面向大规模场景分类的深度学习模型

Deep Learning Models for Large Scale Scene Classification

乔宇

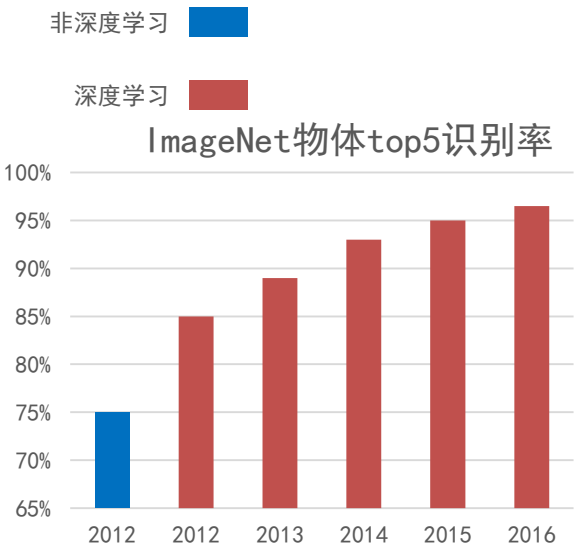
中国科学院-深圳先进技术研究院

2017-April-21

深度学习方法推动视觉技术快速进展

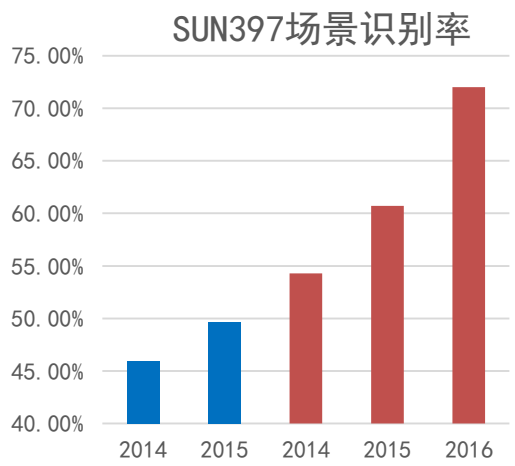
深度神经网络已被成功用于物体识别、场景分类、行为识别等视觉核心任务，极大地推动了计算机视觉技术的发展

视觉三大问题



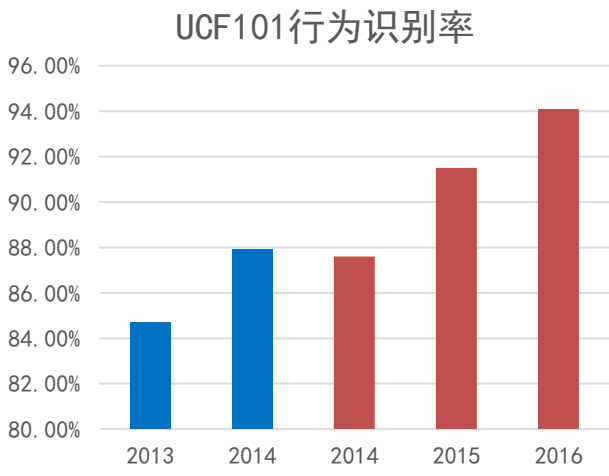
物体识别

“有什么”



场景分类

“在哪里”



行为识别

“干什么”

Classification of scenes is different that of object

Bedroom | Bridge | Church Outdoor | Classroom | Conference Room | Dining Room | Kitchen | Living Room | Restaurant | Tower

Yu QIAO

Scene Datasets

数据库	Scene15	MIT Indoor67	SUN397	Places205	Places401	LSUN	Places365
场景类别	15	67	397	205	401	10	365
数据量	0.5万	1.5万	10.8万	250万	800万	1000万	800万
创建时间	2006	2009	2010	2014	2015	2015	2016
准确率	92.3%	86.7%	73.0%	90.7%	83.1%	91.6%	90.9%

*Place数据库为top 5的正确率

Place2 Scene Dataset in ImageNet'16

http://places.csail.mit.edu/

- Places2 -2016 scene recognition challenge:
  - 365 scene categories, each class containing from 4,000 to 40,000 images.
  - 8M images for training, 36k images for validation and 328k images for testing.
  - Dataset size is much bigger than ImageNet object dataset.

