Application of Deep Learning in Object Detection

Xinyi Zhou[1], Wei Gong[2], WenLong Fu[3], Fengtong Du[4]
[1,2]Information Engineering School, Communication University of China,CUC
[3,4]Neuroscience and Intelligent Media Institute, Communication University of China
Beijng, China
xinyi_M@126.com

Abstract—**This paper deals with the field of computer vision, mainly for the application of deep learning in object detection task. On the one hand, there is a simple summary of the datasets and deep learning algorithms commonly used in computer vision. On the other hand, a new dataset is built according to those commonly used datasets, and choose one of the network called faster r-cnn to work on this new dataset. Through the experiment to strengthen the understanding of these networks, and through the analysis of the results learn the importance of deep learning technology, and the importance of the dataset for deep learning.**

Keywords—deep learning; neural network; faster r-cnn; dataset

I. Introduction

In recent years, with the rapid development of deep learning, a number of research areas have achieved good results, and accompanied by the continuous improvement of convolution neural networks, computer vision has arrived at a new peak. From the ALexNet [1] in 2012 years to the ZF Net [2] in 2013 years, and then to the VGG Net [3], the ResNet [4] and so on, the architecture of convolution neural network is constantly improving. In addition, the return of the convolution neural network also makes the application of computer vision greatly improve, such as face recognition, object detection, object tracking, semantic segmentation, and so on.

Object detection as one of the important applications in the field of computer vision has been the focus of research, and convolution neural network has made great progress in object detection. Object detection is developing from the single object recognition to the multi-object recognition. The meaning of the first is just from an image to identify a single object, it can be said that it is a problem of classification, and the meaning of the later is not only can identify all the objects in an image, including the exact location of the objects. Deep learning has formed a mainstream object recognition algorithm based on R-CNN [5], and these algorithm is refreshing the higher accuracy in a number of famous datasets.

In this paper, we first summarize the some algorithms related to deep learning for object detection, and then apply one of the algorithms to a new dataset to verify its wide applicability.

II. Dataset and NEURAL NETWORK

For deep learning, dataset and neural network are two important parts. The dataset is the fuel for deep learning so that the number and quality of the dataset will affect the accuracy of the neural network output, and the choice of neural network or the network architecture will also affect the accuracy.

A. Dataset

Dataset is one of the foundations of deep learning, for many researchers to get enough data to carry out the experiment just by themselves is a big problem, so we need a lot of open source dataset for everyone to use. Some commonly used datasets in computer vision is the following.

1) ImageNet

The Imagenet dataset [6] has more than 14 million images covering more than 20,000 categories. There are more than a million pictures with explicit class annotations and annotations of object locations in the image. The Imagenet dataset is one of the most widely used datasets in the field of deep learning. Most of the research work such as image classification, location, and detection is based on this dataset. The Imagenet dataset is detailed and is very easy to use. It is very widely used in the field of computer vision research, and has become the "standard" dataset of the current deep learning of image domain to test algorithm performance. There is a well-known challenge called "ImageNet International Computer Vision Challenge" (ILSVRC) [7] based on the Imagenet dataset. It is worth mentioning that the winners of ILSVRC2016 are Chinese teams for all projects.

2) PASCAL VOC

The PASCAL VOC (pattern analysis, statistical modelling and computational learning visual object classes) [8] provides standardized image data sets for object class recognition and provides a common set of tools for accessing the data sets and annotations. The PASCAL VOC dataset includes 20 classes and has a challenge based on this dataset. The PASCAL VOC Challenge [9] is no longer available after 2012, but its dataset is of good quality and well-marked, and enables evaluation and comparison of different methods. And because the amount of data of the PASCAL VOC dataset is small, compared to the imagenet dataset, very suitable for researchers to test network programs. Our dataset is also created based on the PASCAL VOC dataset standard.

3) COCO

COCO (Common Objects in Context) [10] is a new image recognition, segmentation, and captioning dataset, sponsored by Microsoft. COCO dataset has more than 300,000 images covering 80 object categories. The open source of this dataset

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

makes great progress in semantic segmentation in recent years, and it has become a "standard" dataset for the performance of image semantic understanding, and also COCO has its own challenge.

B. *Neural Network*

Deep learning used by the network has been constantly improving, in addition to the changes in the network structure, the more is to do some tune based on the original network or apply some trick to make the network performance to enhance. The more well-known algorithms of object detection are a series of algorithms based on R-CNN, mainly in the following.

