Review: SPPNet —1st Runner Up (Object Detection), 2nd Runner Up (Image Classification) in ILSVRC 2014

SH Tsang Follow

Sep 1, 2018 · 6 min read

In this story, SPPNet is reviewed. SPPNet has introduced a new technique in CNN called **Spatial Pyramid Pooling (SPP)** at the transition of convolutional layer and fully connected layer. This is a work from **Microsoft**.

In **ILSVRC 2014**, SPPNet has got **1st Runner Up in Object Detection**, **2nd Runner Up in Image Classification**, and **5th Place in Localization Task**. And it got 2 papers in **2014 ECCV [1] and 2015 TPAMI [2]** with **over 1000 and 600 citations** respectively. Thus, SPPNet is one of the worth reading deep learning papers. (SH Tsang @ Medium)

. . .

Dataset

Classification: Over 15 millions labeled high-resolution images with around 22,000 categories. ILSVRC uses a subset of ImageNet of around 1000 images in each of **1000 categories**. In all, there are roughly **1.3M/50k/100k images** are used for the **training/validation/testing sets**

Detection: **200 categories**. **450k/20k/40k** images are used for the **training/validation/testing sets**

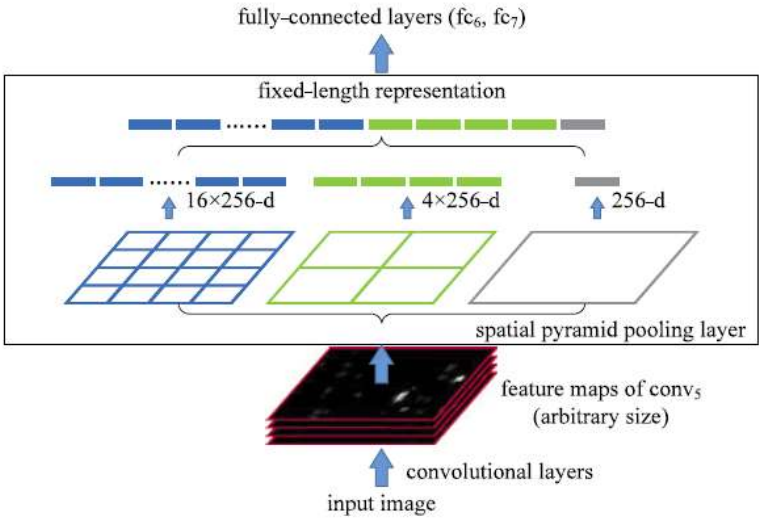
. . .

What are covered

1. **Spatial Pyramid Pooling (SPP)**
2. **Multi-Size Training**
3. **Full Image Representation**
4. **Multi-View Testing**
5. **Comparison with State-of-the-art Approaches (Classification)**
6. **SPPNet in Object Detection**
7. **Comparison with State-of-the-art Approaches (Detection)**

. . .

1. Spatial Pyramid Pooling (SPP)



Three-Level Spatial Pyramid Pooling (SPP) in SPPNet with Pyramid {4×4, 2×2, 1×1}.

Conventionally, at the transition of conv layer and FC layer, there is one single pooling layer or even no pooling layer. In SPPNet, it suggests to have **multiple pooling layers with different scales**.

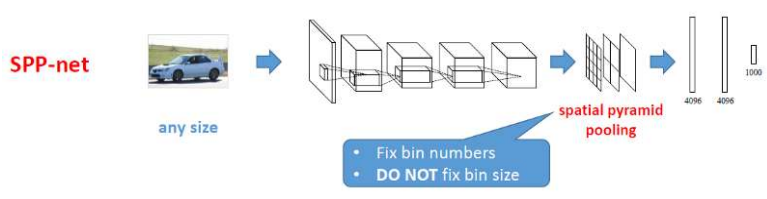
In the above figure, **3-level SPP** is used. Suppose the conv5 layer has 256 feature maps. Then at SPP layer,

1. first, each feature map is **pooled to become one value (grey)**, thus **256-d vector is formed**.
2. Then, each feature map is **pooled to have 4 values (green)**, and form a **4×256-d vector**.
3. Similarly, each feature map is **pooled to have 16 values (blue)**, and form a **16×256-d vector**.
4. The **above 3 vectors are concatenated to form a 1-d vector**.
5. Finally, this **1-d vector is going into FC layers** as usual.

