SINGLE-SCALE SCENE RECOGNITION WITH CNNS: OBJECTS, SCALES AND DATASET BIAS

Sagnik Majumder 2014A8PS464P
B.E. (Hons.) Electronics and Instrumentation
f2014464@pilani.bits-pilani.ac.in
BITS Pilani, Pilani Campus

April 16, 2017

Abstract

In modern scene recognition, the conventional pipeline consists of three stages: train network on small scale images (scenes), train same network on large scale images (objects), classify or recognize the scenes and identifying the constituting objects using the same network. We revisit both the training step and testing step by employing scale-specific networks for both raining and testing. The scale specific networks, ensemble of them or a hybrid architecture, are finally used for testing. Networks, trained on small scale data (scenes), are used for identifying scenes while those, trained on large scale data (objects), are used for extracting object content information in the images. A simple AlexNet ([?]) CNN has been used in the ensemble (large scale network + small scale network). The small scale version has been trained on PLACES205 ([9]) database while the large scale version has been trained on ImageNet (ILSVRC 2012) ([?]) database. Our method reaches an accuracy of 85.00% on the SUN397 ([?]) dataset, increasing the accuracy of the original Herranz et more than 20%, closely approaching human-level performance.

1 Introduction

Objects are integral parts of scenes and a comprehensive understanding of both scenes and objects is very important and relevant for recognition of scenes. CNNs have brought about a revolution in the world of computer vision and, speech and text recognition, due to its high performance statistics and tolerance again distortion, rotation and translation of input vectors.They have shown their promise as a universal representation for recognition. However they have some limitations. CNNs, which are deep structures having multiple layers, work using different levels of abstraction in different layers. Each layer is trained to extract different classes of features from the input in a hierarchical manner. Yet, they fail to recognize the same object when they are fed to the network at different scales. For example a network may recognize a wheel in an image when the image covers the maximum portion of the image i.e. the image has minimum

background clutter or background noise, but it may fail to recognize the same wheel with the car body in background. This is solely due to the fact that CNNs extract different features from the same object when they have different scales. Thus, training the CNN on one dataset which has images of objects of one scale and testing the same network on images with objects of a different scale induces a dataset bias. Thus they lack geometric invariance, which limits their robustness for classification and matching of highly variable scenes. They fail to do the task of recognition for same objects when they are at different scales.This work addresses the following two problems: 1) scale induced dataset bias in multi-scale convolutional neural network (CNN) architectures, and 2) how to combine effectively scene-centric and object-centric knowledge (i.e. Places and ImageNet) in CNNs.

An earlier attempt, Hybrid-CNN, showed that incorporating ImageNet did not help much. Here we propose an alternative method of using a scale-adaptive architecture, which takes the scale of the image into account, resulting in significant recognition gains. The response of ImageNet-CNNs ([**?**]) and Places-CNNs ([9]) at different scales shows that both the networks operate in different scale ranges, so using the same network for all the scales induces dataset bias resulting in limited performance. Thus, adapting the feature extractor to each particular scale (i.e. scale-specific CNNs) is crucial to improve recognition, since the objects in the scenes have their specific range of scales. Experimental results show that the recognition accuracy highly depends on the scale, and that simple yet carefully chosen multi-scale combinations of ImageNet-CNNs and Places-CNNs, can push the recognition accuracy (testing done using 227 X 227 scale only) in SUN397 ([**?**]) up to 85% and more if done using deeper architecture.

Several other algorithms have been developed worldwide to solve these problems. This report talks theoretical and implementation details of the novel algorithm described in the publication 'Scene recognition with CNNs: objects, scales and dataset bias' .

2 Related State of the Art Work

To improve the invariance of CNN activations without degrading their discriminative power, Y. Gong et al. in their publication [2] presents a simple but effective scheme called multi- scale orderless pooling (MOP-CNN). This scheme extracts CNN activations for local patches at multiple scale levels, performs orderless Vectors of Locally Aggregated Descriptors (VLAD), which are a simplified version of Fisher Vectors (FV) pooling, of these activations at each level separately, and concatenates the result. This particular representation has three scale levels, corresponding to CNN activations of the global 256 X 256 image and 128 X 128 and 64 X 64 patches, respectively.The pooling of the activations of these multiple patches, done to summarize the second and third levels by single feature vectors of reasonable dimensionality, is also followed by dimensionality reduction through Principal Component Analysis (PCA). The resulting MOP-CNN representation is used as a generic feature for either supervised or unsupervised recognition tasks, from image classification to instance-level retrieval. Inspired by Spatial Pyramid Matching , which extracts local patches at a single scale but then pools them over regions of increasing scale, ending with the whole image, this paper proposes a kind of reverse SPM idea, where

patches at multiple scales are extracted, starting with the whole image, and then pooling each scale without regard to spatial information.

