Object-Scene Convolutional Neural Networks for Event Recognition in Images

Limin Wang[1,2], Zhe Wang[2], Wenbin Du[2], Yu Qiao[2]

[1]The Chinese University of Hong Kong, Hong Kong
[2]Shenzhen Institutes of Advanced Technology, CAS, China

June 12, 2015

Outline

Outline

Introduction

Figure: Examples of cultural event recognition dataset.

- Event and action recognition has been widely studied in videos, but received less attention in images.
- Event is a complex concept and relevant to many other factors, including objects, human poses, human garments and scene categories.

Motivation

- Object, scene, and event are three highly related concepts in high-level computer vision research.
- Deep learning has turned out to be very effective in the task of object and scene recognition.
- For the tasks of object and recognition, there are very large-scale datasets: ImageNet and Places. But for event recognition, the dataset is relatively small.
- **As event is highly relevant with object and scene, transferring effective representations learned for object and scene recognition will be a reasonable choice.**


J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. *ImageNet: A large-scale hierarchical image database*, CVPR, 2009.


B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *Learning deep features for scene recognition using places database*, NIPS, 2014.

Outline

Overview

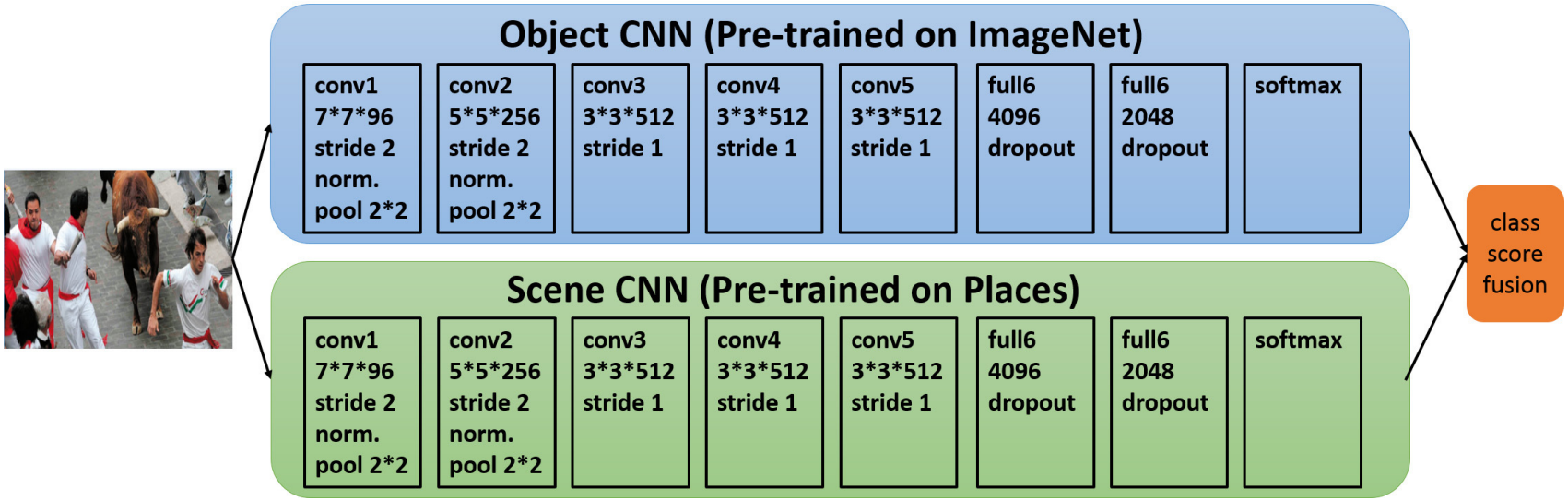


Figure: The architecture of Object-Scene Convolutional Neural Network (OS-CNN) for event recognition.

