International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

Object Detection and Classification using Convolutional Neural Network

Khin Htay[1,*], Mie Mie Aung[2], Yin Cho[2], Moe Moe Thein[3]

[1,*]Faculty of Information Science
[2]Faculty of Information Technology Supporting and Maintenance
[3]Faculty of Computer System and Technology
[1*,2,3]University of Computer Studies (Meiktila), Myanmar

khinhtay.ucs@gmail.com

***Abstract**:*

Owing to the close relationship with the detection of objects in video learning and image recognition, many are attracted. Recent research has focused on traditional defining objects the methods are built on features of handmade trains and shallow architectures. The performance stalls easily making complex sets that combine multiple low image levels with high-level context from detectors and object views classifications with the rapid development of in-depth study, more powerful tools, we can learn semantics architecture in learning the physical things to detect. In this paper, deeper features are introduced to address the problems that exist in the area of physical object detection and classification against traditional architecture. The results show that our proposed model outperform

Keywords: Computer Vision, Object Detection, Convolution Neural Network

1. Introduction

To obtain a complete understanding of the image, we should not just concentrate on classifying different images, but also try to precisely estimate the concepts and locations of objects contained in each image. This task is called an object detection, which usually consists of different subtasks such as face detection, pedestrian detection and skeleton detection. As one of the fundamentals. computer vision problems, object detection is able to provide valuable information for the semantic comprehension of images. and videos, and it is related to many applications, including classification of images. analysis of human actions, face recognition and autonomous driving, recognition of facial expressions, etc. Meanwhile, inheriting from neural networks and related learning is performed in this area. In systems, progress in these fields will develop network algorithms, and will also have a great impact on the object Detection techniques that can be considered as learning systems.

In this paper, we propose a convolutional neural network, which is a kind of deep learning model for object detection. This paper only focusses on detection of humans although it also detects other objects. Nevertheless, the accuracy of human detection rate is much higher than that of other objects.

The rest of the paper is organized as follows. The related work is described in Section 3. Section 3 comes for background theory while the detailed proposed model is mentioned in Section 4. The results are discussed in Section 5 and the conclusion follows in Section 6.

2. Literature Review

Computer vision tool, many are focused on learning of the context and its relation with the discovery of things (Galleguillos and Belongie (2008) [1]; Divvala et al. (2009)[2]; Choi et al. (2010)[3]; Song et al. (2011)[4]). Galleguillos and Belongie [1] a survey that defines three types of context, semantic, spatial, and size context; acting on two levels, global and local; and shows the two mechanisms of integration of information in context.

They show that contextual information can help and successfully disambiguate appearance entries as recognition tasks. In Divvala et al. [2], the authors presented a review. The role of context plays in the use of object detection, the Pascal VOC 2008 data set. Understanding the context as any source of information that may have an impact on a scene and there are things in it, ten sources.

The context was identified, and the experiments were carried out throughout a subset of contextual sources: local pixel context, 2D scene scenes, 3D geometric, semantic, geographical, photogrammetric, and cultural suggestions A Bayesian formulation for the location of the object, size, and face, and their combinations, were used to improve the detection of objects based on the tracks. The work of Choi et al. [3] introduced the SUN data set, and presented a structured context model. Their models are a single graphic and represents positive and negative co-occurrences by data set.

The main weakness of its models is that it only captures the general information of the specific data set. Song et al. [4] proposing a Recurrent contextual system, with the aim of strengthening the detection of objects and categorization of images to the outputs of a task as another context with an SVM formulation and dynamic adjustment of the hyperplane classification. The key to the formulation of the hyperplane is that context activated to support most of the unclear points.

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

Machine learning techniques are widely used for computer vision resolutions Problems. Traditional machine learning techniques can solve simple identifying and identifying small datasets. As a dataset metric has been growing rapidly over the years, a large annotated image dataset, called ImageNet [5], was published. ImageNet contains over 15 million images and each image has an annotation labeled object category at the image level.

Due to strong learning capabilities, CNN begins to become popular strategy to model this amount of data. The first deep CNN, called AlexNet, proposed by Krizhevsky et al. [6] where outperformed tra- method of machine tone learning in ILSVRC-2010 [12] classification of objects benchmark. They implemented CNN on graphics processing unit (GPU) that allows network to train and tested efficiently. They also used a method of regularization, called dropout, which forced the model to know better features and helped to avoid excessive removal.

Due to AlexNet's success in image classification tasks, some researchers CNNs began to apply to the object detection system. Girshick et al. [7] proposes an object discovery framework, called RCNN, which is the first CNN based object detector. The strategy of RCNN is obtained high level of CNN features from the Interest Region (ROI) built by select search method. Then, a multi-class support vector machine (SVM) [8] is trained in CNN features to classify the item category of each item object. Even RCNN is acquired traditionally machine learning techniques in PASCAL VOC [9] dataset, it is a little overhead due to CNN's overhead application each proposed region.

