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Multi-view Testing on Feature Maps

SPP가 입력 이미지에 상관 없이 동일한 크기의 특징 벡터를 뽑아낼 수 있다는 유연함 덕분에 위와 같이 전체 특징맵이 아니라 특정 Window에 있는 특징으로 동일한 크기의 특징 벡터를 뽑아낼 수 있다.

기존의 Standard 10-view 테스트가 아니라 18 view 테스트(단 224 사이즈의 이미지는 제외)를 진행했다고 한다. 이때 크기는 6개의 크기에 대해({224; 256; 300; 360; 448; 560}) 실험을 했다. 그래서 18x5 + 6 = 96 View에 대해서 실험을 했다.

이 방법으로 Top5 에러율이 10.95%에서 9.36%로 감소하기도 했다. 추가적으로 Full image view 2가지(Flipping 포함)를 추가 했을때는 9.14%까지 에러율이 감소했다.

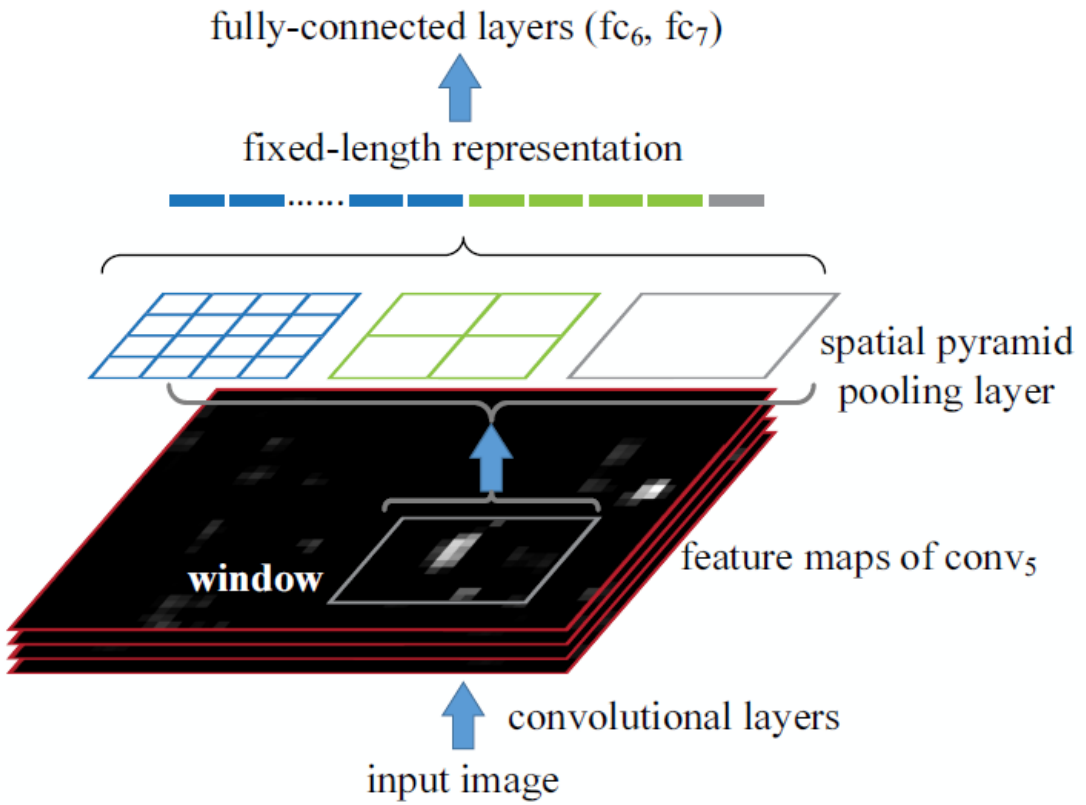


Figure 5: Pooling features from arbitrary windows on feature maps. The feature maps are computed from the entire image. The pooling is performed in candidate windows.