OS-CNN

Object-Scene Convolutional Neural Networks are composed of two nets

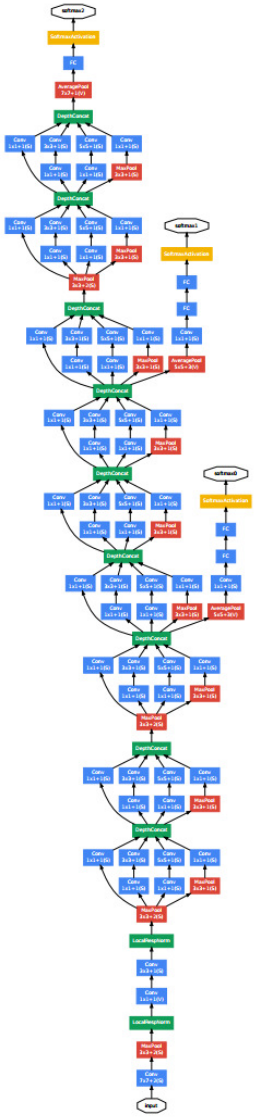
- Object nets: capturing useful information of objects to help event recognition.
- We build object nets based on recent advances on object recognition and pre-train it on the ImageNet dataset.
- Scene nets: extracting scene information to assist event recognition.
- We construct scene nets with the help of recent works on scene recognition and pre-train it on the Places dataset.

Based on previous analysis, event is highly relevant with object and scene. Thus, we combine the recognition scores of both object and scene nets:

$$s(\mathbf{I}) = \alpha_o s_o(\mathbf{I}) + \alpha_s s_s(\mathbf{I}).$$

Network architectures

Arch.	conv1	conv2	conv3	conv4	conv5	full6	full7	full8
CNN-F	64x11x11 st. 4, pad 0 LRN, x2 pool	256x5x5 st. 1, pad 2 LRN, x2 pool	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 x2 pool	4096 drop-out	4096 drop-out	1000 soft-max
CNN-M	96x7x7 st. 2, pad 0 LRN, x2 pool	256x5x5 st. 2, pad 1 LRN, x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x2 pool	4096 drop-out	4096 drop-out	1000 soft-max
CNN-S	96x7x7 st. 2, pad 0 LRN, x3 pool	256x5x5 st. 1, pad 1 x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x3 pool	4096 drop-out	4096 drop-out	1000 soft-max



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Ensemble of multiple CNNs

- In past several years, some successful architectures have been proposed for object recognition:
  - deep CNN, including: AlexNet, ClarifaiNet, OverFeat.
  - very-deep CNN, including: GoogLeNet, VGGNet.
- **Deeper networks obtain higher recognition performance on the dataset of ImageNet**.
- We investigate different network architectures on the task of event recognition and also explore their complementarity by fusing them:

$$s_x(\mathbf{I}) = \beta^{\mathrm{d}} s_x^d(\mathbf{I}) + \beta^{\mathrm{v-d}} s_x^{v-d}(\mathbf{I}).$$

Outline

Experimental setup

- Cultural event recognition task provides a dataset composed of 50 classes, whose images are collected from the Internet.
- The dataset is divided into three parts: development data (5,875 images), validation data (2,332 images), and evaluation data (3,569 images).
- During develop phase, we train our model on the development data and verify the performance of different settings for method on the validation data.
- For final evaluation, we merge the development and validation data into a training data and re-train our model. The settings of our method are the same with the one during develop phase.

Implementation details

- We first pre-train our model on the large datasets: ImageNet and Places, and then fine tune the network weights on the cultural event recognition dataset.
- During training phase, all images are resized to 256×256 and a 224×224 sub-region is cropped and flipped randomly.
- To overcome the problem of overfitting, we set learning rate of hidden layers as 10^{-2} times of final layer.
- The learning rate is initially set to 10^{-2}, decreasing to 10^{-3} after 1.4K, then to 10^{-4} after 2.8K iteration, and training stopped at 4.2K.
- During testing phase, we use a multi-view voting method to classify each image, where we obtain 10 inputs by cropping and flipping four corners and the center of images, and average the recognition scores of these 10 inputs.

