Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

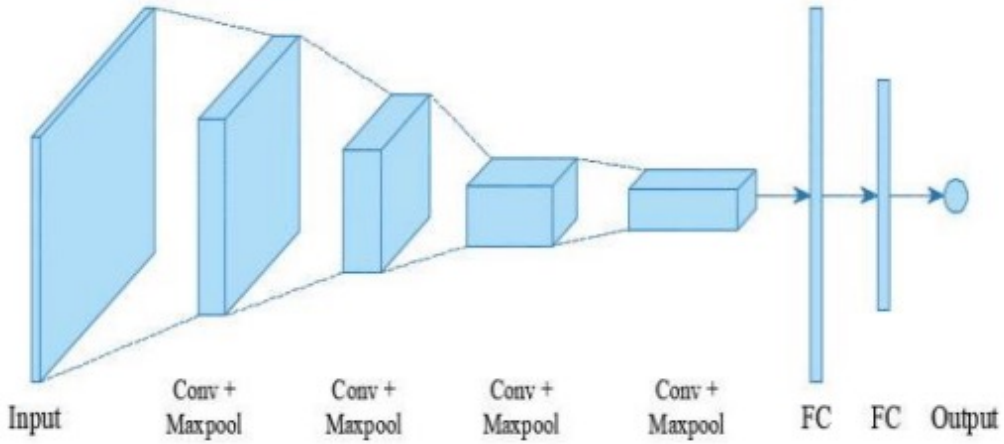
육현준

ETRI
한국전자통신연구원

SPP-net

연구 계기

- 기존 CNN
  - Convolution Layer
  - Fully-Connected Layer → 고정된 크기의 입력 필요



SPP-net

연구 계기

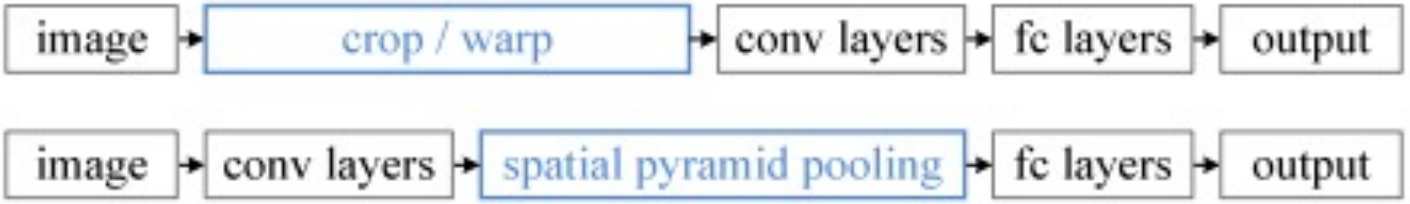
Crop 또는 Warp을 통해 입력 이미지를 임의로 변형

- 이미지 변형으로 정보의 손실 발생