SPP-Net for Image Classification

Experiments on ImageNet 2012 Classification – Summary and Results for ILSVRC 2014

method	test scales	test views	top-1 val	top-5 val	top-5 test
Krizhevsky *et al.* [3]	1	10	40.7	18.2	
Overfeat (fast) [5]	1	-	39.01	16.97	
Overfeat (fast) [5]	6	-	38.12	16.27	
Overfeat (big) [5]	4	-	35.74	14.18	
Howard (base) [36]	3	162	37.0	15.8	
Howard (high-res) [36]	3	162	36.8	16.2	
Zeiler & Fergus (ZF) (fast) [4]	1	10	38.4	16.5	
Zeiler & Fergus (ZF) (big) [4]	1	10	37.5	16.0	
Chatfield *et al.* [6]	1	10	-	13.1	
ours (SPP O-7)	1	10	29.68	10.95	
ours (SPP O-7)	6	96+2full	**27.86**	**9.14**	**9.08**

Table 4: Error rates in ImageNet 2012. All the results are based on **a single network**. The number of views in Overfeat depends on the scales and strides, for which there are several hundreds at the finest scale.

rank	team	top-5 test
1	GoogLeNet [32]	**6.66**
2	VGG [33]	7.32
3	ours	8.06
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.

SPP-Net for Object Detection

생성된 지역 약 2000개에 대해서 각각 CNN을 돌리는 R-CNN과는 달리 SPP-net은 여러 크기의 각 이미지에 대해 한 번 CNN을 돌려 특징 맵을 뽑아 낸 후에 각 Candidate window에 대해 고정된 길이의 특징 벡터를 도출해낸다.

CNN 연산이 전체 네트워크에서 시간을 소비하는 비중이 크기 때문에 R-CNN보다 훨씬 빠르게 Object-Detection을 수행할 수 있다고 한다.

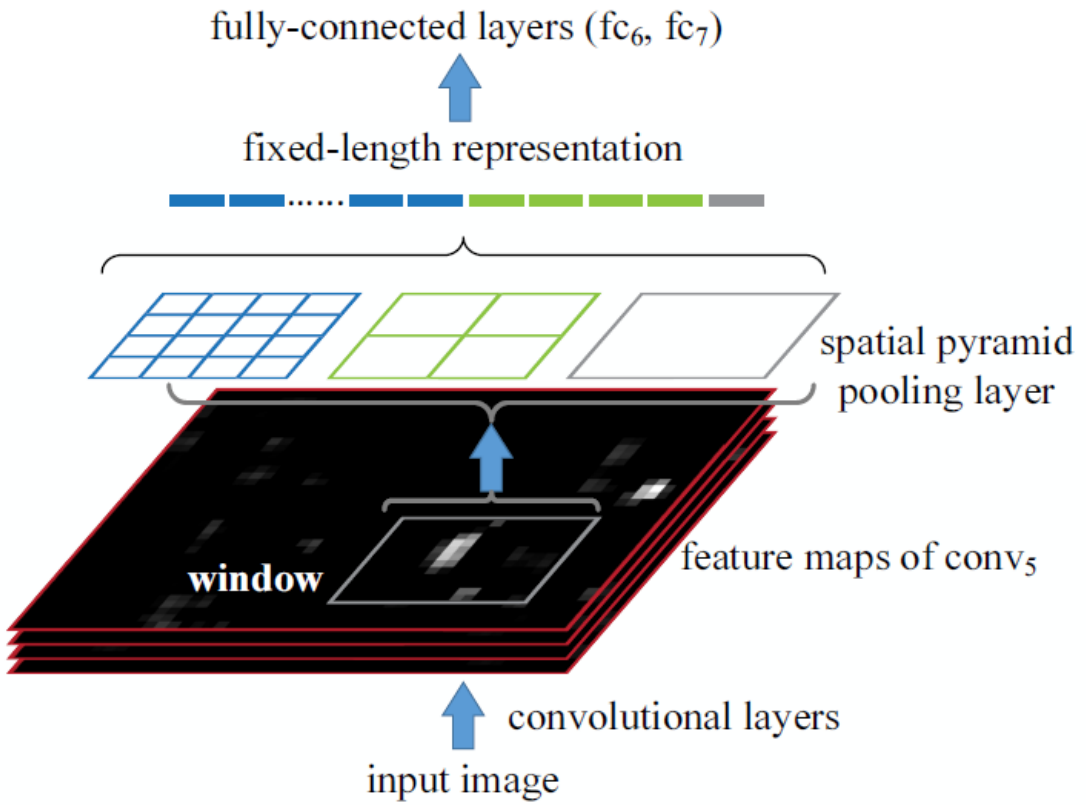


Figure 5: Pooling features from arbitrary windows on feature maps. The feature maps are computed from the entire image. The pooling is performed in candidate windows.

