Exercise 3

TDT4173
About Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Simon Borøy-Johnsen & Magnus Gundersen
MTDT

March 10, 2016

1 Summarize

The paper is about relieving the existing deep convolutional networks(DCN), that does visual recognition, of the constraint of requiring a fixed-size of the input-image. The authors of the article use existing systems, and utilize the technique of ”spatial pyramid pooling” (SPP) to enhance the performance of these nets. The article also contains the description of how the networks were modified, and the source code of the implementation is also published. The authors justify their architectural choices by building the architecture step by step, and also why it helps the system in the training process. The SPP-approach is also applied to object detection, where the goal is to frame an object in an image. The authors compare their own solution to the existing ones, both by examining well-known data sets, and also joining in a competition of best visual recognition.

2 Research goals

The main goal of the authors is to address ”a technical issue in the training and testing of the CNNs: the prevalent CNNs require a fixed input image size” [2]. This issue constructs an artificial requirement when performing image classification. Images come in different scales and sizes. Some methods have to be applied to the images in order to meet the requirement. The most common methods are warping and cropping. The problem with these methods is that a lot of potential useful data might get lost in the process. For example, the cropped image might not contain the whole object the original image is capturing, and warping objects might result in unwanted distortions. The loss of data can result in unwanted loss of recognition accuracy.

The authors also address one of the greatest issues with the R-CNN [1] method for object classification in images; the need to extract thousands of candidate windows from each image, and then run a full feature mapping on each window. This is a tedious process, as feature extraction is the major bottleneck in testing [2].

2.1 Image classification

To overcome the fixed image size restriction, the authors of the article introduce a SPP layer into different deep convolutional network architectures. The SPP layer is added on top of the last convolutional layer. The layer ”pools the features” [2] from the convolutional layers, before it ”generates fixed-length outputs” [2]. These fixed-length outputs are then fed as input to the fully connected layers at the end of the network. By doing this aggregation at a deeper level of the network, the requirement for fixed-sized images can be removed. In addition to removing the fixed image size requirement, SPP provides some additional benefits; SPP uses multi-level pooling, which has been shown has been shown to be robust to object deformations [4]. Because of the flexible input scales of SPP, it can pool features from variable scales.

2.2 Object classification

The approach for improving the object classification is quite similar to that of solving the image classification issue. Instead of extracting the feature maps for each window, SPP extracts the features of the image only once, and then generates the candidate windows. The SPP layer is then applied to each window, just like in the image classification method.

3 Research methodology

The authors address their research goals in a quite direct manner. By literature review, they refer to the strengths of the previously used method of SPP. They also discuss the newer literature that is concerned with using the technique of CNN to produce even better results [3]. The research is continued in an analytical way by breaking down the structure of the existing networks. Here they conclude that the transition between the convolutional layers and the fully connected layers is the problem. Using theoretical review of the architecture behind the networks, they find out that the nature of SPP is perfect for relieving the constraint that was previously discussed.

The next step is to train the network, which the authors does in both a theoretical and experimental way. They propose to use well known frameworks such as Caffe [3] to build the networks, and make them ready for training. The training itself is mostly done in a experimental way, using both image size 180x180 in addition to 224x224. They authors show that both image-sizes produce the same input to the fully connected layer, hereby accomplishing the research goal.

4 Results and evaluation

The authors address their results in several ways, but they all use some kind of comparison to other classification methods.

4.1 Image classification

First, they compared the four before-mentioned image classification methods without SPP to the same methods with both single- and multi-sized SPP between the convolutional layers and the fully connected layers. The experiments show considerable improvements over the methods without SPP, both for top-1 and top-5 error.

The authors ran the different classification methods on three different data sets; ImageNet 2012, Caltech101, and Pascal VOC 2007. The results (mean average pricesion (mAP), accuracy and error rates) for each of the methods were then compared.

SPP	ZF-5	Covnet*-5	Overfeat-5	Overfeat-7
No SPP	35.99	34.93	34.13	32.01
SPP single-sized	34.98	34.38	32.87	30.36
SPP multi-sized	34.60	33.94	32.26	29.68

Table 1: Error rates in the validation set og ImageNet 2012

The ImageNet 2012 data set was used for comparing the four classification methods, using the error rates as measurement. The Caltech101 and Pascal VOC 2007 data sets were used for comparing the results when performing the training at different layers. On the Pascal VOC 2007 set, the authors used mAP as the measurement, and on the Caltech101 data set, classification accuracy was used.