- Scene recognition is challenging:
  - The concept of scene is more subjective and high level than object.
  - Larger intra-class variations (visual inconsistency).
  - Smaller inter-class variations (label ambiguity).

B. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva , *Places2: A Large-Scale Database for Scene Understanding*, in *Arxiv, 2015.*

Challenge 1- Visual Inconsistency

Kitchen

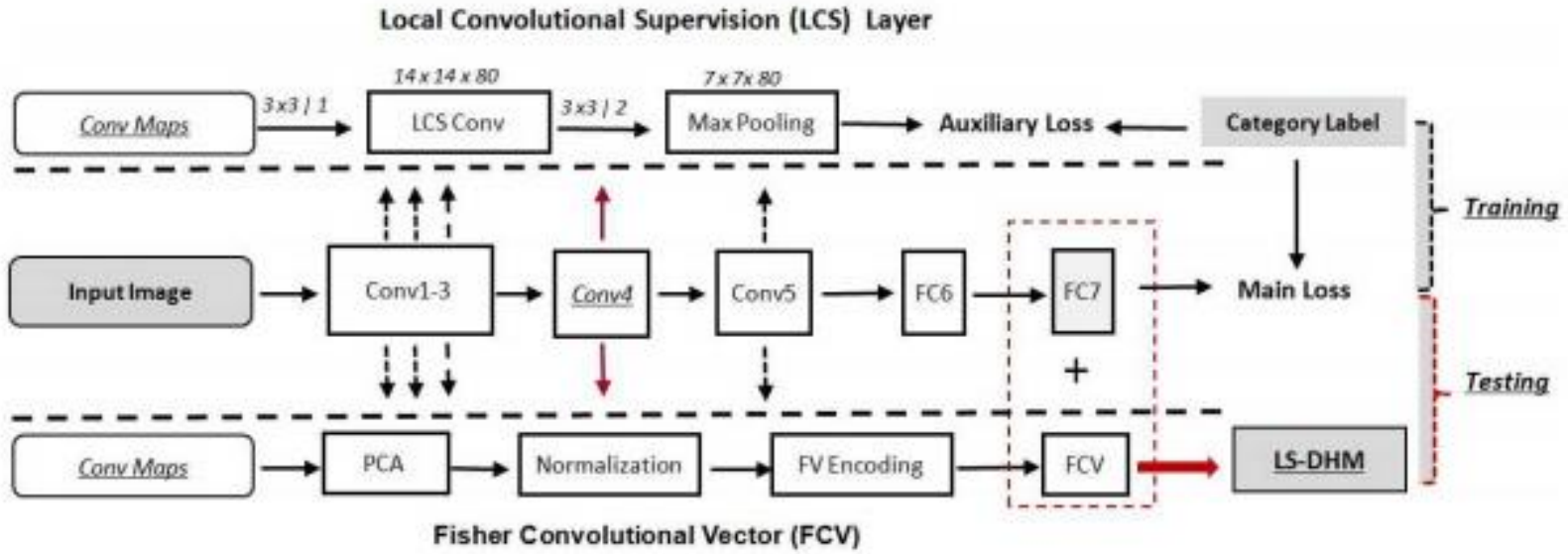
Coffe Shop

Chanllenge2-Label Ambiguity Example

cubicle office

office cubicles

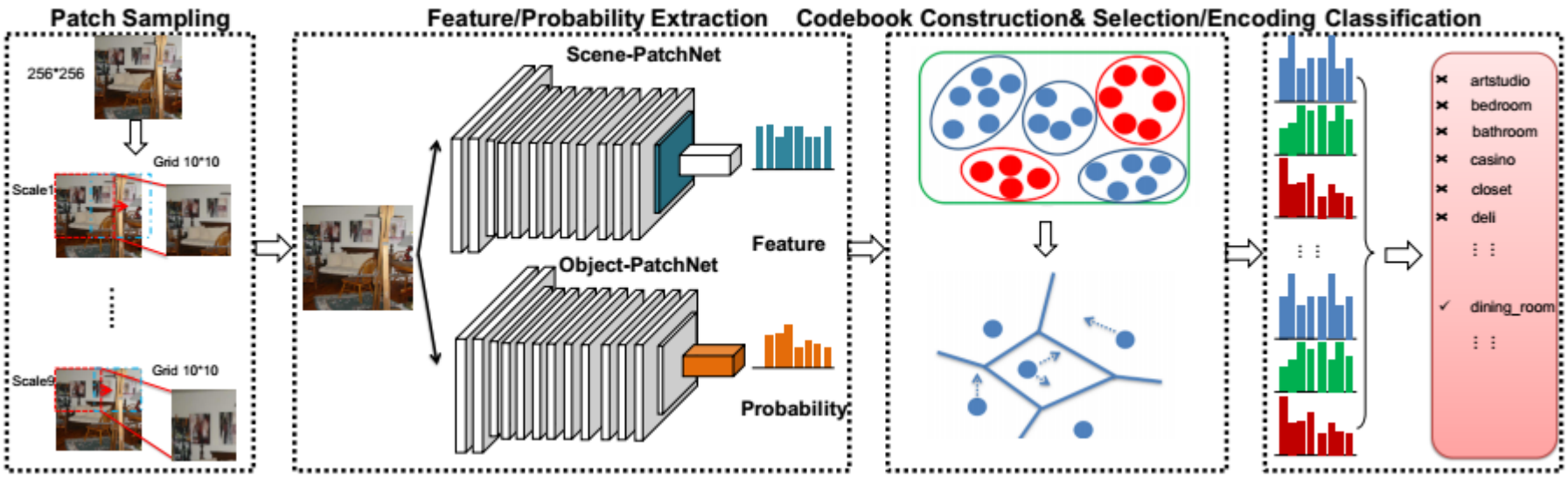