With SPP, we don't need to crop the image to fixed size, like AlexNet, before going into CNN. **Any image sizes can be inputted.**

• • •

2. Multi-Size Training

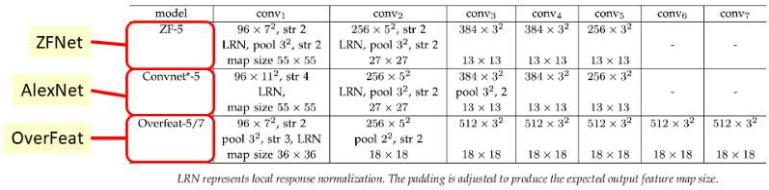


SPPNet supports any sizes due to the use of SPP

With SPP, variable sizes are accepted as input, different sizes should be inputted into network to increase the robustness of the network during training.

However, for the effectiveness of the training process, **only 224×224 and 180×180 images are used as input.** Two networks, 180-network and 240-network are trained with shared parameters.

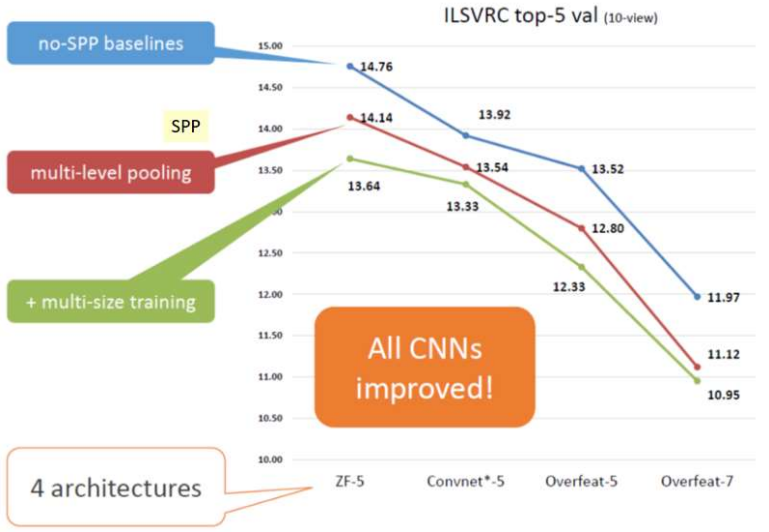
Authors replicated ZFNet [3], AlexNet [4] and Overfeat [5] with modifications as below (The number after it is the conv layer number):



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

LRN represents local response normalization. The padding is adjusted to produce the expected output feature map size.

Replicated Model as Baseline



Top-5 Error Rates for SPP and Multi-Size Training

4-level SPPNet is used here with the pyramid **{6×6, 3×3, 2×2, 1×1}.**

As shown above, **with SPP only, the error rates have been improved for all models. With Multi-Size Training, the error rates improve further.** (10-view means the 10-crop testing from [four corners + 1 center] and the corresponding horizontal flips)

• • •

3. Full Image Representation

As full image can also input into CNN with the use of SPP, authors **compare full image input with only using 1 center crop input**:

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**

The images are resized so $\min(w, h) = 256$. *The crop view is the central* 224×224 *of the image.*

Top-1 Error Rates for Full Image Representation

Top-1 error rates are all improved with full image as input.

• • •

4. Multi-View Testing

With full image support by using SPP, multi-view testing can be facilitated easily.

1. Authors resize the image to **6 scales: {224, 256, 300, 360, 448, 560}**
2. For each scale, **18 views are generated: {1 center, 4 corners, 4 on the middle of each side} and the corresponding flips.** Thus, there are **in total 96 views**.
3. And **2 full-image views plus corresponding flips** are also included.

	method	test scales	test views	top-1 val	top-5 val	top-5 test
AlexNet	Krizhevsky *et al.* [3]	1	10	40.7	18.2	
OverFeat	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	
ZFNet	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	
SPPNet using OverFeat-7	ours (SPP O-7)	1	10	29.68	10.95	
	ours (SPP O-7)	6	96+2full	**27.86**	**9.14**	**9.08**

Error Rates in ILSVRC 2012 (All are Single Model Results)

SPPNet using OverFeat-7 obtains 9.14/9.08% Top-5 Error Rate on validation/test set which is **the only one under 10%** in the table.

