ISSN 1870-4069

Evolution of Modern Deep Learning Methods of Object Recognition

Pritom Hazarika, Ms Madhu Kumari

NIT Hamirpur, Himachal Pradesh, India

mimikaan@gmail.com, madhu.jaglan@gmail.com

Abstract. Vision is a very complex work of our brain. 90% of the information entered to our brain is related to vision. In computer vision engineers and scientists are trying to give the computers the ability of vision. Object recognition is one of the most exciting fields of computer vision and AI. There is no rarity of problems and challenges in object recognition, from image classification to key-point detection. But like many other problems in the world, there is still no obvious and "Best" way to resolve these problems. But the recent advancement in GPU technology propelled the success and accuracy of deep learning algorithms. In this paper, we will study about the evolution of deep learning methods in object recognition. We will also go through some of the classical machine learning methods for object recognition and talk about multimodal approaches to solving these issues.

Keywords. Computer vision, AI, object recognition, GPU, deep learning, machine learning.

1 Introduction

Object recognition is one of the most exciting fields of computer vision and AI. There are countless practical application of object recognition to solve real-world problems. Face detection, people counting, anomaly detection, web image classification, self-driving cars, video surveillance are some practical example where object recognition is used. In recent years image classification techniques surpassed the human ability.

There are multiple subfields of object recognition. Image classification, where an image is classified into many different categories. Object Localization, which is similar to classification. Localization predicts the location of a dominant object inside the image. Object detection, it is the combination of localization and classification. It is the process of finding and classifying multiple numbers of objects on an image. Instance Segmentation, here we not only find objects inside an image but label each pixel of an image by object class and object in the region. Key-point detection, detect locations of a set of predefined key points of an object in an image, such as the human body or a human face.

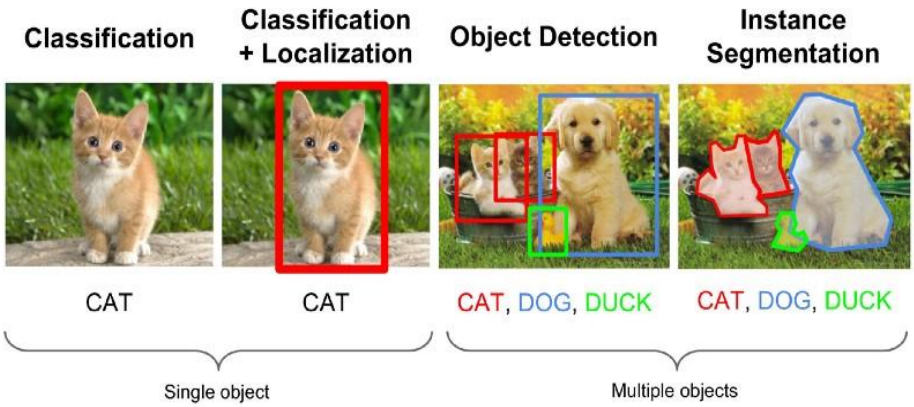

Fig. 1. Object recognition problems.

In 2001 Paul Viola and Michael Jones [10] invented a simultaneous face detection algorithm allowing for a human figure to be identified through their facial traits. Navneet Dalal and Bill Triggs [11] published a Histograms of Oriented Gradients (HOG) in 2005 which theories a feature detector for the recognition of pedestrians in security system circuits. The modern era of object recognition starts with the development of the convolutional neural network (CNN).

In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton introduce a new algorithm called AlexNet [1] ensuring an 85% level of accuracy. One of the first deep learning method used for object detection was Overfeat [7] in 2013 where P Sermanet proposed a multi-scale sliding window algorithm using CNN. Quickly after that Ross Girshick, et al. proposed a Region based convolutional neural network which is a combination of heuristic region proposal method and CNN feature extractor [2].

In 2015 Ross Girshick proposed Fast R-CNN [3]. It applied the CNN on the complete image and then used both Region of Interest (RoI) Pooling on the feature map with a final feedforward network for classification and regression to generate object proposal, unlike R-CNN which use selective search independently and use Support Vector machine as a classifier.