4.1.1 ImageNet 2012

When comparing results using the ImageNet 2012 data set, the authors used both single- and multi-sized trained SPP networks.

Using standard 10-view testing, the authors were able to reduce classification error by up to 2.33% points. The improvements were all in the range [0.38, 2.33].

Table 1 shows the different results from the validation set of the ImageNet 2012 data set. For fair comparison, the authors cropped all images to 224x224 pixels. The authors also investigated the accuracy achieved by using the whole image, re-sized so that $min(width, height) = 256$. Combining six different full-size image scales, using eighteen different views for each scale (96 in total), plus two full-image scales, the authors were able to reduce the error rate to 9.14%.

4.1.2 Pascal VOC 2007

The authors used five different configurations when comparing results for the Pascal VOC 2007 data set; ZF-5 cropped to 224x224, ZF-5 with SPP cropped to 224x224, ZF-5 with SPP full-size image 224x-, ZF-5 with SPP full-size image 392x-, and Overfeat-7 with SPP full-size image 364x-.

All tests showed that the deeper the layer the training was performed on, the better the mAP was. On the ZF-5 cropped to 224x224, the results from performing the training of the last layer were 15.94% points better than when training on the fourth convolutional layer. The best mAP was achieved when training the network after the last fully connected layer using the Overfeat-7 method. A mAP of 82.44% was achieved.

4.1.3 Caltech101

The authors used the same configurations for the Caltech101 data set as for the Pascal VOC 2007 data set.

Also here, Overfeat-7 achieved the best scores, but this time, the best results were achieved when training after the SPP layer. The authors discussed this briefly, and concluded that this was most like caused by the fully connected layers being too category specialized, thus being less accurate.

4.2 Object classification

When comparing the SPP networks to R-CNN, the most important improvement was the classification speed. The R-CNN used more than 14 seconds (14.46) to classify one image. When testing the SPP network using the ZF-5 method and a single image size, the SPP network could classify image in 0.142 seconds, 102 times faster. Unfortunately, the one-scale version of SPP was 1.2 percent inferior to the R-CNN method. By using a five-scale version of the SPP network, classifying one image took 0.293 seconds (38 times faster), and achieved similar results as R-CNN. By combining two similar SPP networks, trained with different random initializations, the two networks performed better than R-CNN in 17 of 20 categories.

4.3 Competition results

The authors participated in ILSVRC 2014, a classification competition. Using the same network as when testing for ImageNet 2012, they achieved a third place for image classification. The authors also participated in the object classification competition, achieving a second place.

4.4 Evaluation

The author's evaluation of their results is based both on the experimental results of the known data sets, and the positions that the team got in the competition. In both cases the results were quite astonishing, and they are able to justify their results in a good way. The authors did this by pointing to the methodology used, and putting emphasis on the fact that the original constraint of having a fixed image size now is gone.

5 Discussion

The authors only discussed their results on the fly, after each result was presented. Almost all discussions ended in favor of the authors, except for a few times. However, every weakness was due to the authors not utilizing the full potential of the SPP network, so the authors were very quick to explain what could be done to patch the weaknesses. All in all, the authors were entirely positive to their approach. This is expected, as their results were quite astonishing.

6 Personal note

We did like the article, mainly because of its content. The research seem cutting edge, and both of us have been using ANNs for image recognition before. It was therefore interesting to see what modern nets actually can accomplish, and that fact that relatively easy-to-understand modifications can have such a large impact on performance.

The article was written in a logical way, and was easy to follow. We did however not like the unilateral approach of the evaluation and discussion made by the authors. They did not discuss nor point to any strong weaknesses of their system, which we think a good article should do. We would also have liked if the authors did point our interesting topics for further research.

References

[1] Ross Girshick et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*. Tech. rep. UC Berkeley, 2013.

[2] Kaiming He et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (Sept. 2013).

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. Tech. rep. Proc. Adv. Neural Inf. Process. Syst., 2012.

[4] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*. Tech. rep. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.