International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 3 170 - 174

A Survey on Object Recognition Using Deep Neural Networks

[1]Wangkheimayum Madal, [2]Dr. Lakhmi Prasad Saikia
[1]M.Tech, Computer Sc & Engg, Assam downtown University,India
[2]Professor, Computer Sc& Engg, Assam downtown University, India
[1]*madalwangkheimayum@gmail.com* :[2]*lp_saikia@yahoo.co.in*

***Abstract*—** Deep Neural Networks as a means of objects detection and recognition is an active area of research and several discoveries have been made in this field. Here we will be discussing briefly about the history of research in the field of computer vision, mainly for the application of deep learning in object detection task and describe several of the recent advances in this field. This paper describes a simple summary of the datasets and deep learning algorithms commonly used in computer vision, some of the applications of this field have been provided.

Keywords- —*Convolutional, Neural Networks, Datasets,Deep learning,faster r-cnn,SSD(Single Shot MultiBox Detector),YOLO(You Only Look Once).*

*****

I. INTRODUCTION

In Artificial Intelligence, Object Recognition and Object Detection still remains a major challenging task. Many recognition systems have been developed to recognize and classify images over a decade. In recent times, several outstanding results have been achieved in this field of research and computer vision reached its zenith. ALexNet [1] in 2012, ZF Net [2] in 2013, and then VGG Net [3], ResNet [4], etc. are worth mentioning. The architecture of convolution neural network is improving constantly with the application of convolution neural network in the field of computer vision greatly improve, such as object detection, object recognition, object tracking, face recognition, and so on.

The focus of several researches in the field of computer vision has been on Object detection as it has several applications in it, and convolution neural network has made great progress in object detection. From single object recognition to multi object recognition, researches in this field are making huge progress. Traditionally, algorithms were designed to look for predetermined image features. It was quite a back-aching task as the developer needed to have a thorough knowledge about the data of the image and had to engineer each and every feature detection algorithm in this traditional approach. A weakness of this system was that it was vulnerable to small ambiguities which may be present in the concerned image. These problems have been removed with the new approach from Neural Network as it has the following advantages: firstly, they are data driven self-adaptive algorithms; no prior knowledge of the data or underlying properties is needed, secondly, they can approximate any function with arbitrary accuracy [5] [6] [7]; as any classification task is essentially the task of determining the underlying function, this property is important and thirdly, neural networks can estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analysis [8]. In this paper, we will summarize some of the algorithms related to the recent improvements in the deep learning of object detection which led to fantastic results in Object Recognition.

II. HISTORY OF NEURAL NETWORKS

Neurophysiologist Warren McCulloch and mathematician Walter Pitts can be considered as pioneers in the field of Neural Networks with their experiments in 1943 in which they modeled a simple neural network with electrical circuits. The neuron took inputs and depending on the weighted sum, it would give out a binary output. In 1950s neural networks were simulated on larger scale with the coming of more powerful computers. In 1955, IBM organized a group to study pattern recognition, information theory and switching circuit theory, headed by Nathanial Rochester [10]. Alongside the research on Artificial Neural Networks, basic research on layout of neurons inside the brain was also being conducted. The idea of a Convoluted Neural Networks can be traced to Hubel and Wiesels 1962 work on the cat's primary visual cortex. It identified orientation-selective simple cells with local receptive fields, whose role is similar to the Feature Extractors, and complex cells, whose role is similar to the Pooling units. The first such model to be simulated on a computer was FukushimasNeocognitron [11], which used a layer-wise, unsupervised competitive learning algorithm for the feature extractors, and a separately-trained supervised linear classifier for the output layer. In 1985, Yann Le Cun proposed an algorithm to train Neural Networks. The innovation [12] was to simplify the architecture and to use the back-propagation algorithm to train the entire system. The approach was very successful for tasks such as OCR and handwriting recognition. By the late 90s it was reading over 10% of all the checks in the US. This motivated Microsoft to deploy Convolutional Neural Networks in a number of OCR and handwriting recognition systems including for Arabic and Chinese characters. Supervised Convolutional Neural Networks (ConvNet) have also been used for object detection in images, including faces

with record accuracy and real-time performance. Google recently deployed a Convolutional Neural Networks (ConvNet) to detect faces and license plate in Street View images to protect privacy.

More recently, a lot of development has occurred in this field leading to a number of improvements in the performances and accuracy. In ILSVRC-2012 (Large Scale Visual Recognition Challenge) the task was to assign labels to an image. The winning algorithm produced the result [14], the accuracy in the task was as described in the image caption was 83%. Two years since then, in ILSVRC-2014, the winning team from Google had an accuracy of 93.3% [13].