YOLO is published in 2015 by Joseph Redmon which use CNN [5]. YOLO's level of accuracy in facial recognition exceeded 95% and it was also very fast, which allows for the very first time to use facial recognition in real time. Subsequently, Shaoqing Ren co-authored by Girshick proposed the third iteration of the R-CNN series the Faster R-CNN [4]. In Faster R-CNN they make the model trainable end to end by adding region proposal network (RPN).

2 Challenges

There are a lot of issues and challenges related to object recognition.

2.1 Variable Number of Objects

Usually while training a machine learning modal we need to represent data into fixed-sized vectors. If we don't know the number of object in an image, we can't tell the correct number of output. It will create a problem determining the vector size. Post-processing is required to solve this problem but post-processing increase the complexity of the model.

2.2 Sizing

All the objects in an image are not of the same size. The size difference of the objects is a big challenge in object recognition model. When doing classification we generally want to recognize the object covering the large portion of the image. But sometimes some object covers a comparatively small portion of an image. This problem could be solved using variable size sliding windows but this solution is very inefficient.

2.3 Modelling

Solving two problem at the same time is another challenge. Combining classification and localization into a single model is a challenge.

2.4 Illumination

Depending on the lighting condition the same object may look different on different images. The system must be able to recognize the object irrespective of the lighting condition.

2.5 Occlusion

Sometimes objects on the image are not completely visible. Some objects are partially covered by some other object. Model must be able to handle these situation.

Noise, blurry picture, deformation, interclass variation, background clutter etc. are also some problems faced in object recognition.

3 Object Recognition Methods

3.1 Classical Methods

Over the years there are many different types of methods proposed to solve object recognition problems. But two methods stands out of all. First on is in 2001 by Paul Viola. He published a paper "Robust Real-Time Object Detection". This approach is fast and relatively simple. It is being applied in point and shoot cameras to detect faces in real time. This method creates different binary classifier by using Haar feature then

these binary classifiers are assessed with a multi-scale sliding window in cascade and dropped early in case of a negative classification.

The second method is the Histogram of oriented Gradient or HOG feature proposed by Navneet Dalal and Bill Triggs. They use Support Vector Machine (SVM) for classification. It involves a multiscale sliding window similar to Paul's method. In terms of accuracy, it is superior to the first method but it is much slower than Paul's method.

3.2 Deep Learning Methods

Deep learning revolutionize the field of machine learning, especially computer vision. Deep learning models have outperformed the other classical methods of object recognition. Modern history or object recognition stared in 2012 with the development of the convolutional neural network. It all started when AlexNet won the ILSVRC 2012 by a large margin. AlexNet was based on the decades-old LeNet, combined with data augmentation, rectified linear unit (ReLU), dropout, and GPU implementation. It proved the effectiveness of a convolutional neural network, it opened a new era for computer vision.

OverFeat: In 2013 P Sermanet from NYU proposed a multiscale sliding window algorithm using AlexNet to extract feature from an input image.

R-CNN: Region-based convolutional neural network (R-CNN) is a natural combination of heuristic region proposal method and CNN feature extractor. From an image, possible objects are extracted using a region proposal method like selective search. Those regions are then cropped and warped to a fixed size. CNN is used to extract features from each region. Then a support vector machine model is trained to classify each region. Training an R-CNN is a difficult process although it can achieve great results.

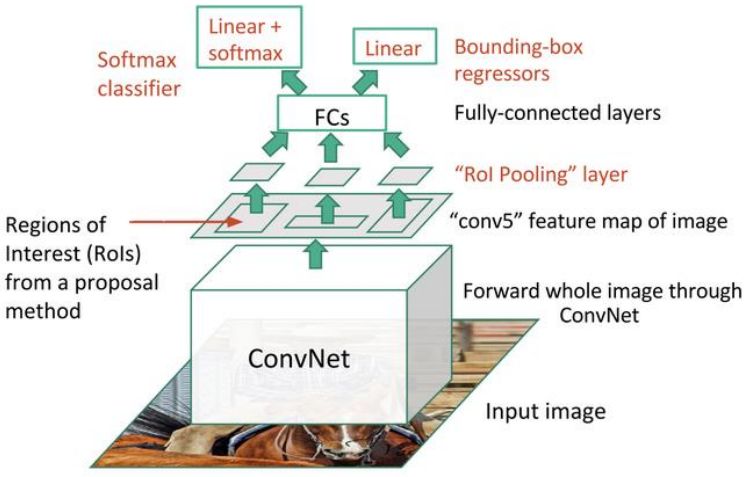

Fig. 2. Fast R-CNN.