3. Background Theory

4. Proposed Object Detection Model

A neural network consists of several different layers such as the input layer, at least one hidden layer, and an output layer. They are best used in object detection for recognizing patterns such as edges (vertical/horizontal), shapes, colors, and textures. The hidden layers are convolutional layers in this type of neural network which acts like a filter that first receives input, transforms it using a specific pattern/feature, and sends it to the next layer. With more convolutional layers, each time a new input is sent to the next convolutional layer, they are changed in different ways. For example, in the first convolutional layer, the filter may identify shape/color in a region (i.e. brown), and the next one may be able to conclude the object it really is (i.e. an ear or paw), and the last convolutional layer may classify the object as a dog. Basically, as more and more layers the input goes through, the more sophisticated patterns the future ones can detect.

The vision of the computer is around decades and there are many different uses in the real world.

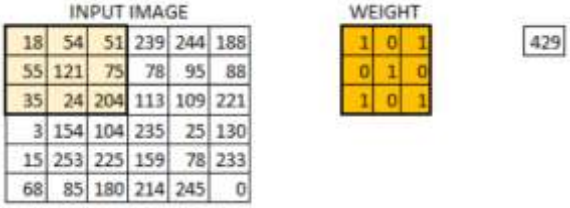
The whole concept of computer vision depends heavily on machine learning, especially today, and as new technologies emerge. When a machine has the purpose of classifying objects within an image, video, or real-time webcam, it should train with labeled data. This labeled data on object detection is the labeled pictures itself. Before the program training, we had to have lots of labeled pictures of specific things (in this case, we used 6 different types of cards and there were hundreds of pictures of them). This means that after model training, its goal is to identify specific types of cards in their labels in real-time. Now, when you actually train the program, you are trying to learn it through a supervised study. Because it is labeled images (the output), it usually learns through trial and error, continuously trying to predict the best outcome (identifying cards). The main component while the model is training is the loss function which is a constant measure of the accuracy of the model. The purpose of the training is to optimize the accuracy, so the loss of function keeps track of the error and the more the model trains, the less the loss will be! This is because the machine is constantly learning to make better decisions to reduce the error, just like us people, we are constantly learning to make ourselves better in achieving our goals.

In our convolutional neural network, we have three basic components to define a basic convolutional network.

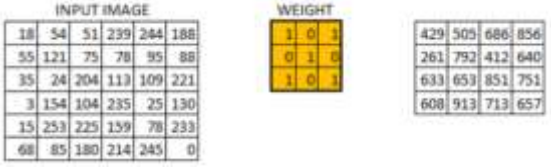
1. The convolutional layer
2. The Pooling layer, and
3. The output layers

4.1. Convolution Layer

In this layer, what happens is exactly what we saw in case 5 above. Suppose we have an image of size 6*6. We define a weight matrix which extracts certain features from the images.



(a) Step 1: extracting certain features for first pass



(b) Step 2: transformed features

Figure 1. Concept of stride and padding in convolution layer

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

The 6*6 image is now converted into a 4*4 image. Think of weight matrix like a paint brush painting a wall. The brush first paints the wall horizontally and then comes down and paints the next row horizontally. Pixel values are used again when the weight matrix moves along the image. This basically enables parameter sharing in a convolutional neural network.

The weight matrix behaves like a filter in an image extracting particular information from the original image matrix. A weight combination might be extracting edges, while another one might a particular color, while another one might just blur the unwanted noise.

The weights are learnt such that the loss function is minimized similar to an MLP. Therefore, weights are learnt to extract features from the original image which help the network in correct prediction. When we have multiple convolutional layers, the initial layer extract more generic features, while as the network gets deeper, the features extracted by the weight matrices are more and more complex and more suited to the problem at hand.

4.2. Pooling Layer

Sometimes, when the images are too large, we will need to reduce the number of trainable parameters. Then it is time to introduce grouping layers between the subsequent layers of the convention. The grouping is done with the sole purpose of reducing the size of the image. The grouping is done independently with each dim dimension; therefore, the depth of the image remains constant. The most common type of grouping layer commonly used is the maximum grouping.

4.3 The Output layer

After some layers of convolution and filling, we will need the output in the form of a class. Conventions and group layers can only extract attributes and reduce the number of parameters of the original images. However, in order to generate the final result, we need to apply to a fully connected layer to produce a result equal to the number of classes we need. It is difficult to reach that number with only the convention layers. Convolution layers create 3D activation maps, whereas we need only output if an image belongs to a particular class or not. The output layer has a function as the categorical cross entropy, to calculate the prediction error. Once the step progresses forward, back propagation begins to update the weight and biases for reducing errors and losses.

4.4 CNN Model works with those layers

The CNN you can see now consists of a variety of convolutional and pooling layers. Let's look at how the network looks.

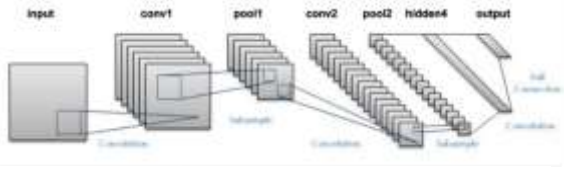


Figure 1. CNN Model [referenced by [16]

Our proposed CNN model for object classification is working with following steps.