Approach 1: Locally-Supervised Deep Hybrid Model



- Convolutional features are important and complementary with fully connected features for scene classification.
- Local supervision help to enhance the discriminative ability of convolutional features.

•Sheng Guo, Weilin Huang, Limin Wang, and Yu Qiao, "Locally Supervised Deep Hybrid Model for Scene Recognition, " Vol. 26, No.2, 808 – 820, IEEE Transactions on Image Processing (**T-IP**), 2017

Approach 2: Aggregating Local Patches

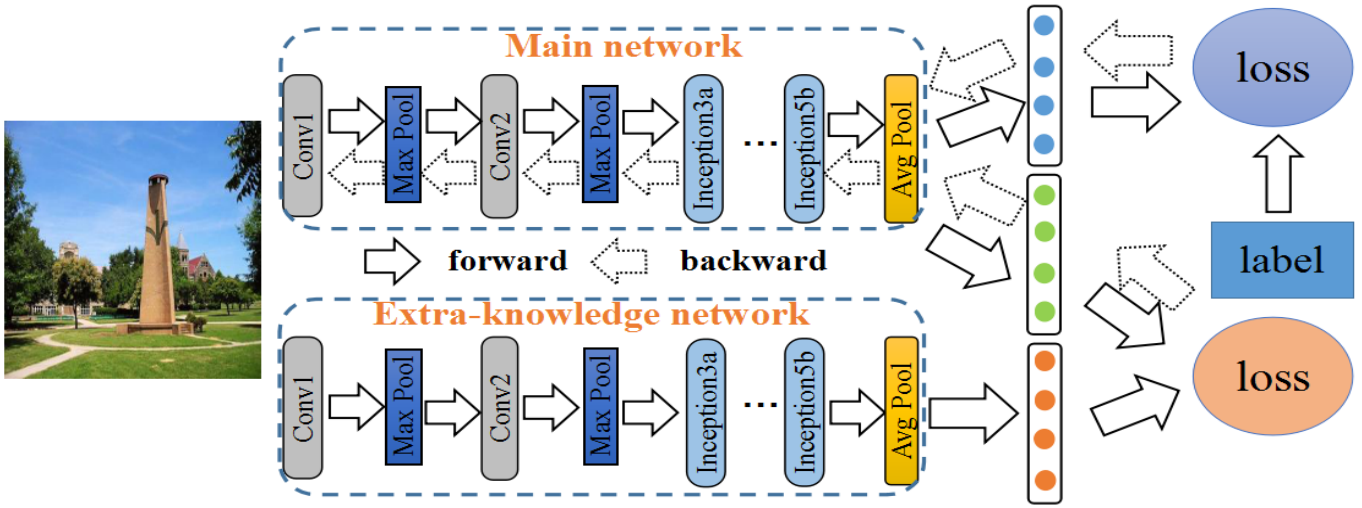


Patch and object yield important cues for scene classification.