Fast R-CNN: Fast R-CNN is similar to R-CNN. Like R-CNN, it also uses selective search to extract possible object. But difference comes in feature extraction step. Instead of applying SVM in individual region it applies CNN in the entire image to extract feature. And applies the Region of Interest (ROI) pooling on the feature map with a final feed forward network for classification and regression. The biggest disadvantage of this system is, it is still relied on selective search for region proposal.

YOLO: In 2015 Joseph Redmon published a paper You Only Look Once: Unified, Real-Time Object Detection (YOLO). YOLO is a development of multibox which is a CNN based region proposal solution. It Convert multibox from a region proposal system to object recognition system by adding softmax layer parallel to the box regressor and box classifier layer, to directly predicts the object class. It gives great result as well as high speed.

Faster R-CNN: Faster R-CNN is a Fast R-CNN where selective search is replaced by Region Proposal Network (RPN) for region proposal. RPN is also inspired by multibox. This makes the modal completely trainable from end to end.

SSD: Single shot detector uses the RPN of Faster R-CNN [6]. In Faster RCNN, RPN is used to give object confidence score but here it directly uses the RPN to classify object inside the prior box.

Mask R-CNN: It is a modified Faster R-CNN for segmentation. Here a Branch is added for predicting class specific object mask [8]. Mask RCNN replace RoIPooling with RoIAlign since the previous technique was not designed for pixel to pixel alignment.

3.3 Multimodal Methods

Deep learning methods of object recognition have impressive results. But to train a deep learning model we need a lot of data. For classification problems we need labelled data. But labelling image is time consuming task. Images from internet are sometimes not very well labelled or inaccurately. Multimodal machine learning is the solution to this problem. Modality refers to the type of information or data representation format in which information is stored. The way we perceived the world multimodal. We see things, we hear sounds, smell odours, feel texture. This is how we infer knowledge from the world. Multimodal machine learning is based on the same concept. Information from different modality compliments each other. For example, image classification and image captioning models relies on labelled input data. But labels maybe incorrect or unavailable. In such cases descriptions, tags available along with the image could be used to train the model. There are some core technical challenges in multimodal learning [9]:

- Representation,
- Alignment,
- Translation,
- Fusion,
- Co-learning.

Due to the heterogeneity of multimodal data, it is difficult to construct a representation to represent the data from different modal which can exploits the complementarity and redundancy of multimodality. There are two major types of multimodal representation –joint representation and coordinated representation. Joint representation projects unimodal representations together into a multimodal space.

Mathematically, joint representation is expressed as

$$X_m=f(x_1,x_2,\ldots\ldots,x_n). \quad (1)$$

Instead of projecting modality into joint space, coordinated representation learn separate representation for each modality but coordinate them through a constraint. Mathematically, coordinated representation is expressed as

$$f(x_1)\sim g(x_2). \quad (2)$$

Multimodal deep Boltzmann Machine is an example of joint representation. It is a graphical model based representation which stack restricted Boltzmann machines as building blocks. Canonical correlation analysis (CCA) is a correlated representation model. It computes linear projection which maximize the correlation between two random modalities and enforces orthogonality of the new space.

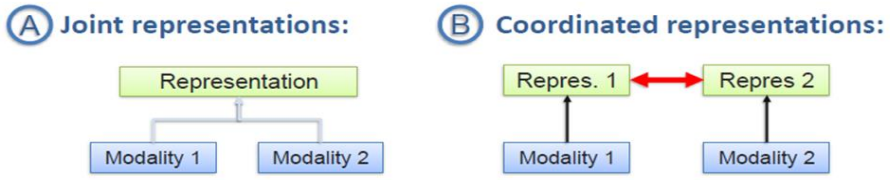

Fig. 3. Multimodal representation.

Task of identifying direct relationship between elements from different modality is known as alignment. Multimodal alignment are categorize into two parts – implicit and explicit. In explicit alignment the goal is to directly find correspondences between elements of different modalities, whereas in implicit alignment the alignment is used as an intermediate step for another task. There are two types of algorithm that handles explicit alignment- supervised and unsupervised. There are two types of implicit alignment – graphical model and neural network.