Additionally, in [9], Zhou et al. trained a Hybrid-CNN, by combining the training set of Places-CNN ([9]) and training set of ImageNet-CNN ([**?**]). They removed the overlapping scene categories from the training set of ImageNet ([**?**]), and then the training set of Hybrid-CNN has 3.5 million images from 1183 categories. Hybrid-CNN is trained over 700,000 iterations, under the same network architecture of Places-CNN ([9]) and ImageNet-CNN ([**?**]). The accuracy on the validation set was 52.3%. They evaluated the deep feature (FC 7) from Hybrid-CNN on some popular benchmarks. Combining the two datasets yields a little increase in performance for a few.

The publication [3] by Herranz et al. addresses two main problems with CNNs:

1. scale-induced dataset bias which might be introduced in multi-scale CNNs, which are trained using database having images of a certain scale and tested on images having a different scale

2. combining the object-centric and scene-centric data of 2 different CNNs (Places and ImageNet) into one single architecture.

It does so by using different specific networks for different scales of the same object in the image with the aim to build a scale-tolerant architecture that can recognize objects both in the class of images which portray an object only and the class of images which portray an object in the middle of a scene.

Two different architectures:

1. ImageNet (ImageNet CNNs)([**?**]) which are trained using ImageNet ([**?**]) database (ILSVRC 2012 in particular) which contains images with the object only and

2. Places (Places CNNs) ([9]) which is trained using Places database which contain images with small-sized objects with a background.

The publication [3] is finally able to build a scale-invariant model of scale-specific networks by

1. using a hybrid of ImageNet ([**?**]) and Places networks ([9])

2. fine-tuning the networks

Summary of the Results of the Algorithms on Popular Datasets

2.1 Gong et al.

Given below is a brief summary of the results of the algorithm on popular and challenging datasets:

1. level1 + level2 + level3 (MOP-CNN) with 12,288 features achieves 51.98% accuracy on SUN397 ([**?**]) dataset

2. level1 + level2 + level3 (MOP-CNN) with 12,288 features achieves 68.88% accuracy on MIT 67 Indoors ([**?**]) dataset

3. level1 + level2 + level3 (MOP-CNN) with 12,288 features achieves 57.93% accuracy on ILSVRC 2012/2013 ([**?**]).

Clearly, it can be concluded that the algorithm, despite showing promising performance on a few less significant datasets, fails to perform well for the challenging and diverse databases like SUN397 ([**?**]) and ILSVRC 2012/2013 ([**?**]).

2.2 Zhou et al.

Given below is a brief summary of the results of the algorithm on some datasets:

1. SUN397 ([**?**])- 53.860.21%
2. MIT Indoor 67 ([**?**])- 70.80%
3. Scene15 ([**?**]) - 91.590.48%
4. SUN Attribute ([**?**])- 91.56%
5. Action40 - 55.280.64%

Here, it is seen that though the algorithm performs well for the simple datasets (marked in boldface), fails miserably in the case of complex datasets like SUN397.

2.3 Herranz et al.

It's methods finally led to the state-of-the-art scene recognition.

Other notable works, which were refered to, are [8], [1], [9], [5], [6], [7].

3 Proposed Approach

3.1 Architecture Models

AlexNet([**?**])(shown in Fig. 1)

3.2 Libraries and Toolkits

LabelMe([**?**]) and Caffe([4]) by Berkley Vision and Learning Center

3.3 Description of the work

Training and Testing Pipeline

Training of the small scale network was done on the PLACES205 ([9]) dataset andthat of the large scale network was done on IMAGENET (ILSVRC 2012) ([**?**]) dataset.

Testing was done on SUN397 ([**?**]) dataset.