III. DATASET AND NEURAL NETWORK

For deep learning, dataset and neural network are two important parts. Availability of lot of dataset through internet and emergence of high processing CPU and GPU at reasonable price which speed up training of dataset to neural network make the rapid progress in the field of 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

One of the difficulties faced in the early experiments of Deep Learning was the limited availability of labeled data sets. Many image datasets have now been created and are growing rapidly to meet the demand for larger data sets by the Image and Vision Research Community. The following is a list of data sets frequently used for testing object classification algorithms.

1) ImageNet: The Imagenet dataset [15] 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) [16] 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) 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 [17] 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.

3) COCO: COCO (Common Objects in Context) [18] is a new image recognition, segmentation, and captioning dataset, sponsored by Microsoft. COCO dataset has more than 328,000 images covering 91 object categories. The open source of this dataset 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.

4) CIFAR-10 and CIFAR-100: These subsets are derived from the Tiny Image Dataset, with the images being labelled more accurately. The CIFAR-10 set [19] has 6000 examples of each of 10 classes and the CIFAR-100 set has 600 examples of each of 100 classes. Each image has a resolution of 32×32

5) STL-10: The STL-10 dataset [20] is derived from the Imagenet. It has 10 classes with 1300 images in each class. Apart from these it has 100000 unlabeled images for unsupervised learning which belong to one of the 10 classes. The resolution of each image is 96×96.

6) Street View House Numbers: SVHN [21] is a real world image dataset with minimal requirement on data preprocessing and formatting. It can be seen as similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images. The resolution of the images is 32×32.

7) MNIST: The MNIST database [22] of handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image of resolution 28×28.

8) NORB [23]: This database is intended for experiments in 3D object recognition from shape. It contains images of 50 toys belonging to 5 generic categories: four-legged animals, human figures, airplanes, trucks, and cars. The objects were imaged by two cameras under 6 lighting conditions, 9 elevations (30 to 70 degrees), and 18 azimuths (0 to 340). The training set is composed of 5 instances of each category and the test set of the remaining 5 instances, making the total number of image pairs 50.

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 [24]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 [25] 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 [26] 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).

5) SSD: Single Shot MultiBox Detector(SSD) method for detecting objects in images using a single deep neural network. Discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. SSD [27] is simple relative to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. This makes SSD easy to train and straightforward to integrate into systems that require a detection component. Experimental results on the PASCAL VOC, COCO, and ILSVRC datasets confirm that SSD has competitive accuracy to methods that utilize an additional object proposal step and is much faster, while providing a unified framework for both training and inference. For 300 × 300 input, SSD achieves 74.3% mAP1 on VOC2007 test at 59 FPS on a Nvidia Titan X and for 512 × 512 input, SSD achieves 76.9% mAP, outperforming a comparable state-of-the-art Faster R-CNN model.

6) YOLO: You Only Look Once (YOLO), a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, it frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. YOLO [28] architecture is extremely fast. YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

Table I shows the mean average precision (mAP), FPS, and number of bounding boxes by each detection system on PASCAL VOC 2007 dataset.

TABLE I. COMPARING THE PERFORMANCE AND SPEED OF OBJECT DETECTOR SYSTEM ON PASCAL VOC 2007

System	*mAP*	*FPS*	*Number of boxes*
RCNN	58.5	6	-
Fast-RCNN	70	0.5	-
Faster-RCNN(VGG16)	73.2	7	300
YOLO	63.4	45	98
Fast YOLO	52.7	155	98
SSD300(VGG)	72.1	58	7308
SSD500(VGG16)	75.1	23	20097

IV. EMERGING APPLICATIONS

Having demonstrated a high level of accuracy, Convoluted Neural Networks are seeing applications in many fields. Such as -

1) *Image Recognition* [29] - Neural Networks have been already deployed in Image Recognition Applications. The Google Image Search is based on [14].
2) *Speech Recognition* [30] - Most current speech recognition systems use Hidden Markov Models (HMMs) to deal with the temporal variability of speech and Gaussian Mixture Models (GMMs) to determine how well each state of each HMMs fits a frame or a short window of frames of coefficients that represents the acoustic input. Deep neural networks with many hidden layers, that are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin.
3) *Image Compression* - Neural Networks have a property of creating a lower dimensional internal representation of input. This has been tapped to create algorithms for image compression. These techniques fall into three main categories - direct development of neural learning algorithms for image compression, neural network implementation of traditional image compression algorithms, and indirect applications of neural networks to assist with those existing image compression techniques [31].
4) *Medical Diagnosis* - There are vast amounts of medical data in store today, in the form of medical images, doctors' notes, and structured lab tests. Convoluted Neural Networks have been used to analyze such data. For example in medical image analysis, it is common to design a group of specific features for a high-level task such as classification and segmentation. But detailed annotation of medical images is often an ambiguous and challenging task. In [32] it shown that deep neural networks have been effectively used to perform this tasks.

