CSE616 Neural Networks and Their Applications Project 2 Submission

Ayman Wagih Mohsen (2000728)

June 2, 2022

4 Main Results

4.1 Experimental Results

Fast R-CNN uses 3 pretrained ImageNet Models:
1. The small (S) size CaffeNet, which is based on AlexNet.
2. The medium (M) size VGG_CNN_M_1024.
3. The large (L) size VGG16.

Training and testing are single sale and the length of the shortest image side s=600.

4.2 VOC 2007, 2010 and 2012 results

The PASCAL Virtual Object Classes (VOC) project is an object classification competition that had a series of challenges spanning from 2005 till 2012.
The mean average precision (mAP) of Fast R-CNN on VOC 2012 was the highest at its time (65.7%). It was also the fastest contestant on that dataset.
Fast R-CNN was surpassed only by SegDeepM on VOC 2010.
As for VOC 2007 Fast R-CNN is a large improvement over the other contestants.

4.3 Training and testing time

Truncated singular value decomposition (SVD) helps in fully connected (FC) layers. For example if the fc6 layer has weights in the shape of a 25088×4096 matrix, then by using SVD we can use the largest 1024 singular values only. Saving test time at the cost of a little precision.

4.4 Which layers to fine-tune?

With very deep networks it is not enough to fine tune just the fully connected layers. In the fast R-CNN study using the large VGG16 net, it was revealed that it is best to fine-tune the last 9 conv layers out of the total 13 ($conv3_1$ and

Table 1: Fast R-CNN train and test time compared to R-CNN

	Fast R-CNN			R-CNN			SPPnet
	S	**M**	**L**	**S**	**M**	**L**	†**L**
train time (h)	**1.2**	2.0	9.5	22	28	84	25
train speedup	**18.3**×	14.0×	8.8×	1×	1×	1×	3.4×
test rate (s/im)	0.10	0.15	0.32	9.8	12.1	47.0	2.3
▷ with SVD	**0.06**	0.08	0.22	-	-	-	-
test speedup	98×	80×	146×	1×	1×	1×	20×
▷ with SVD	169×	150×	**213**×	-	-	-	-
VOC07 mAP	57.1	59.2	**66.9**	58.5	60.2	66.0	63.1
▷ with SVD	56.5	58.7	66.6	-	-	-	-

up). And this increased mAP from 61.4% to 66.9%. Fine-tuning starting from earlier layers has more impact on train time than on mAP.
As for medium and small nets, it was shown that fine-tuning conv layers has too little of an impact on the mAP. So fine-tuning the FC layers only is enough.

5 Design evaluation

5.1 Does multi-task training help?

Multi-task learning is when a network is trained for multiple tasks simultaneously to save redundant calculations at test time, espacially in the early layers. The loss function becomes a weighted sum of sub-task loss functions. From the fast R-CNN study, it turns out that multi-task training for tasks A and B sometimes helps the network get better at task A than just training for task A. When the network was jointly trained for object classification and bounding box regression, it got better by 0.8 to 1.1 mAP at classification than training just for classification. This positive effect is consistent assuming the tasks are somehow related.

5.2 Scale invariance: to brute force or finesse?

Image scale is the length of its shortest side. Brute-force learning here means that all images are in single scale (of s=600). The aspect ration is kept the same. While multi-scale learning uses image pyramids at scales of s=480, 576, 688, 864, 1200. The fast R-CNN concludes that there is no need to train at multiple scales, as the gain in mAP is negligible compared to the increase in training memory requirements. This applies especially to the deep models like VGG16.

5.3 Do we need more training data?

When the number of images in VOC 2007 dataset was tripled to 16500 by augmenting with VOC 2012 dataset, the mAP improved from 66.9% to only 70.0%. When the VOC 2010 dataset was augmented with VOC 2007 to 21500 images and the learning rate is lowered by $0.1\times$ each 40k iterations, mAP improved more significantly from 66.1% to 68.8%. The same was done with VOC 2012 dataset and mAP improved from 65.7% to 68.4%.

5.4 Do SVMs outperform softmax?

R-CNN uses softmax, while fast R-CNN was tested with softmax and SVM. Softmax slightly outperforms support vector machines (SVMs) by 0.1 to 0.8 mAP.

5.5 Are more proposals always better?

A proposal here means a candidade region of interest (RoI). Object detectors can use a sparse set of object proposals (e.g. selective search) or a dense set (e.g. DPM). For sparse set, as the proposal count increases, mAP rises then falls slightly, peaking around 3000 or 4000 proposals. With densely generated boxes is was concluded that mAP drops with larger proposal count.

5.6 Preliminary MS COCO results

The large VGG16 net based fast R-CNN was also trained on Microsoft COCO dataset with 80k images for 240k iterations, giving mAP of 35.9%.

6 Conclusion

An improvement on Region-based CNN (R-CNN) was proposed called Fast R-CNN. The new object detection technique is orders of magnitude faster at training and testing, allowing for more experimentation to study the effects of various hyperparameters on mean average precision (mAP). The speedup was achieved as follows: instead of passing each region of interest (RoI) again and again through the CNN, the CNN is run on the whole image followed by a novel stage called ROIPooling. This eliminates redundant calculations.

References

[1] Ross Girshick (ICCV 2015) "Fast R-CNN" https://arxiv.org/abs/1504.08083