Changes

The changes that have been introduced in addition to the implementation, mentioned in the publication [3] are as follows:

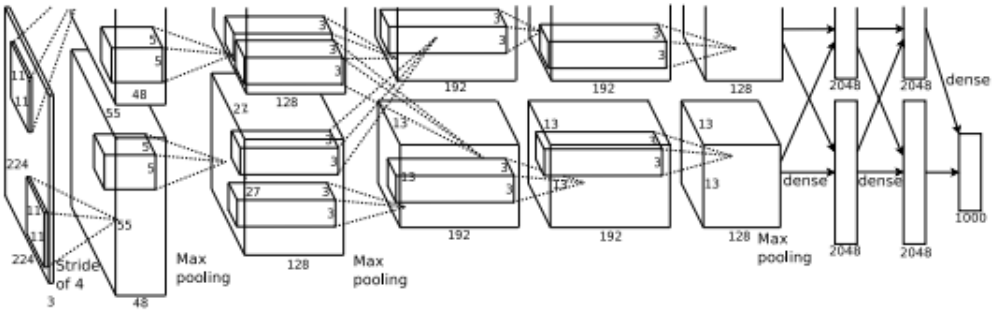

Figure 1: **An image of a generic CNN showing the Convolutions and Pooling operations ([?])**

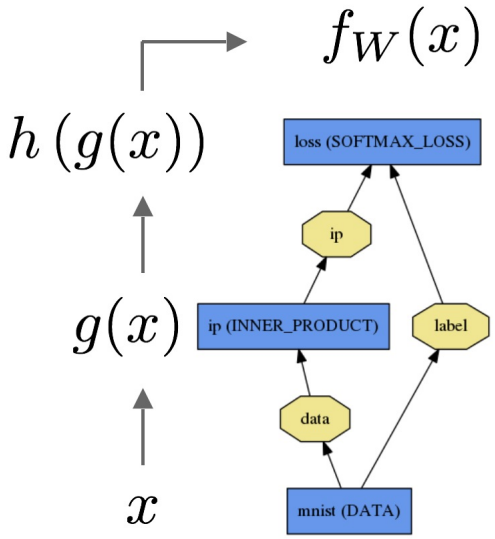

Figure 2: **An image showing forward pass ([4])**

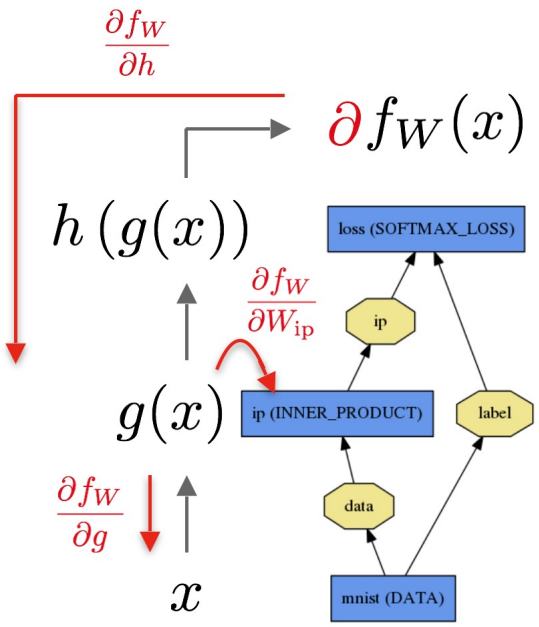

Figure 3: **An image showing backward pass ([4])**

1. An algorithm to calculate the the overall accuracy in recognizing the scenes has been implemented, which was not mentioned in the paper. After trial-running the algorithm, as mentioned in the paper for many times, it was observed that the PlacesCNN was very well tuned and adapted specifically to calculate the overall accuracy of the algorithm.

   The accuracy calculating algorithm has been implemented in python. The way it works is as follows:

   The algorithm uses the PlacesCNN that is well adapted to small scale images and classifies the images having a scale of 227 X 227. Classification is done by feeding the image through the PlacesCNN network, which in turn moves the image features down the network and it finally classifies the image as a member of one of the scene classes, as given in the file **'categoryIndex_places205.csv'** ([9]). The code compares the classification result, as produced by the network, with the classification labels, as given in the above-mentioned text file.

   This method turns out to be very fast and efficient in classifying the scene images as a member of one of the classes, which are mentioned in the **.csv** file. This also gives a testing accuracy of **85%**, when tested with **20 images** from the classses **'Airport Terminal'** and **'Alley'** of the **'SUN 2012'** data set.