•Zhe Wang, Limin Wang, Yali Wang, Bowen Zhang, and Yu Qiao, "Weakly Supervised PatchNets: Describing and Aggregating Local Patches for Scene Recognition, " *IEEE Transactions on Image Processing* (**T-IP**), Vol. 26, No.4, pp. 2028 – 2041, 2017

Approach 3: Knowledge Guided Disambiguation

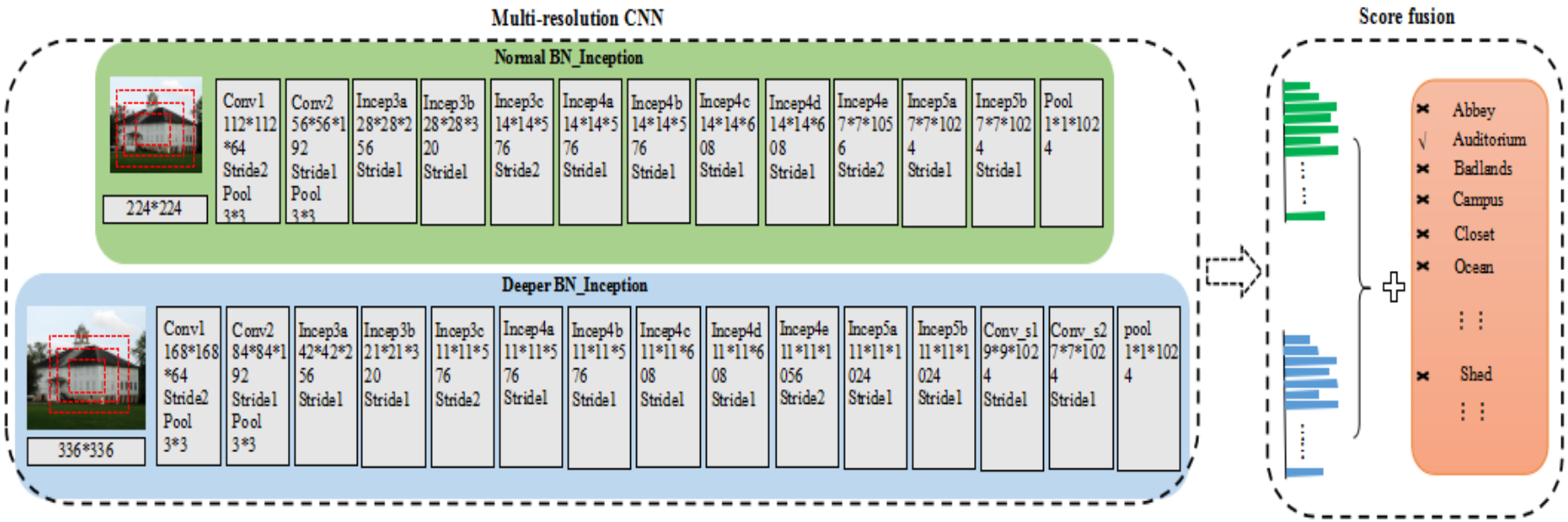
Knowledge from networks trained on other datasets help to relieve scene category disambiguation



- In previous scenario, all the images belonging to the same super category are constrained to have the same label, without considering the difference between images.
- We propose to automatically assign a soft code to each image, which is able to better encode the visual information of natural images.
- In the soft code space, the images from easily confused categories are equipped with similar codes.
- Finally, we design a multi-task framework to predict both hard code and soft code

•Limin Wang, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao, " Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNs, " *IEEE Transactions on Image Processing* (**T-IP**), Vol. 26, No.4, pp. 2055 – 2068, 2017

Appraoch 3:Multi-Resolution CNNs



Implementation details

- **Architectures:**
  - Low resolution: image (256*256), crop(224*224), inception2 network [2]
  - High resolution, image (384*384), crop(336*336), inception2+2 convs
- **Knowledge networks:**
  - Object nets: inception2 trained with ImageNet
  - Scene nets: inception2 trained with Places205
  - Currently, knowledge disambiguation only for low resolution CNNs
- **Implementation details:**
  - Resample images to balance the class distribution
  - Data augmentation: fixed crop, scale jittering, horizontal flipping [1,6]
  - Toolbox: we use a multi-GPU extension of Caffe, which is publicly available: https://github.com/yjxiong/caffe.git