Effectiveness of OS-CNN

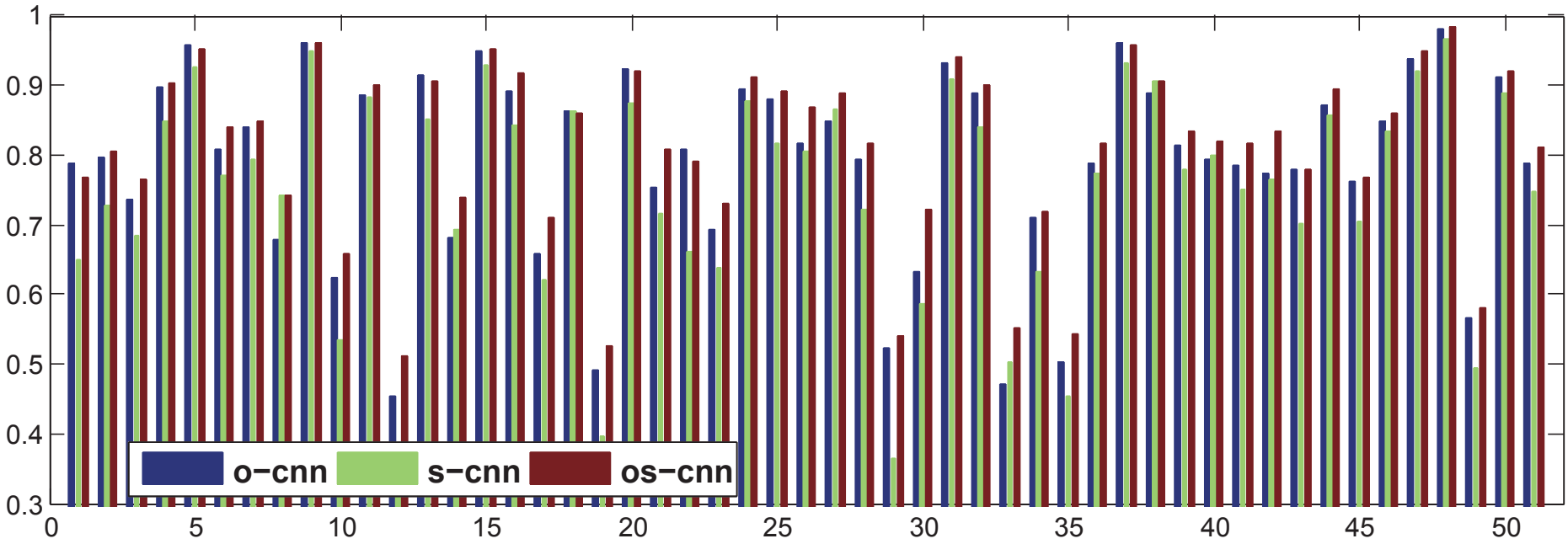


Figure: Comparison of the performance of object nets, scene nets, and OS-CNN

Evaluation of different architectures

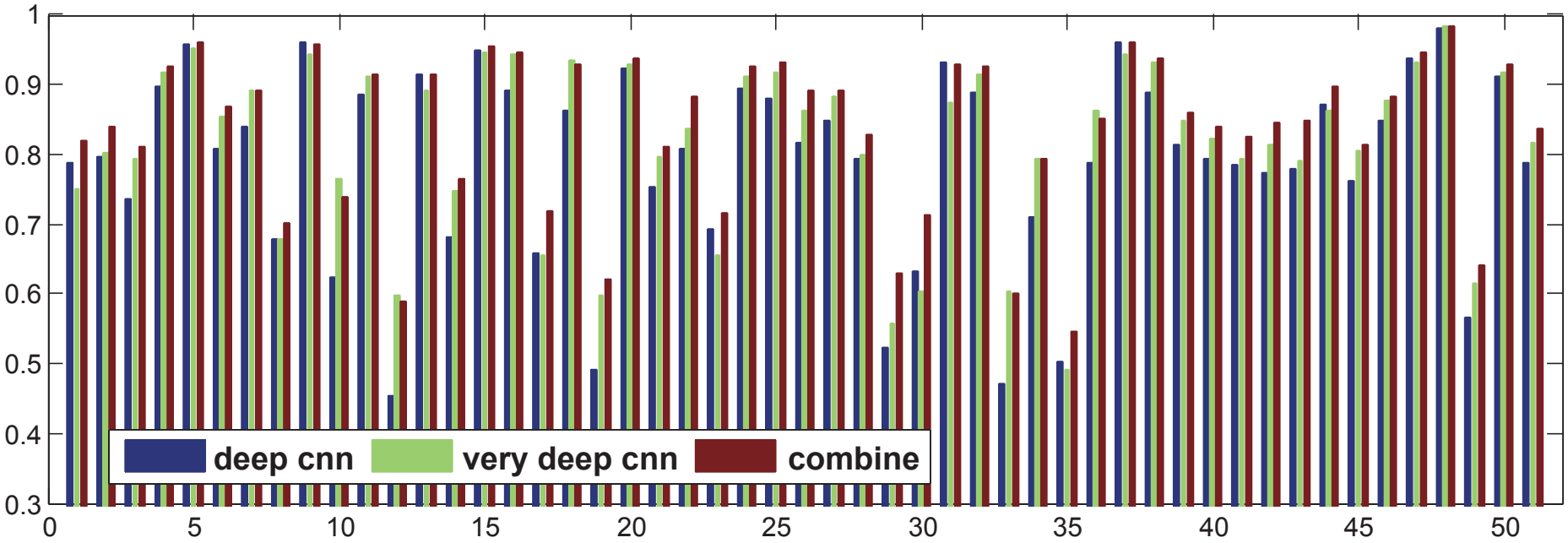


Figure: Results of **object nets** using different architectures.