SPP-net

연구 계기

- Spatial Pyramid Pooling
  - ➢ 입력 이미지의 크기에 상관없이 항상 같은 크기의 벡터를 생성



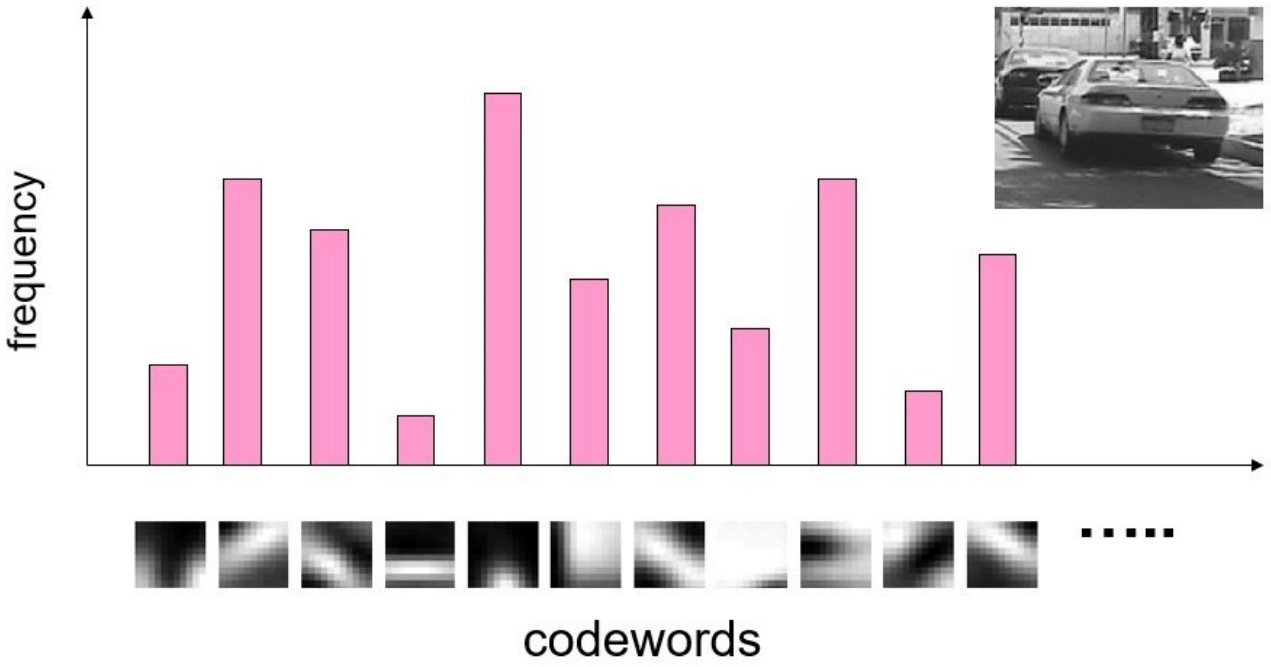
입력 이미지 변형 불필요

➢ 정보의 손실 x

SPP-net

Bag of Visual Words

Feature Extraction → K-Means → histogram



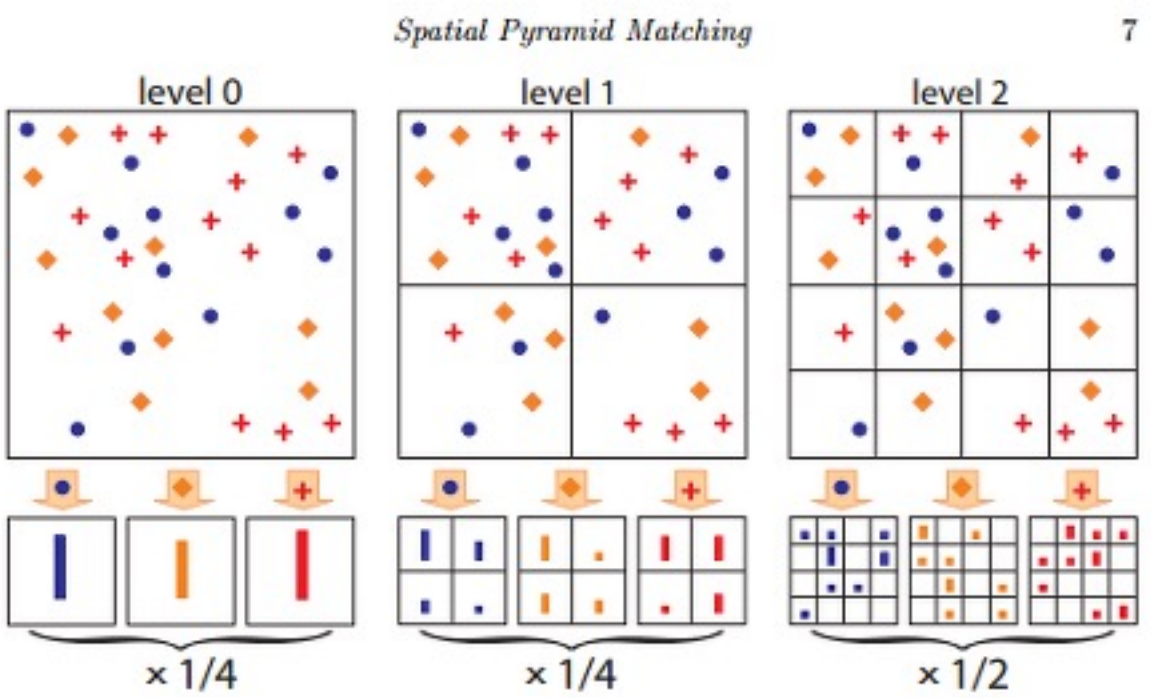
SPP-net

Bag of Visual Words

단점 : 위치정보를 잃는다.