• • •

5. Comparison with State-of-the-art Approaches (Classification)

11 models of SPPNet are used in testing. The outputs are averaged to get a more accurate prediction. This is a boosting or ensemble technique used in many CNN models such as LeNet, AlexNet, ZFNet.

team	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

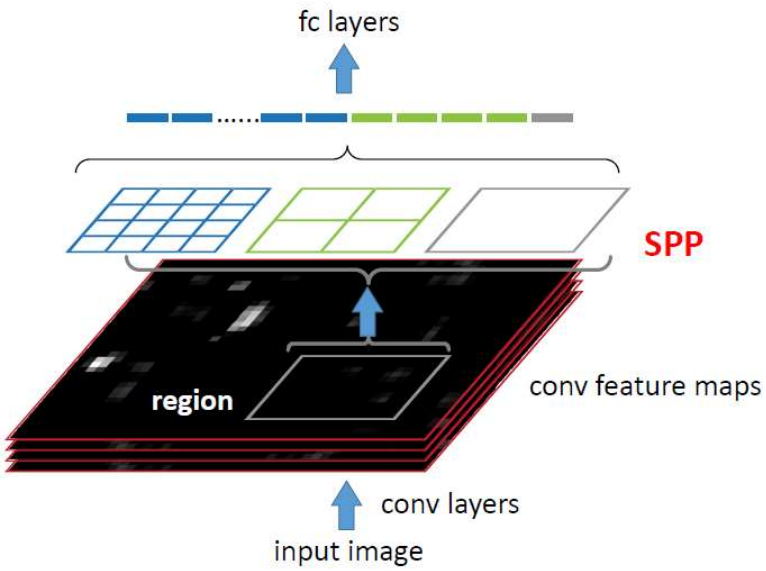
2nd Runner Up in Image Classification (ILSVRC 2014)

8.06% error rate is obtained. Unfortunately, VGGNet and GoogLeNet have better performance with the use of deep models. **Finally, SPPNet can only got 2nd runner-up in the classification task.**

• • •

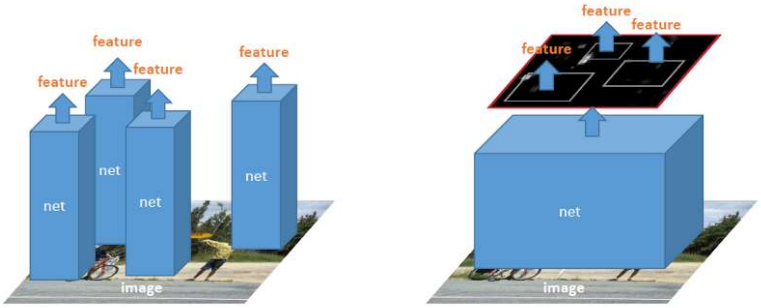
6. SPPNet in Object Detection

1. **Selective Search [6] is used to generate about 2k region proposals (bounding boxes) just like in R-CNN [7].**
2. **The input image goes through SPPNet using ZFNet by ONLY one time.**
3. **At the last conv layer, feature maps bounded by each region proposal is going into SPP layer then FC layer** as shown below:



SPPNet for Object Detection

Compared with R-CNN, **SPPNet processes the image at conv layers for only one time while R-CNN processes the image at conv layers for 2k times** since there are 2k region proposal. The image below illustrates the idea:



R-CNN (Left) and SPPet (Right)

After the FC layer for each bounding box, SVM and bounding box regressor are also needed, which is not an end-to-end learning architecture.

. . .

