주제: **SPP-Net**

논문저자가 누구**?**

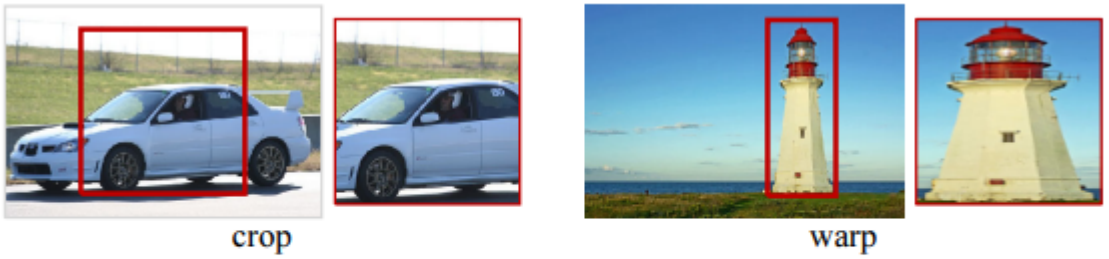
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

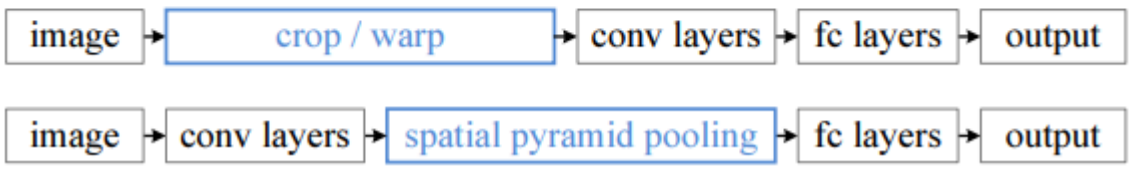
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun

- 인용수 5831회의 후덜덜한 논문 (2021-03-37 기준)
- 1저자인 He는 ResNet 만든 사람임…!

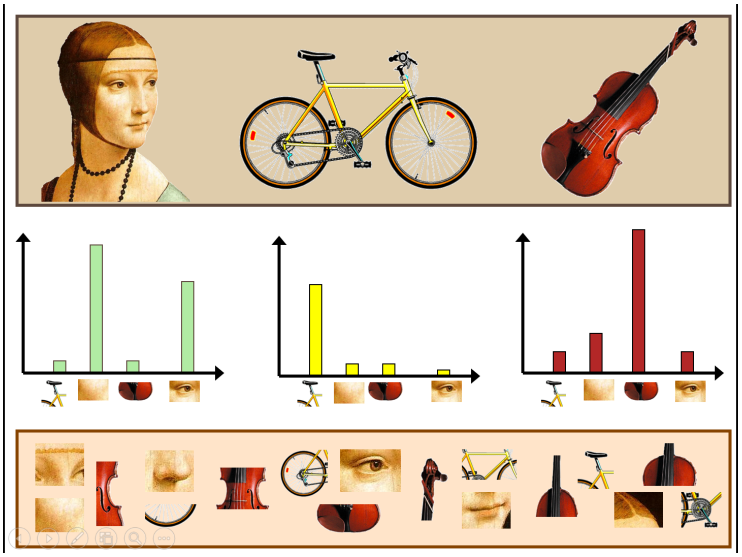
배경

- 과거에 사용된 CNN 구조들을 트레이닝하고 테스트할 때 발생하는 기술적 문제: 고정된 input image size (e.g., 224x224) 필요 => Fully-connected layer에서 고정된 길이의 입력이 필요하기 때문에 고정된 크기의 이미지를 필요로 함
    - 이미지 왜곡 가능성: crop / warp 과정에서 이미지가 짤리거나 왜곡되면서 손상되는 현상 발생 -> 인식 저하로 이어질 수 있음
    - 다양한 스케일의 사물에 적합하지 않을 수 있음
- 해결책: SPP layer를 통해서 임의의 크기의 이미지에 대해서 항상 같은 길이의 벡터를 만들고자 함
    - 입력 이미지의 크기를 조절하지 않을 채로 conv layer를 통과시키면 원본 이미지의 특징을 그대로 간직한 피쳐맵 && 사물의 크기 변화에 더 견고한 모델을 얻을 수 있음
    - Image classification이나 object detection 등의 테스크들에 일반적으로 적용 가능해짐