1) R-CNN

Paper which the R-CNN (Regions with Convolutional Neural Network) is in has been the state-of-art papers in field of object detection in 2014 years. The idea of this paper has changed the general idea of object detection. Later, algorithms in many literatures on deep learning of object detection basically inherited this idea which is the core algorithm for object detection with deep learning. One of the most noteworthy points of this paper is that the CNN is applied to the candidate box to extract the feature vector, and the second is to propose a way to effectively train large CNNs. It is supervised pre-training on large dataset such ILSVRC, and then do some fine-tuning training in a specific range on a small dataset such PASCAL.

2) SPP-Net

SPP-Net [11] is an improvement based on the R-CNN with faster speed. SPP-Net proposed a spatial pyramid pooling (SPP) layer that removes restrictions on network fixed size. SPP-Net only needs to run the convolution layer once (the whole image, regardless of size), and then use the SPP layer to extract features, compared to the R-CNN, to avoid repeat convolution operation the candidate area, reducing the number of convolution times. The speed for SPP-Net calculating the convolution on the Pascal VOC 2007 dataset by 30-170 times faster than the R-CNN, and the overall speed is 24-64 times faster than the R-CNN.

3) Fast R-CNN

For the shortcomings of R-CNN and SPP-Net, Fast R-CNN [12] did the following improvements: higher detection quality (mAP) than R-CNN and SPP-Net; write the loss function of multiple tasks together to achieve single-level training process; in the training can update all the layers; do not need to store features in the disk. Fast R-CNN can improve the speed of training deeper neural networks, such as VGG16. Compared to R-CNN, The speed for Fast R-CNN training stage is 9 times faster and the speed for test is 213 times faster. The speed for Fast R-CNN training stage is 3 times faster than SPP-net and the speed for test is 10 times faster, the accuracy rate also have a certain increase.

4) Faster R-CNN

The emergence of SPP-net and Fast R-CNN has greatly reduced the running time of the object detection network. However, the time they take for the regional proposal method is too long, and the task of getting regional proposal is a bottleneck. Faster R-CNN [13] presents a solution to this problem by converting traditional practices (such as Selective Search, SS) to use a deep network to compute a proposal box (such as Region Proposal Network, RPN). In this paper, our experiment chooses the Faster R-CNN network. Table I shows the comparison of mean Average Precision (mAP) of above four kinds of network structure on the VOC2007 dataset.

TABLE I. MEAN AVERAGE PRECISION (MAP) OF DIFFERENT NETWORK ON THE VOC2007 DATASET

Network	***R-CNN***	***SPP-Net***	***Fast R-CNN***	***Faster R-CNN***
VOC07 mAP	0.66	0.631	0.669	0.732

III. APPLICATION OF FASTER R-CNN ON NEW DATASET

In the experiment, we must need a new dataset, the dataset format as seem as VOC data set format. We have created a football game image dataset, which has the four categories of objects that is player, football, soccer goal, corner flag. As shown in Figure 1.

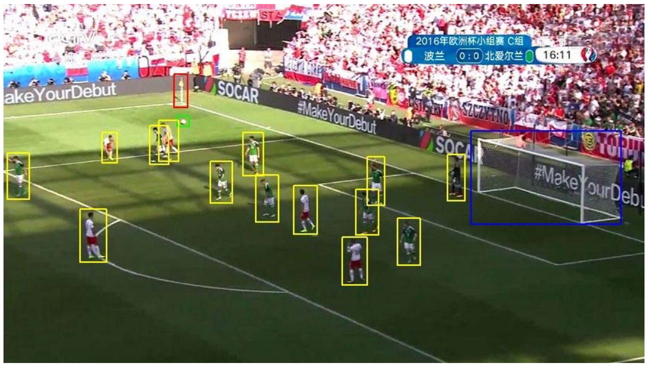


Fig. 1. A marked image

Red marked in the figure is the corner flag, blue marked is soccer goal, green marked is football, and yellow marked is player. In the process of marking the image, in addition to marking each of the four categories of object, but also marked the location of the object which is shown with form of bounding box, and also marked whether objects is clipped or not, whether the object is well recognized, and marked the angle of the player, and marked whether the image is a long view or a close view. These annotated messages are saved in xml format. Take the simplest Figure 2 as an example, and its label of the xml format is shown below.