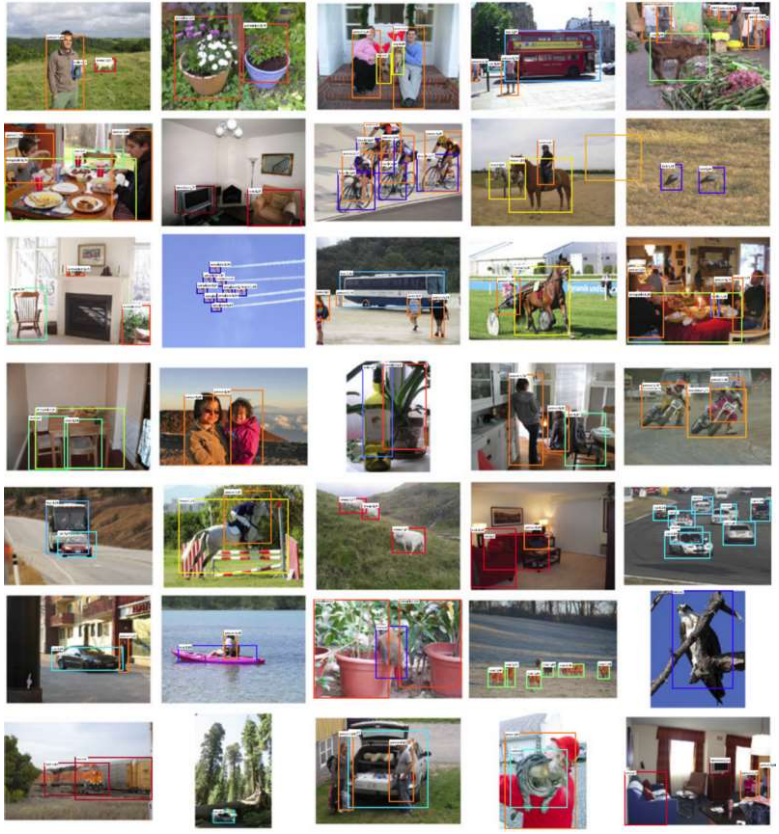
7. Comparison with State-of-the-art Approaches (Detection)

7.1 VOC 2007

	SPP-net 1-scale	SPP-net 5-scale	RCNN
mAP	58.0	**59.2**	58.5
GPU time / img	**0.14s**	0.38s	9s
speed-up	**64x**	**24x**	-

VOC 2007

VOC 2007 Results



Some Amazing Results in VOC 2017

In VOC 2007 as shown above, **compared with R-CNN, SPPNet with 5 scales obtained higher mAP of 59.2%.**

7.2 ILSVRC 2014

	mAP
NUS	**37.2**
ours, multi SPP-nets	**35.1**
UvA	32.0
ours, 1 SPP-net	31.8
Southeast-CASIA	30.4
1-HKUST	28.8
CASIA_CRIPAC_2	28.6

SPPNet got 1st Runner-Up in ILSVRC 2014 Object Detection

In ILSVRC 2014, SPPNet obtains 35.1% mAP and got 1st runner-up in object detection task.

Actually, Microsoft has contributed many state-of-the-art deep learning approaches in ILSVRC afterwards such as PReLUNet and ResNet. I will review them later on!

Of course, other networks would also be reviewed, please stay tuned!!!!

. . .

References

1. [2014 ECCV] [SPPNet]
   Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
2. [2015 TPAMI] [SPPNet]
   Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
3. [2014 ECCV] [ZFNet]
   Visualizing and Understanding Convolutional Networks
4. [2012 NIPS] [AlexNet]
   ImageNet Classification with Deep Convolutional Neural Networks
5. [2014 ICLR] [OverFeat]
   OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks
6. [2013 IJCV] [Selective Search]
   Selective Search for Object Recognition
7. [2014 CVPR] [R-CNN]
   Rich feature hierarchies for accurate object detection and semantic segmentation

My Reviews

1. Review of ZFNet—Winner of ILSVRC 2013 (Image Classification)
2. Review: AlexNet, CaffeNet—Winner of ILSVRC 2012 (Image Classification)
3. Review: OverFeat—Winner of ILSVRC 2013 Localization Task (Object Detection)
4. Review: R-CNN (Object Detection)