SPP-Net: an extension of Bag-of-Words model

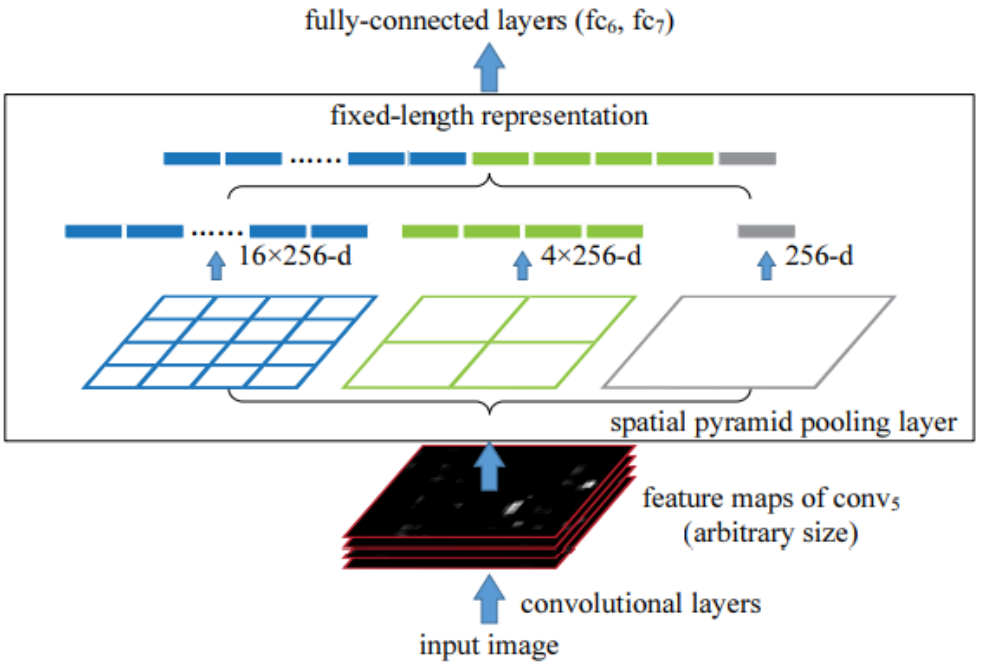


- *(출처: Recognizing and Learning Object Categories; ICCV 2005 short course)*
- Feature extraction - 이미지로부터 피쳐를 추출(SIFT 등)
- Clustering - 피쳐들을 클러스터링하여 코드값을 구함
    - 코드값 = 분류된 클러스터들의 대표값
- Codebook - 코드값들이 모인 코드북을 생성
- Image Representation - 이미지를 코드값들의 히스토그램으로 표현
    - 이미지마다 어떤 피처의 값이 어느정도 있는지 확인 가능
- Learning and Recognition - svm 등의 분류기로 학습하여 이미지를 분류

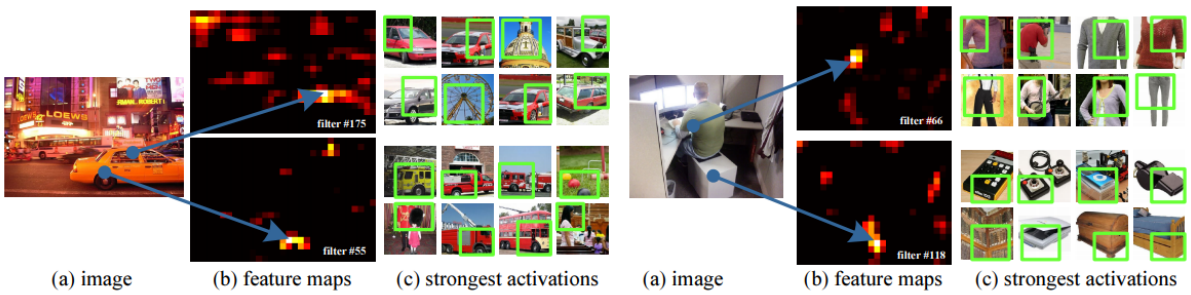
SPP-Net - 전체 알고리즘

1. 먼저 전체 이미지를 미리 학습된 CNN을 통과시켜 피쳐맵을 추출
2. Selective Search를 통해서 찾은 제 각기 크기와 비율이 다른 ROI들에 SPP를 적용하여 고정된 크기의 feature vector 추출
3. 그 다음 fully connected layer들을 통과 시킴
4. 앞서 추출한 벡터로 각 이미지 클래스 별로 binary SVM Classifier를 학습시킴
5. 마찬가지로 앞서 추출한 벡터로 bounding box regressor를 학습시킴