• We pass an input image into the first convolutional layer. The unified output is obtained as a map of activation. Filters applied to the convolution layer capture the relevant features from the input image to forward more.

• Each filter should provide different features to assist the correct guess in the class. In case we need to keep the image size, we use the same padding (zero padding), other smart padding is used because it helps to reduce the number of features.

• Pool pools are added to further reduce the number of parameters

• Several conventions and pooling layers are added before prediction is made. The convolutional layer helps to extract features. As we get to the network, the features are more precise than a superficial network where the features obtained are more common.

• The output layer on a CNN as mentioned previously is a fully connected layer, where inputs from other layers are pressed and transmitted so the output transforms to the number of classes as desired by the network.

• The output is then generated by the output layer and compared to the output layer for error generation. The loss function is defined in the fully connected output layer to calculate the mean square loss. The error gradient is then calculated.

• The error will then be backpropagated to update the filters (weights) and bias values.

• A training cycle completed with a single forward and backward pass.

5. Discussion and Simulations

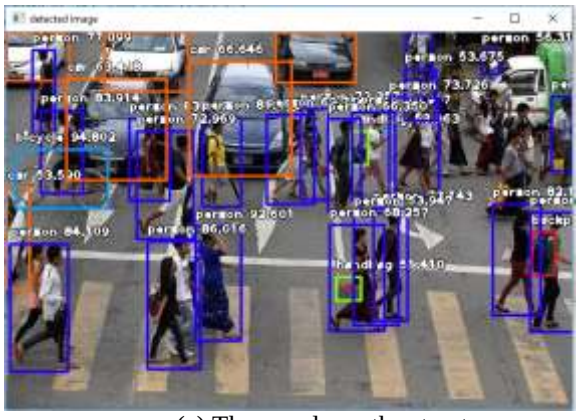
Our network is trained and tested on a personal computer with single NVIDIA X GPU. For the datasets, in order to fairly compare our model to original SSD, we follow the training setting of the dataset which splits the validation set into two sets; one contains 4000 images and the other contains 15,000 images. Then, we add a 15,000-image set into the training set for training.

5.1 Development Setting

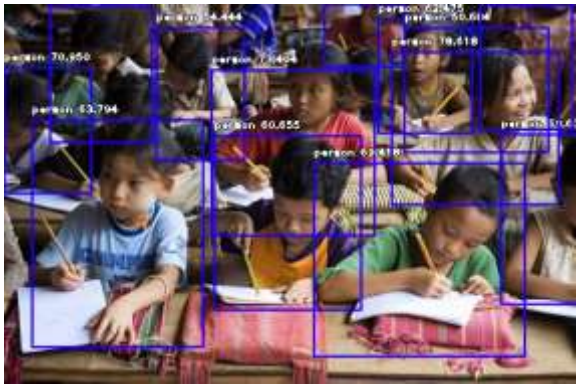
The proposed system is implemented with Python Programming and uses TensorFlow model combination with our proposed CNN model. According to the results

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

shown in Figure 2, we achieve superior detection results with accurate results.



(a) The people on the street



(b) the children in the classroom
Figure 2. The development results

6. Conclusions

In this paper, we had proposed an object detection and classification model with cutting-edge technology called convolution neural network with three layers. The detailed process of each layer has been discussed. The system was implemented with Python programming and we gain better promising results in detection in different kinds of objects. As future plan, we like to extend this work with better alternative techniques to dig better results in both images and videos.

7. References

[1] Galleguillos, C., Belongie, S., 2008. Context based object categorization: A critical survey, in: ECCV.

[2] Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A., Hebert, M., 2009. An empirical study of context in object detection, in: IEEE Conference on Computer VIsion and Pattern Recognition (CVPR).

[3] Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S., 2010. Exploiting hierarchical context on a large database of object categories, in: IEEE Conference on Computer VIsion and Pattern Recognition (CVPR).

[4] Song, Z., Chen, Q., Huang, Z., Hua, Y., Yan, S., 2011. Contextualizing object detection and classification, in: CVPR.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 248–255.

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 2012, pp. 1097–1105.

[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 580–587.

[8] C. Cortes, V. Vapnik, Support vector machine, Mach. Learn. 20 (3) (1995) 273–297.

[9] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[10] T. Kong, A. Yao, Y. Chen, F. Sun, HyperNet: towards accurate region proposal generation and joint object detection, IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 845–853.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 2921–2929.

[12] L. Herranz, S. Jiang, X. Li, Scene recognition with CNNs: objects, scales and dataset bias, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 571–579.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[14] J. Redmon, A. Farhadi, YOLO9000: Better, Faster, Stronger, arXiv preprint arXiv:1612.08242 2016.

[15] X. Zeng et al., Crafting GBD-net for object detection, IEEE Trans. Pattern Anal. Mach. Intell., vol. PP, no. 99, 1.

[16]https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/