Scene recognition with CNNs: objects, scales and dataset bias
Luis Herranz, Shuqiang Jiang, Xiangyang Li
Institute of Computer Technology, Chinese Academy of Sciences, Beijing, China
{luis.herranz,shuqiang.jiang,xiangyang.li}@vipl.ict.ac.cn

Motivation
- Scaling (of patches) changes the data distribution.
- This induces a scale-related bias if the CNN model is fixed.
- However, this bias is ignored in most works.

Contributions
- Study the scale-induced bias
- Multi-scale architecture using scale-specific CNNs.

Previous works
1. Scene recognition:
   - Holistic scene-CNNs[1]. Trained on Places-CNN. *Only global scale(s), object-like scales are ignored.
   - Multi-scale local CNN pooling[2,3,4]. ImageNet-CNNs on multiple scales and aggregated using pooling (e.g. VLAD, FV).
   - The CNN model is fixed black box (scale is ignored).

2. Combining object data and scene data in a CNN:
   - Hybrid-CNN[1]. Trained with all ImageNet and Places data. *Object and scene images are rescaled equally (scale is ignored).

Introduction
We address two problems:
1. Effectively implement multiscale CNN architectures for scene recognition.
2. Effectively combine Places and ImageNet

The distributions of objects in object datasets and scene datasets are very different
- Scale is one of the main factors

Effects of scaling:
- Changes the distribution of visual features
- Content in patches shifts from scenes to objects

Looking at the number of objects in the scale

Scale-induced bias

How to correct scale-induced bias?
Scale-specific CNNs (instead of a fixed one) adapted to the patches at each scale. We study two ways:
- Switch Places-CNNs/ImageNet-CNNs, for global/local scales, respectively.
- Fine tune with patches extracted at the target scale

Multi-scale architecture with scale-specific CNNs

MIT Indoor 67 (dual architecture)
(dual architecture, fine tuned)

SUN397 (dual architecture) (spliced architecture)

References

Experimental results

Architecture
Data
MIT Indoor 67 SUN 397
Alex VGG Alex VGG
Best single
IN 66.64 76.42 32.42 79.71
PL 77.78 80.90 58.88 66.23
Dual
IN 71.87 79.04 56.62 84.07
Dual hybrid INPL 78.28 83.59 64.10 80.20
Three
INPL 78.28 86.04 63.05 70.17
Full (7 scales) IN 74.35 70.22 58.71 55.18
PL 78.21 77.81 63.81 58.00
Full hybrid (spliced) INPL 80.97 80.22 65.38 63.19
Double full hybrid INPL 80.97 80.22 66.28 62.01
Human (good)[4] - - 68.5% -
Human (expert)[5] - - 70.6% -

Scene recognition
Object recognition
Scaled versions

References