SPP-Net - Contribution

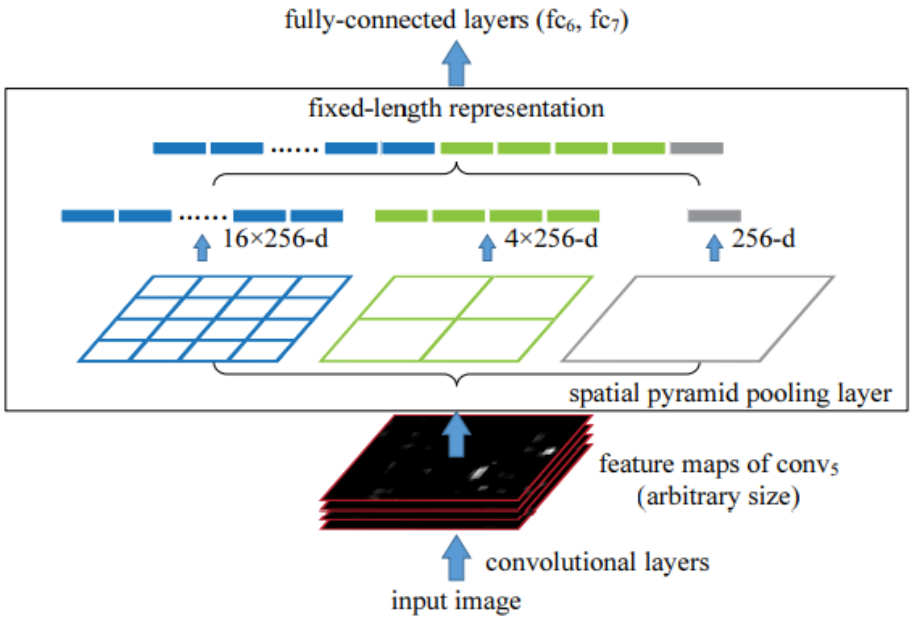


- 고정된 input 크기가 아닌 임의의 크기 이미지에 대해 CNN으로 학습을 하고 conv 5 layer에서 추출한 피쳐맵을 input으로 받음
    - R-CNN과 다르게 전체 이미지에 대해 한 번만 수행하므로 수 백배 빠름
- 다양한 크기의 bin을 통해 고정된 길이의 결과값을 만듦과 동시에, 다양한 scale의 특징을 추출
    - bin 크기는 개인이 설정하기 나름
    - 다양한 특징들을 가진 고정된 길이의 fixed-length representation 추출 가능
- Classification과 detection 모두에서 강점을 보임
- ECCV 2014 초기 버전 발행 후, ILSVRC 2014에 출전하여 좋은 성적을 보이면서 피쳐맵에 대한 multi-view test가 classification task에 좋은 성능을 가져옴을 알게 됨
    - 특정 scale에 대해 overfitting되는 다른 모델들의 문제점을 해결



- 임의의 크기의 이미지에 대해 conv층을 거쳐서 만들어진 피쳐맵을 시각화한 것
- 피쳐맵에 대해서 특정 필터들에 의해 activate된 이미지를 볼 수 있음
- 이미지에서 특징을 찾아내는 필터에 대한 반응의 강도와 그 공간적인 위치를 나타내는 이미지