V. CONCLUSION

We have summarized the recent advancements made in the field of Deep Neural Network for Object recognition and the importance of deep learning technology applications and the impact of dataset for deep learning has also been expressed briefly. Reliability is must in the datasets being used, as annotating them gets difficult when they get larger. Crowd sourcing has been used to create big datasets - like TinyImage dataset [33], MS-COCO [38] and ImageNet [15] but still have many ambiguities that have removed manually. Better crowd sourcing strategies have to develop. Training of Neural Networks requires a huge amount of computational resource. Efforts have to be made to make the code more efficient and compatible with new upcoming High Performance Computational Platforms

Recently, huge achievements have been made in the technology of deep learning in computer vision tasks like image classification, object detection and face identification. 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 fuel to the rocket for deep learning, it is the very foundation of the success of deep learning.

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.
[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, 2016.
[5] G. Cybenko (1989),"Approximation by superpositions of a sigmoidal function", Math. Contr. Signals Syst., vol. 2, pp. 303314
[6] K. Hornik (1991),"Approximation capabilities of multilayer feedforward networks", Neural Networks, vol. 4, pp. 251257.
[7] K. Hornik, M. Stinchcombe, and H. White (1989),"Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, pp. 359366.
[8] M. D. Richard and R. Lippmann (1991),"Neural network classifiers estimate Bayesian a posteriori probabilities", Neural Computation, vol. 3, pp. 461483.
[9] McCulloch, W. and Pitts, W. (1943),"A logical calculus of the ideas immanent in nervous activity", Bulletin of Mathematical Biophysics.
[10] Crevier, Daniel (1993), p. 39"AI: The Tumultuous Search for Artificial Intelligence", ISBN 0-465-02997-3.

[11] Kunihiko Fukushima,"Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position",Biological Cybernetics 1980.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Handwritten digit recognition with a backpropagation network, in NIPS89.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, SanjeevSatheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei (2014), "ImageNet Large Scale Visual Recognition Challenge", arXiv:1409.0575.

[14] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012),"ImageNet classification with deep convolutional neural networks", NIPS2012.

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

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

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

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

[19] Alex Krizhevsky (2009),"Learning Multiple Layers of Features from Tiny Images".

[20] Adam Coates, Honglak Lee, Andrew Y. Ng (2011)"An Analysis of Single Layer Networks in Unsupervised Feature Learning", AISTATS.

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng (2011), "Reading Digits in Natural Images with Unsupervised Feature Learning NIPS Workshop on Deep Learning and Unsupervised Feature Learning".

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner (1998),"Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324.

[23] Y. LeCun, F.J. Huang, L. Bottou (2004),"Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting", CVPR.

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

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

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

[27] Wei Liu, DragomirAnguelov, DumitruErhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg ,"SSD: Single Shot MultiBox Detector" in University of Michigan, Google Inc.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and AliFarhadi" You Only Look Once: Unified, Real-Time Object Detection", at University of Washington, Allen Institute for AI

[29] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, Marc Aurelio Ranzato, Andrew Senior, P. Tucker, Ke Yang, A. Y. Ng (2012), "Large Scale Distributed Deep Networks Jeffrey", NIPS 2012

[30] G. Hinton, Li Deng, Dong Yu, G. Dahl, A. R. Mohamed, N. Jaitly, Andrew Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury (2012), "Deep Neural Networks for Acoustic Modeling in Speech Recognition", Google

[31] J. Jiang (1999),"Image compression with neural networks a survey", Signal Processing: Image Communication, vol 14, issue 9 pg 737-760

[32] Yan Xu, Tao Mo, Qiwei Feng, PeilinZhong, Maode Lai, Eric I-chao Chang (2014), "DEEP LEARNING OF FEATURE REPRESENTATION WITH MULTIPLE INSTANCE LEARNING FOR MEDICAL IMAGE ANALYSIS", 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)

[33] A. Torralba, R. Fergus and W.T. Freeman (2008),"80 million tiny images: a large dataset for non-parametric object and scene recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30(11), pp. 1958-1970.

.