2. The second change that was made to the original implementation was that while using the ImageNet CNN ([**?**]) to classify the small-scale images, a system was implemented so that the network classifies all the sub-crops

from the large-scale images and tries to recognize the objects that are part of the sub-crops. In this way, the network is able to recognize a lot of objects, that are part of the scene images, and is able to provide valuable information about the object content of the image, which again was one of the primary objectives of the paper ([3]).

The performance of the algorithm could not be evaluated and quantified properly in this respect because the images in the database are not annotated and labelled because of which a comparative Analysis can not be carried out. Another problem was the lack of a robust object detection system. It was manually observed that even the object classification performance of the ImageNet ([**?**]) network was quite high and was comparable with the state-of-the-art performance, as specified in literature.

4 Datasets

1. **IMAGENET (ILSVRC 2012) ([?])**

   ImageNet is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+). In ImageNet, we aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion, we hope ImageNet ([**?**]) will offer tens of millions of cleanly sorted images for most of the concepts in the WordNet hierarchy.

   Overall

   - Total number of non-empty synsets: 21841
   - Total number of images: 14,197,122
   - Number of images with bounding box annotations: 1,034,908
   - Number of synsets with SIFT features: 1000
   - Number of images with SIFT features: 1.2 million

2. **PLACES ([9])**

   Scene recognition is one of the hallmark tasks of computer vision, allowing defining a context for object recognition. MIT CSAIL introduced a new scene-centric database called Places, with 205 scene categories and 2.5 millions of images with a category label. Using convolutional neural network (CNN), they learned deep scene features for scene recognition tasks, and establish new state-of-the-art performances on scene-centric benchmarks. They provided the Places Database and the trained CNNs for academic research and education purposes. In this particular work, Places205 ([9]) has been used to train the Places CNN ([9]). Places205: MIT CSAIL release 2.5 million images from 205 scene categories to the public.

3. **SUN397 ([?])**

Class	Accuracy
Airport Terminal	86%
Alley	89%
Access_Road	80%

Table 1: **Table showing testing accuracy on SUN397 ([?]) database.**

Architecture	Pretraining Datasets	Scales	AlexNet ([?]) Accuracy on SUN397
Hybrid_CNN	IN_PL	1	53.86%
Implemented_Architecture	IN_PL	1	85%

Table 2: **Table showing comparison of 2 architectures on SUN397 ([?]) database.**

The goal of the SUN database project was to provide researchers in computer vision, human perception, cognition and neuroscience, machine learning and data mining, computer graphics and robotics, with a comprehensive collection of images covering a large variety of environmental scenes, places and the objects within. To build the core of the dataset, the researchers counted all the entries that corresponded to names of scenes, places and environments (any concrete noun which could reasonably complete the phrase I am in a place, or Lets go to the place), using WordNet English dictionary. Once we a vocabulary for scenes was established, the researchers collected images belonging to each scene category using online image search engines by querying for each scene category term. SUN397 ([**?**]) has a total of 397 scene categories.

5 Results and Evaluation Criteria

Testing was done on SUN397 ([**?**]) database. **20 Different images** from the class Airport Terminal were chosen. The images were chosen randomly. They were rescaled and cropped to different sizes, as mentioned in the literature. The crops and rescaled images were then fed into the respective networks: ImageNet CNN ([**?**]) and Places CNN ([9]). The **testing accuracy** was found to hit a value of **86%**. Multiple test sets, each with 20 different images but from a different class each time, were set up. Each time a different accuracy was obtained. The **'alley'** class, when tested, gave an accuracy of **89%**, while the **'Access_Road'** class, when tested, gave an accuracy of **80%**. A tabular representation of the test statistics is given below Table 1. The final overall accuracy was calculated to be **85%** (Refer to Fig. 4) using code. The training error and accuracy (Fig. 5) and validation error and accuracy (Fig. 6) are given below.