SPP-Net for Object Detection

Detection Algorithm

1. Candidate Windows를 생성하는 방법은 각 이미지에 대해서 Fast Selective search를 수행해서 2,000개의 Windows를 생성해낸다.

2. 각 Candidate Windows에 대해서 4-level Spatial pyramid를 적용한다(1x1, 2x2, 3x3, 6x6으로 총 50 bins). 그래서 12,800 차원의 벡터가 만들어진다(256x50).

3. 이 벡터들이 완전 연결 계층의 입력으로 들어가고 그래서 나온 결과로 각 카테고리에 대한 이진 SVM classifier를 훈련시킨다.

4. SVM을 훈련시킬 때는 Ground-truth windows들을 Positive로, Positive window와 많아 봤자 30%로 겹치는 window는 Negative로 설정한다.

5. Negative 샘플 중에서 다른 Negative 샘플과 70% 겹치는 window가 있다면 제거한다.

6. 이렇게 만들어진 Negative 샘플에 대해 Hard negative mining을 수행한다.

7. CNN에서 완전 연결 계층에 대해서만 Fine-tuning을 진행했다. Fine-tuning 동안에 Positive 샘플은 Ground-truth windows와 0.5~1만큼 겹치는 샘플을, Negative 샘플은 0.1~0.5미만으로 겹치는 샘플로 설정.

8. 바운딩 박스 회귀도 수행했다. 회귀를 위해 사용된 특징은 마지막 컨볼루션 계층에서 도출된 특징들이고 회귀가 수행되는 windows들은 Groud-truth window와 적어도 0.5이상 겹치는 windows들이다.

SPP-Net for Object Detection

Detection Results

	SPP (1-sc) (ZF-5)	SPP (5-sc) (ZF-5)	R-CNN (Alex-5)
$pool_5$	43.0	44.9	44.2
fc_6	42.5	44.8	46.2
$ftfc_6$	52.3	53.7	53.1
$ftfc_7$	54.5	55.2	54.2
$ftfc_7$ bb	58.0	**59.2**	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**×	-

Table 9: Detection results (mAP) on Pascal VOC 2007. "ft" and "bb" denote fine-tuning and bounding box regression.

	SPP (1-sc) (ZF-5)	SPP (5-sc) (ZF-5)	R-CNN (ZF-5)
$ftfc_7$	54.5	55.2	55.1
$ftfc_7$ bb	58.0	**59.2**	**59.2**
conv time (GPU)	0.053s	0.293s	14.37s
fc time (GPU)	0.089s	0.089s	0.089s
total time (GPU)	0.142s	0.382s	14.46s
speedup (*vs.* RCNN)	**102**×	**38**×	-

Table 10: Detection results (mAP) on Pascal VOC 2007, **using the same pre-trained model** of SPP (ZF-5).

위에서 1-sc는 단일 스케일, 5-sc는 5개의 스케일을 의미한다.

SPP-Net for Object Detection

Model Combination for Detection

method	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
SPP-net (1)	59.2	**68.6**	69.7	57.1	41.2	40.5	66.3	71.3	72.5	34.4	**67.3**	61.7	63.1	71.0	69.8	57.6	29.7	59.0	50.2	65.2	68.0
SPP-net (2)	59.1	65.7	71.4	57.4	**42.4**	39.9	67.0	71.4	70.6	32.4	66.7	61.7	64.8	71.7	70.4	56.5	30.8	59.9	53.2	63.9	64.6
combination	**60.9**	68.5	**71.7**	**58.7**	41.9	**42.5**	**67.7**	**72.1**	**73.8**	**34.7**	67.0	**63.4**	**66.0**	**72.5**	**71.3**	**58.9**	**32.8**	**60.9**	**56.1**	**67.9**	**68.8**

Table 12: Detection results on VOC 2007 using model combination. The results of both models use "ftfc$_7$ bb".

앙상블 같은 Model combination 기법들은 테스트 성능을 더 높이기도 한다. 저자들도 Detection에서의 Model combination을 시도했다. 두 가지 모델이 있을 때, 각 모델의 결과로 테스트 이미지의 각 Candidate windows에 대해 Scoring을 하고 NMS을 두 결과에 대해서 수행한다.

SPP-Net for Object Detection

ILSVRC 2014 Detection

rank	team	mAP
1	NUS	**37.21**
2	ours	35.11
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.

Conclusion

SPP-layer를 통해서 입력 크기나 종횡비에 상관 없이 모델을 훈련시키는 방법을 제시함으로써 정확도와 속도 측면에서 기존의 방법보다 성능이 좋은 새로운 방법을 제시하고자 함.