SPP-net

Spatial Pyramid Matching



지역 위치정보 보존

SPP-net

Spatial Pyramid Pooling

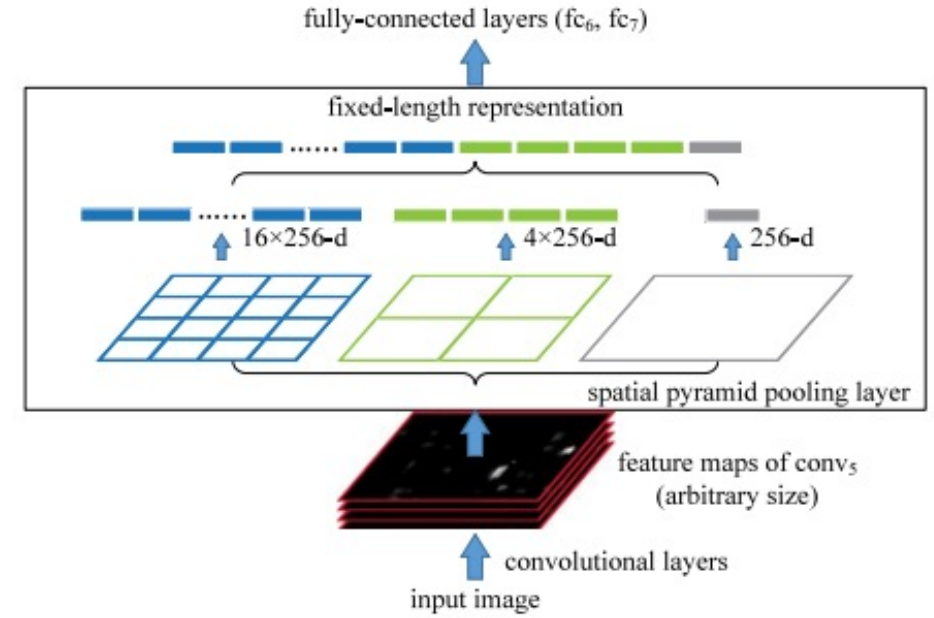


Fig. 3. A network structure with a *spatial pyramid pooling layer*. Here 256 is the filter number of the $conv_5$ layer, and $conv_5$ is the last convolutional layer.

- bin의 개수 고정
  - M
- 마지막 Convolution Layer의 filter의 개수 고정
  - K

- 결과적으로 KM 차원의 고정된 output을 얻을 수 있다.

SPP-net <Classification>

Multi-Size Training

- 1 epoch
  - ➢ (180 x 180) 크기의 이미지로 학습
- 2 epoch
  - ➢ (224 x 224) 크기의 이미지로 학습

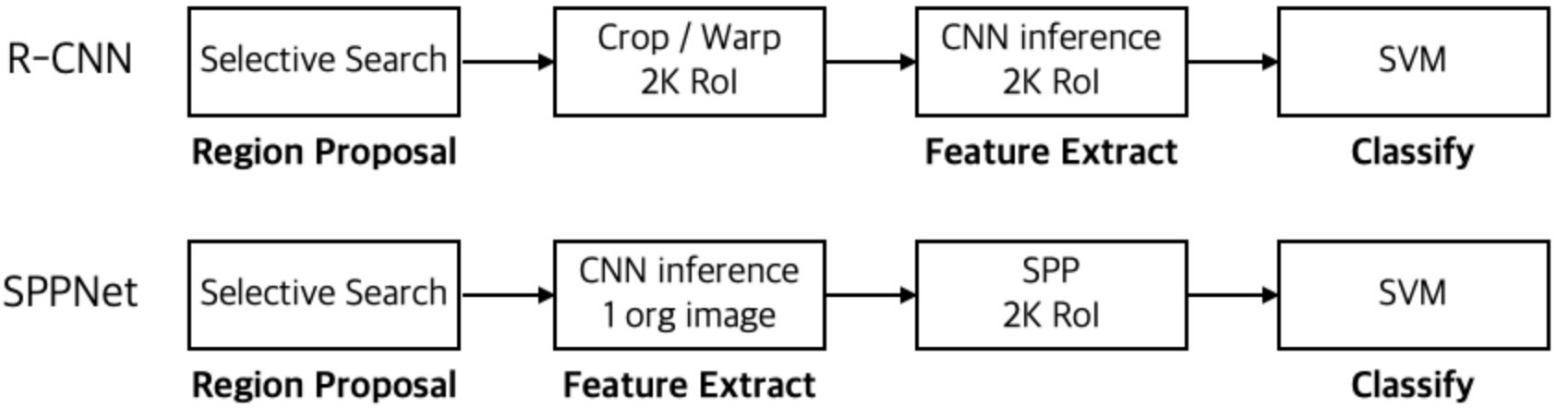
TABLE 2
Error Rates in the Validation Set of ImageNet 2012

		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)

All the results are obtained using standard 10-view testing. In the brackets are the gains over the "no SPP" baselines.

SPP-net <Detection>



R-CNN

- Crop/Warp : 이미지 변형
- 이미지에서 Object가 존재할 후보영역 2000개 추출
- 한 이미지당 CNN을 2000번 반복 → 매우 느리다.

SPP-net

- 마지막 Convolution Layer의 Feature map에서 Selective Search 진행
- SPP Layer → 2000개의 동일한 크기의 벡터 생성

➢ 한 이미지당 CNN을 단 한번 통과 → 속도 향상

SPP-net <Detection>

TABLE 10
Detection Results (mAP) on Pascal VOC 2007, **Using the Same Pre-Trained Model** of SPP (ZF-5)

	SPP (1-sc) (ZF-5)	SPP (5-sc) (ZF-5)	R-CNN (ZF-5)
$ftfc_7$	54.5	<u>55.2</u>	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**×	-

- R-CNN보다 매우 빠르고 비슷한 성능

참고문헌

Sakib, S.; Ahmed, N.; Kabir, A.J.; Ahmed, H. An Overview of Convolutional Neural Network: Its Architecture and Applications. Preprints 2018, 2018110546 (doi: 10.20944/preprints201811.0546.v4).

K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

https://towardsdatascience.com/bag-of-visual-words-in-a-nutshell-9ceea97ce0fb

S. Lazebnik, C. Schmid and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 2169-2178, doi: 10.1109/CVPR.2006.68.