Fusion is the task of joining information from two or more modality to perform prediction. Multimodal fusion could be classified in to two broad category. Model-agnostic and model based. Model –agnostic fusion can be divided as early fusion, late fusion and hybrid fusion. Model based fusion can be categories as multiple kernel learning (MKL), graphical models and neural networks. MKL approach is a popular method for fusing visual descriptor for object detection.

Translation is the task of translating data from one modality to another. Data from different modality are heterogeneous and relationship between modalities is often open ended and subjective. There could be multiple correct answers for the same problem. Translation could be categorized as example based and generative. Example based

model used a dictionary while translating between modals. Generative models constructs models that is able to produce translation.

The last challenge of multimodal machine learning is Co-Learning it is concern with transfer of knowledge between multiple modality. Co-learning is categorized into parallel, non-parallel and hybrid data. For object recognition task primarily non-parallel and hybrid data is used. Frome et al. used transfer learning method and used text to improve visual representations for image classification by synchronizing CNN visual feature with word2vec textual feature. Zero shot learning is another popular algorithm for image classification. Here to recognize any concept the model need not see any explicit example of that concept.

4 Important Dataset

Table 1. Dataset table.

Name	Images	Classes
ImageNet	450K	200
COCO	120K	80
Pascal VOC	12K	20
Oxford-IIIT Pet	7K	37
KITTI Vision	7K	3
MNIST	70K	10
Open Images Dataset	900K	5K
SVHN	630K	10

5 Conclusion

Object recognition gives the computer and robots the ability to see. Computer scientists have been working in computer vision technology since 1966 when students of MIT are asked to solve the human vision problem as a summer project. Since then there has been a lot of progress in this field. At the core of all computer vision problem is the task of classification. Since the images are represented as a 3D array of numbers, with integers from 0 to 255, there is a semantic gap. Besides, there are challenges such as Illumination, Deformation, Occlusion, Background Clutter and Interclass variation.

In recent years classical machine learning methods for object recognition are being outperformed by deep learning methods which have achieved accuracies that are far beyond that of classical ML methods. Deep learning methods scale effectively with data. The feature engineering is not required in deep learning as it was required in classical ML methods. Classical ML methods also have some advantages over deep learning methods. Classical methods work better on small data. It is also cheap financially and computationally compared to deep learning methods. Also, classical ML methods are easier to understand.

Multimodal methods took advantage of redundant information from different modalities. The use of new sensing modalities, in particular depth and thermal cameras, has seen some development in the recent years [e.g., Fehr and Burkhardt (2008) and Correa et al. (2012)]. Classical ML and deep learning methods are applied to multiple modalities of data to solve new problems such as caption generation, depth sensing, segmentation using thermal and depth camera.

Still, there are some problems which we believe have not been addressed, or addressed partially, and may be interesting relevant research directions. Open world learning and active vision is one such area. Another area which is not being addressed completely is Pixel-Level Detection (Segmentation) and Background Objects. Another basic dilemma during the detection process is, should we detect the object first or the parts first? And there are no clear solution exists yet.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), Vol. 1, Curran Associates Inc., USA, 1097–1105 (2012)
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14), IEEE Computer Society, Washington, DC, USA, 580–587 (2014) DOI: https://doi.org/10.1109/CVPR.2014.81
3. Girshick, R.: Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15), IEEE Computer Society, Washington, DC, USA, 1440–1448 (2015) DOI=http://dx.doi.org/10.1109/ICCV. 2015.169
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), Vol. 1, MIT Press, Cambridge, MA, USA, 91–99 (2015)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 779–788 (2016)
6. Liu., W. Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: Single Shot MultiBox Detector. ECCV (2016)
7. Sermanet, P., Eigen, D., Zhang, X., et al.: OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks CoRR. Vol. abs/1312.6229 (2013)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In Proceedings IEEE International Conference on Computer Vision (ICCV), Venice, pp. 2980–2988 (2017)
9. Baltrušaitis, T., Ahuja, C., Morency, L.: Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence
10. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. Int. J. Comput. Vision 57, 2, 137–154 (2004) DOI: https://doi.org/10.1023/B:VISI.0000013087.49260.fb

11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In Proceeding IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, vol. 1, pp. 886–893 (2005) DOI: 10.1109/CVPR.2005.177