6 Observations

The observations were as follows

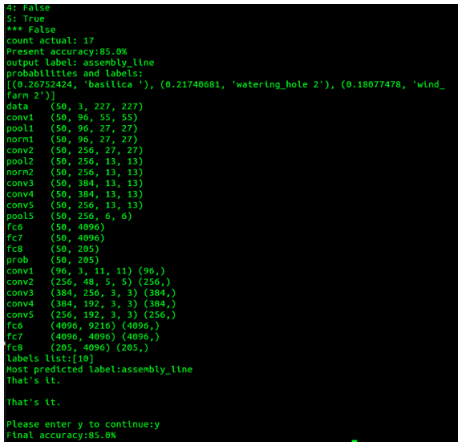

Figure 4: **An image showing the overall test accuracy of 85% on SUN397 ([?]) database**

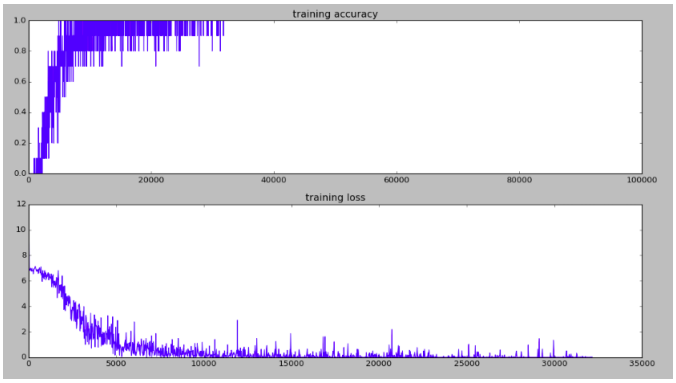

Figure 5: **An image showing training loss and accuracy**

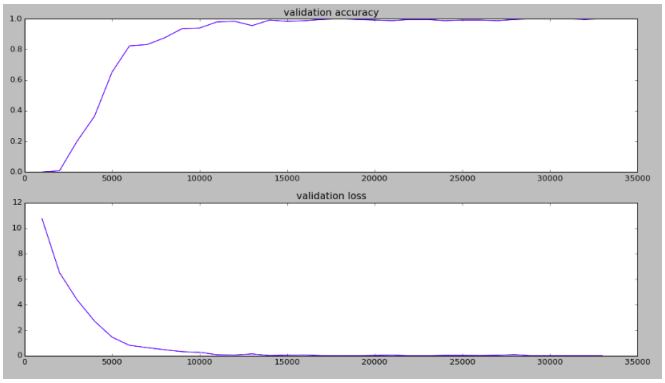

Figure 6: **An image showing validionation loss and accuracy**

1. The overall accuracy (85 %) on SUN397 ([**?**]) be determined very well determined by using the Places ([9]) CNN on the images of the scale 227 X 227, as the crop of a 227 X 227 image is same as that of the image itself.

2. A lot of information can be obtained by using the ImageNet ([**?**]) CNN on the crops of images as the ImageNet ([**?**]) CNN is often able to identify individual objects in the images.

3. The ImageNet CNN and the Places CNN are very good learning networks as they both of them, in the form of an ensemble network, give very less training error and very high training accuracy on their respective datasets which are ImageNet or ILSVRC 2012 and PLACES database respectively. They also give very low validation error and very high validation accuracy on the same datasets. These two phenomena are supported by the graphs.

4. The database bias is eliminated to a very large extent as the ensemble performs very well on both large scale and small scale images because the ensemble is made of networks which are fine-tuned and well-adapted to both large and small scales of data.

5. The overall accuracy on SUN397 ([**?**]) could have been increased by a certain amount had an object detection system been designed which could detect the constituent objects in an image and then those specific crops, which contain the objects only, could have been feed into the ImageNet CNN. In that case, the classification outputs (labels) of the ImageNet CNN could have been very well compared to the actual object labels and thus the overall accuracy could have been computed. But in that case, the database would have needed to be annotated.

7 Conclusion

We have proposed a novel approach for scene recognition, that uses a single scale to predict the accuracy. The overall accuracy, achieved on SUN397 dataset, is 85%, which is more than 20% higher than the one attained by Herranz et al. ([3]) using an architecture of the same complexity, which was 53.86%. This method is also fast and robust, and can be applied in real-time scene recognition systems with good accuracy.

References

[1] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos. Scene classification with semantic fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2015.

[2] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014.

[3] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. Scene recognition with cnns: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016.

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] Xiu-Shen Wei, Bin-Bin Gao, and Jianxin Wu. Deep spatial pyramid ensemble for cultural event recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 38–44, 2015.

[7] Ruobing Wu, Baoyuan Wang, Wenping Wang, and Yizhou Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1287–1295, 2015.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[9] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.