Evaluation of different architectures (cont'd)

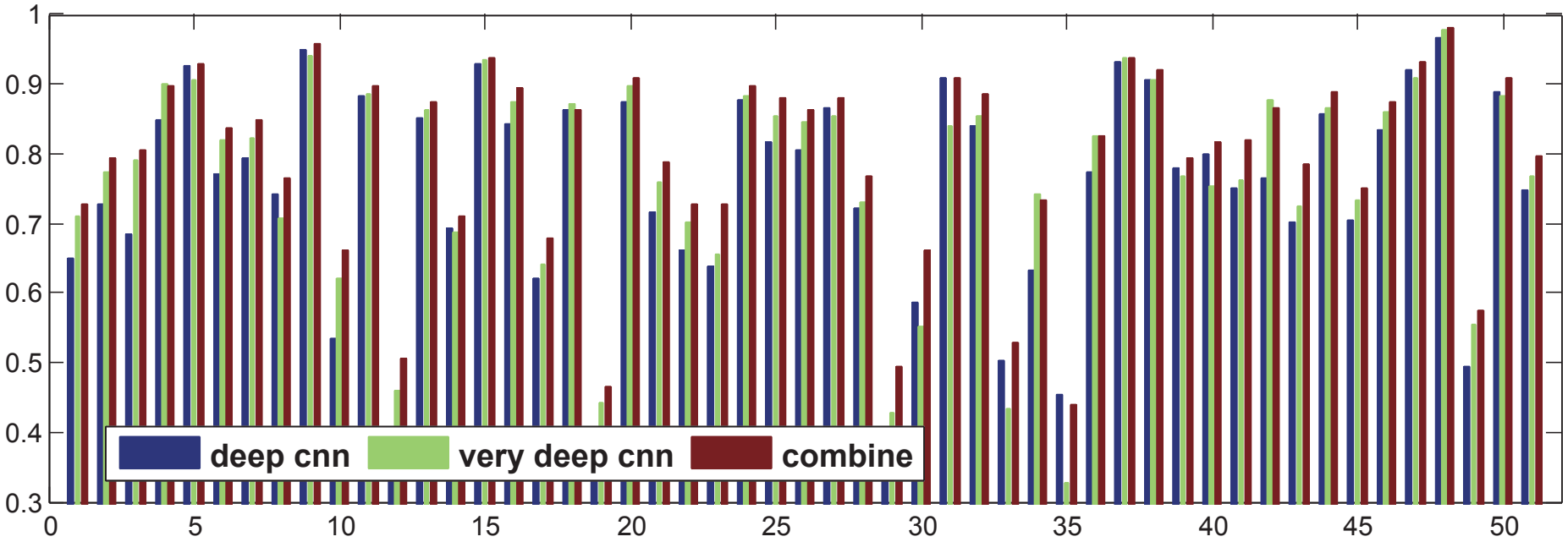


Figure: Results of **scene nets** using different architectures.

Challenge solution and result

- Based on previous numerical evaluation, we conclude that (i) object ne is better than scene net. (ii) very-deep architecture outperforms deep architecture.
- Hence, we introduce another object net with very-deep architecture into our OS-CNN framework.
- Challenge solution: five-stream CNNs, namely object net: ClarifaiNet, GoogLeNet, VGGNet, and scene net: AlexNet, GoogLeNet.

Rank	Team	Score
1	**MMLAB (Ours)**	85.5%
2	UPC-STP	76.7%
3	MIPAL_SNU	73.5%
4	SBU_CS	61.0%
5	MasterBlaster	58.2%
6	Nyx	31.9%

Table: Comparison the performance.

Outline

Conclusions

- We have presented a new architecture for event recognition, called *object-scene convolutional neural networks* (OS-CNN), by capturing effective information from the perspectives of object and scene.
- From our experimental results, object nets outperform scene nets on event recognition, and the combination of them further improve performance.
- We also investigated different network architectures for the design of OS-CNN, from deep networks such as AlexNet, to very-deep networks such as GoogLeNet.
- From our results, deeper architectures achieve better performance and the fusion of different network architectures assists to boost recognition performance.

Thank you!

Email: wlm011@ie.cuhk.edu.hk