SPP-Net - Spatial Pyramid Pooling Layer



- Input = conv layer에서 추출된 피쳐맵
- Bin을 일정한 수(M개)로 고정시키고 각 공간 bin에 필터들(k개)의 반응정도를 pooling
    - 예시에 있는 4x4, 2x2, 1x1 영역을 각각 하나의 피라미드로 부름
    - 피라미드 한칸 = bin
    - 각 bin에서 가장 큰 값만 추출하는 max pooling 수행한 후 그 결과를 쭉 이어붙여줌
- 그렇게 되면 총 k*M 차원의 고정길이 벡터가 만들어짐
    - bin은 4x4, 2x3, 1x1으로 21개
    - 필터는 conv5의 필터 수로 256개
    - -> 즉 21*256차원의 고정길이 벡터가 만들어짐
- 입력 이미지의 크기와는 상관없이 미리 설정한 bin의 개수와 CNN 채널 값으로 SPP의 출력이 결정되므로, 항상 동일한 크기의 결과를 리턴
    - bin의 크기를 늘리면 좀 더 다양한 특징들을 이끌어내고 성능을 향상 시킬 수 있지만, 복잡도가 증가함

SPP-Net - Training the Network

- Single-size training: 앞서 소개했듯이 spp layer에 필요한 bin의 사이즈를 미지 계산한다. 여러 크기 조합의 빈을 통해서 다중 수준의 pooling을 가능하게 하여 정확도를 높이고자 함
- Multi-size training: 여기서는 180x180과 224x224 두 종류에 대해서 학습을 진행함. 224의 사이즈를 180의 사이즈로 resize해서 내용과 배치를 동일하게 만듦. 그리고 두 스케일에 대해서 같은 parameter를 공유하도록 두 개의 네트워크를 학습시킴. 임의의 크기에 대해서도 적용이 가능

SPP-Net: Classification - Dataset

- ImageNet 2012의 1000-category train set으로 학습
- 분류할 이미지를 256x256의 사이즈로 크기 조절
- 조절한 이미지에 대해 224x224의 사이즈로 crop(코너4개와 센터 및 좌우대칭)하여 augmentation
- SPPnet을 거친 feature vector에 softmax 스코어를 계산하여 분류에 사용

SPP-Net: Classification - Baseline

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

- SPPNet의 장점은 conv네트워크 구조에 독립적이라는 것
- 그래서 기존에 있던 4개의 네트워크 구조에 대해서 SPPNet을 적용하여 개선됨을 실험하고자 함

SPP-Net: Classification - Baseline + SPP

		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)

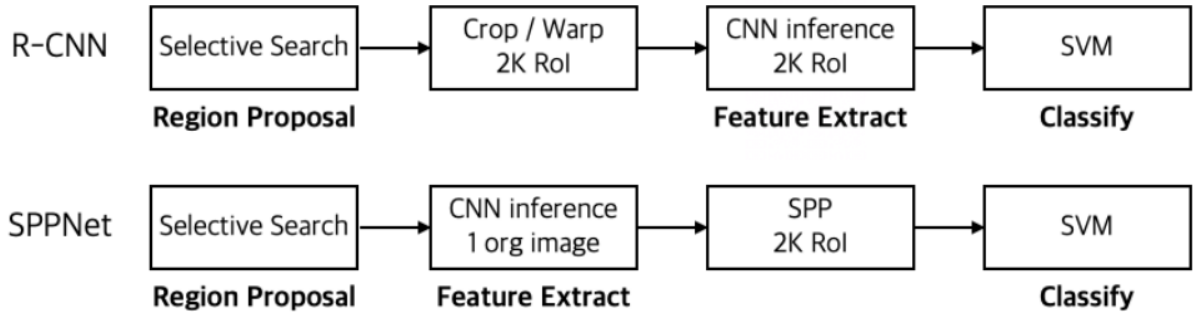
- 총 4-level pyramid를 사용해서 50개의 bin을 사용
- SPP를 사용하지 않은 것보다 다중 수준의 pooling을 진행하면 parameter를 더 사용하면서 object 변형이나 공간에 대해서 robust하기 때문에 에러율이 향상됨
- 그리고 multi-size 학습을 하게 되면, no SPP나 single-size보다 향상됨

SPP-Net: Classification - representational power of full image

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**

- crop하지 않은 full-size 이미지를 사용했을 때가 에러율이 향상됨 => 완전한 content 확보의 중요성

R-CNN vs. SPP-Net



한계점

1. end-to-end 방식이 아니라 학습에 여러 단계가 필요함 (fine-tuning, SVM training, Bounding Box Regression)
2. 최종 classification이 binary SVM, Region proposal은 selective search 이용
3. fine tuning시에 spp를 거치기 이전의 conv 레이어들을 학습시키지 못함