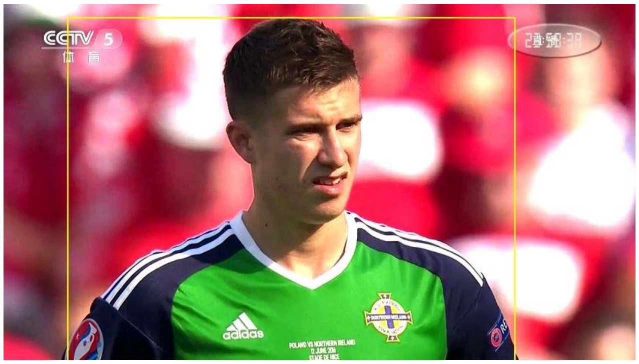
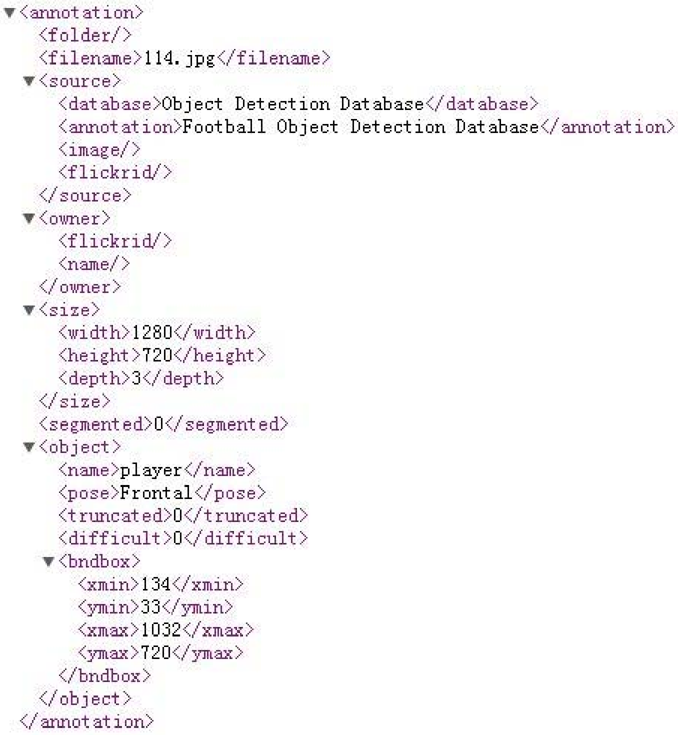


Fig. 2. A simple marked image



```
▼<annotation>
   <folder/>
   <filename>114.jpg</filename>
  ▼<source>
     <database>Object Detection Database</database>
     <annotation>Football Object Detection Database</annotation>
     <image/>
     <flickrid/>
   </source>
  ▼<owner>
     <flickrid/>
     <name/>
   </owner>
  ▼<size>
     <width>1280</width>
     <height>720</height>
     <depth>3</depth>
   </size>
   <segmented>0</segmented>
  ▼<object>
     <name>player</name>
     <pose>Frontal</pose>
     <truncated>0</truncated>
     <difficult>0</difficult>
    ▼<bndbox>
       <xmin>134</xmin>
       <ymin>33</ymin>
       <xmax>1032</xmax>
       <ymax>720</ymax>
     </bndbox>
   </object>
 </annotation>
```

Fig. 3. Label information

From the label information as show in Figure 3 can be seen some other information, including the name of the image and the name of the dataset. The size and depth of the image are recorded under the 'size' tab. The information under the 'object' tab is the content we marked before. Then 'name' tab records the class of the object in the range of those four category; 'pose' tab records the person's perspective, which is divided into Frontal, SideFaceLeft, SideFaceRight, Rear and Unspecified; 'truncated' tab records whether the object is clipped or truncated, with clipping or truncation set to 1 and no set to 0; 'difficult' tab records whether the object is well recognized, if it is clear and easy to identify set to 0, otherwise set to 1; 'bndbox' tab records the specific location of the object.

All the images are the frame from the football game videos, and then sorting so that these images are suit for object recognition task of the football game. The entire football dataset includes 5357 images, in which the training set has 3587 images, the test set has 1770 images. After having a new dataset, then we choose the faster r-cnn network working on this football dataset.

IV. RESULTS

After training, some effect on the testset is shown below:

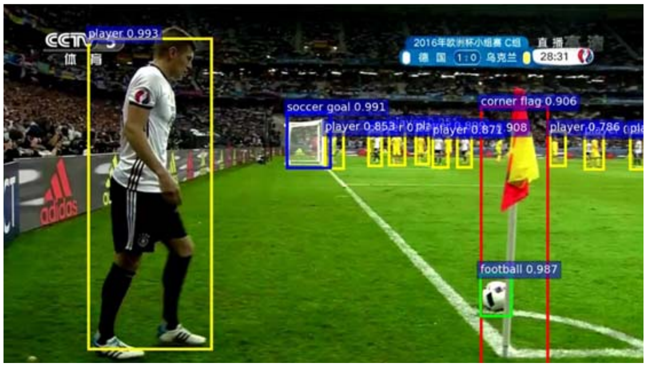


Fig. 4. Recognize all classes

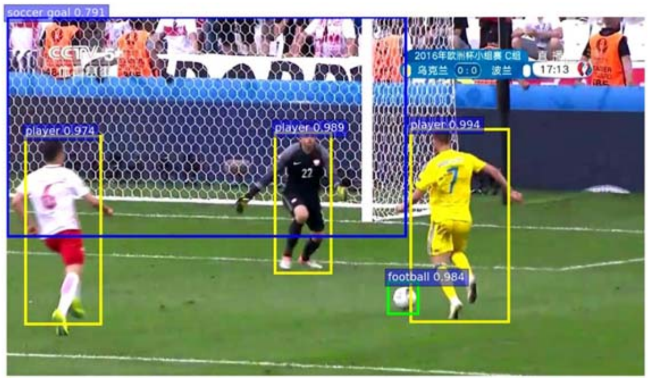


Fig. 5. Recognize player, football and soccer goal

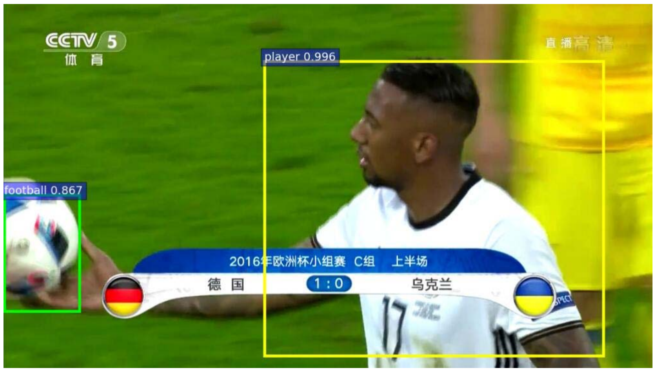


Fig. 6. Recognize player and football

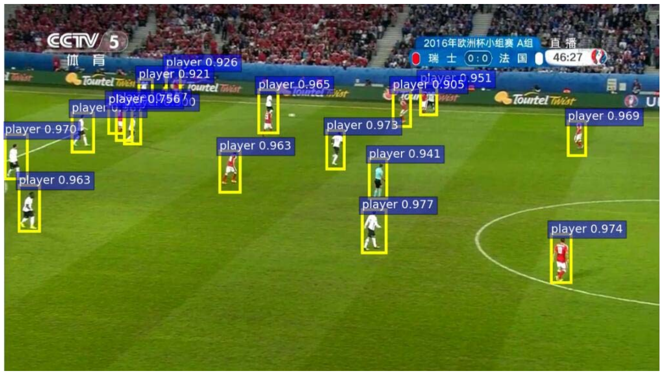


Fig. 7. Recognize all players

From these pictures can see intuitively the player could be identified very good, and other classes can also be recognized well when the image is close view. The Mean Average Precision (MAP) for each class is shown in the following table:

TABLE II. THE MEAN AVERAGE PRECISION (MAP) FOR EACH CLASS

Class	***player***	***soccer goal***	***corner flag***	***football***
mAP	0.7902	0.8377	0.3508	0.4752

From this table we can see the correct rate of class player and soccer goal is high, and the correct rate of class corner flag and football is low. The reason of the two classes having low mAP may be as follows.

A. Uneven Quantity

Compared to player and soccer goal, the quantity of corner flag and football is less, especially corner flag.

B. Uneven Size

Because the dataset has a large number of long view pictures, compared to player and soccer goal, the size of corner flag and football are too small in the long view images. So, even in the process of marking the dataset of those small objects, bounding box range is not enough accurate so that making recognition of the two types is relatively poor for long view image. Although the size of player is also relatively small in the long view pictures, but the number is very large, which to some extent improves the accuracy of recognition.

V. CONCLUSION

This paper expresses the importance of deep learning technology applications and the impact of dataset for deep learning through the use of the faster r-cnn on new datasets. In recent years, the technology of deep learning in image classification, object detection and face identification and many other computer vision tasks have achieved great success. Experimental data shows that the technology of deep learning is an effective tool to pass the man-made feature relying on the drive of experience to the learning relying on the drive of data. Large data is the base of the success of deep learning, large data just as fuel to the rocket for deep learning. More and more applications are continually accumulating increasingly rich application data, which is critical to the further development and application of deep learning. However, the quality of the data affects the deep learning in deed, of course, in addition to these real data, maybe we can also consider some of synthetic data to increase the amount of data in the further.

ACKNOWLEDGMENT

We would like to thank the Beijing Innovation Collaborative Institute(Video intelligent creation platform and demonstration application,16I -04-BJ)

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS,2012.

[2] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In ECCV,

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, 2016.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009.

[7] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. FeiFei. ImageNet Large Scale Visual Recognition Competition 2012 (ILSVRC2012).

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV, 2010.

[10] Lin, Tsung Yi, et al. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014. Springer International Publishing, 2014:740-755.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014.

[12] R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.