L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, Towards Good Practices for Very Deep Two-Stream ConvNets, in *Arxiv, 2015.*

Results on MITIndoor67 and SUN397

Model	MIT Indoor67	SUN397
ImageNet-VGGNet-16 [38]	67.7%	51.7%
Places205-AlexNet [22]	68.2%	54.3%
Places205-GoogLeNet [10]	74.0%	58.8%
Places205-CNDS-8 [54]	76.1%	60.7%
Places205-VGGNet-16 [55]	81.2%	66.9%
Places365-VGGNet-16 [23]	76.5%	63.2%
Hybrid1365-VGGNet-16 [23]	77.6%	61.7%
DAG-VGGNet19 [7]	77.5%	56.2%
MS-DSP [56]	78.3%	59.8%
LS-DHM [10]	83.8%	67.6%
VSAD [11]	84.9%	71.7%
Multiple Models [57]	86.0%	70.7%
Three [58]	86.0%	70.2%
Places365-Deeper-BN-Inception (B2)	84.8%	71.7%
Places401-Deeper-BN-Inception (B2)	**86.7%**	**72.0%**

Results on Place2

Method	Imagenet(top1/top5)	Places(top1/top5)	Places2(top1/top5)
AlexNet	40.7%/18.2%	50.0%/-	57.0%/-
VGGNet	27.0%/8.8%	39.4%/11.5%	52.4%/-
Normal BN-Inception	24.7%/7.2%	38.1%/11.3%	48.8%/17.4%
Deeper BN-Inception	23.7%/6.6%	37.8%/10.7%	48.0%/16.7%
Multi-resolution CNN	**21.8%/6.0%**	**36.4%/10.4%**	**47.4%/16.3%**

Performance of Multi-Resolution CNNs on the validation data from the datasets of ImageNet, Places and Places2.

LSUN Challenge

- Winner of LSUN Challenge in CVPR 2016 :
  - 10M images, 10 category

Rank	Team	Year	Top1 Accuracy
1	SIAT_MMLAB	2016	**91.6%**
2	SJTU-ReadSense	2016	90.4%
3	TEG Rangers	2016	88.7%
4	ds-cube	2016	83.0%
1	Google	2015	91.2%

ImageNet Scene Classification Task

- Ranking 2nd in ImageNet 2015 Scene Classification Task

Rank	Team	Top1
1	WM	16.9%
2	SIAT_MMLAB(our)	17.4%
3	Qualcomm	17.6%
4	Trimps-Soushen	18.0%
5	NTU_Rose	19.3%

Conclusions

- Large scale scene datasets with many categories come along with increased ambiguity between the class labels (e.g. baseball field vs. stadium baseball).
  - Knowledge guided disambiguation aims to regularize CNN training with extra knowledge and improve the generalization capacity.

- Scene or Places, defined by containing objects, spatial layout, human events, and global contexts, are more high-level concepts.
  - Object and semantic regions yield useful cues for scene classification together with global layout.
  - Multi-Resolution CNNs take images of different sizes as input and capture visual information from different levels.

模型和代码公开

场景理解与分类

- MR-CNNs (2nd in scene classification task ImageNet 2016, 1st in LSUN 2016)
- Weakly Supervised PatchNets (Top performance in MIT Indoor67 and SUN397)

行为识别和检测

- Temporal Segment Networks (NO1 in ActivityNet 2016)
- MV-CNNS(Speed:300帧/s)
- Trajectory-Pooled Deep-Convolutional Descriptors （Top performance in UCF101 and HMDB51)

人脸检测与识别

- MJ-CNN face detection (top performance in FDDB & WIDE)
- HFA-CNN face recognition (single model 99% in LFW)

场景文字检测与识别

- Connectionist Text Proposal Network for Scene Text Detection (Top performance in ICDAR)

下载地址

http://mmlab.siat.ac.cn/yuqiao/Codes.html

Thank you!

Q&A