ILSVRC 2014
rank #2 in detection
#3 in classification

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He[1], Xiangyu Zhang[2], Shaoqing Ren[3], Jian Sun[1]
[1]Microsoft Research, [2]Xi'an Jiaotong University, [3]University of Science and Technology of China

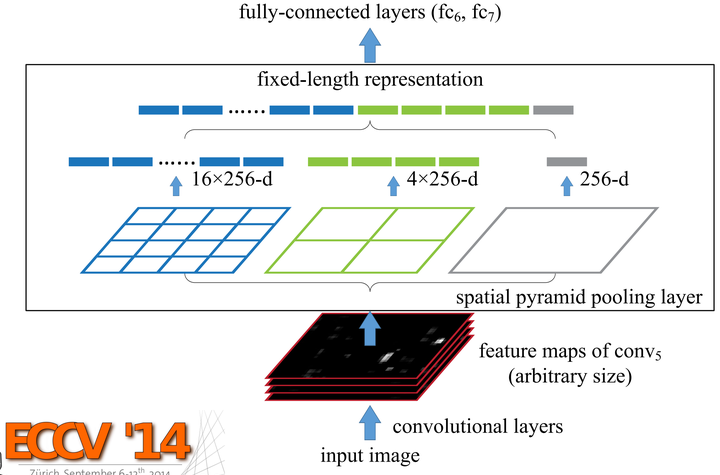
Microsoft Research

Highlights of SPP-net

- Classification: improves all CNN architectures
- Detection: 24-64x faster than R-CNN
- **ILSVRC 2014**: **#2** in detection, **#3** in classification. All details disclosed.

What is SPP-net?

- SPP-net is a new network with **Spatial Pyramid Pooling** (SPP)



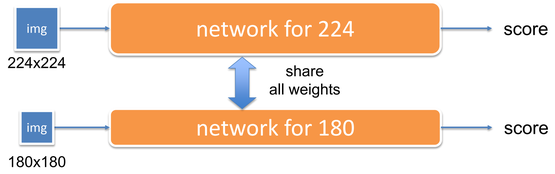
Training SPP-net

- **Single-size training**: simply modify the configuration file

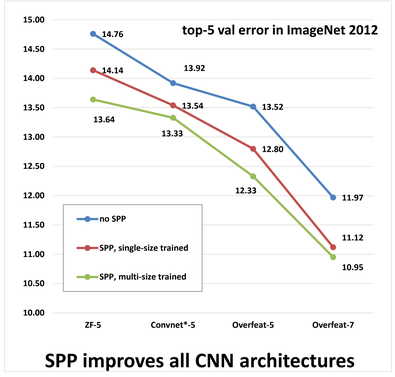
[pool3x3]	[pool2x2]	[pool1x1]	[fc6]
type=pool	type=pool	type=pool	type=fc
pool=max	pool=max	pool=max	outputs=4096
inputs=conv5	inputs=conv5	inputs=conv5	inputs=pool3x3,pool2x2,pool1x1
sizeX=5	sizeX=7	sizeX=13	
stride=4	stride=6	stride=13	

example for a 13x13 feature map

- **Multi-size training**: Multiple networks sharing all weights; each network for a single size. Improves scale-invariance.



ILSVRC 2014 Classification



SPP improves all CNN architectures

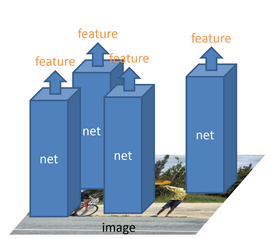
single-model	top-5 val	
Convnet [1]	18.2	ILSVRC 12
ZF [2]	16.0	ILSVRC 13
Howard's [3]	15.8	
Overfeat [4]	14.18	
ours (10-view)	10.95	
ours (96+2-full-view)	**9.14**	

rank #3 in ILSVRC 2014

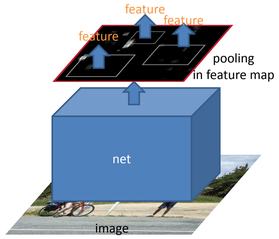
multi-model	top-5 test
GoogLeNet	**6.66**
Oxford VGG	7.32
ours	**8.06**
Howard	8.11
DeeperVision	9.50
NUS-BST	9.79
TTIC_ECP	10.22
...	

Fast CNN-based Object Detection

- **R-CNN vs. SPP-net**: image regions *vs*. feature map regions
- With features => fine-tuning, SVM, bbox regression



R-CNN: 2000 runs of nets



SPP-Net: 1 net on full image

VOC 2007	SPP-net (1-scale)	SPP-net (5-scale)	RCNN
FT fc7	54.5	**55.2**	54.2
FT fc7 + bbox	58.0	**59.2**	58.5
GPU feature time	0.142s	0.382s	9.03s
speed-up vs. RCNN	**64x**	**24x**	-

ILSVRC 2014 Detection

rank #2 in ILSVRC 2014

	mAP
NUS	**37.21**
ours	**35.11**
UvA	32.02
ours (single-model)	31.84
Southeast-CASIA	30.47
1-HKUST	28.86
CASIA_CRIPAC_2	28.61

internal-data track

more practical than R-CNN

	SPP-net	RCNN
GPU feature time / img	0.6s	32s
40k testing imgs	8 GPU hours	15 GPU **days**

*conv feature extracting time (Ovefeat-7 architecture, 1-model)

Code, network config, technical report with all details:
http://research.microsoft.com/en-us/um/people/kahe/

References
[1] Imagenet classification with deep convolutional neural networks. NIPS 12
[2] Visualizing and Understanding Convolutional Neural Networks. arXiv 13
[3] Some Improvements on Deep Convolutional Neural Network Based Image Classification. arXiv 13
[4] OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. arXiv 13
[5] Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 14

ECCV '14
Zürich, September 6-12th, 2014