OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

DESCRIBING IMAGE USING DEEP LEARNING

Shreyas Ghosalkar[1], Rajashri Jagadale[2], Kranti Ranpise[3], Ronak Chabukswar[4], Prof. V.T. Pokale[5]

Student, Computer Engineering, VIIT, Pune, India[1 2 3 4]

Asst. Professor, Computer Engineering, VIIT, Pune, India[5]

shreyasghosalkar@gmail.com, rajshree.jagdake25@gmail.com

ABSTRACT: **Describing an image is an effortless task for human beings. But achieving the same by computers, without the aid of humans is a challenging task. To correctly describe an image, accurate recognition of objects, their attributes, relationships and scene information is required. This idea of describing contents of an image can have a lot of use in social networking apps or software that want to process visual data. In this paper, we aim to generate sentences based on images using deep neural networks. We have used MS-COCO dataset to train the Convolutional Neural Network, and have used PASCAL VOC to fine-tune it. The labels are everyday occurring objects. The sentences are generated using a fixed template. It detects maximum of four objects and describes the localization between them [above, below, left and right]. This system achieves an accuracy of 72.7\% when evaluated against a subset of MS-COCO dataset for object classification.**

KEYWORDS: *Object Classification, Deep Learning, Convolutional Neural Network, Batch Normalization, ReLU*

I INTRODUCTION

Humans can easily process visual information they see in the surrounding and infer upon its contents. Humans can infer and conclude certain things about what is seen from a very young age. But it is difficult for computers and systems to identify such contents and infer accurately. But current researches and development in the image processing have provided powerful algorithms that can be used to train the machines to understand an image like humans do. This as a model can be further used on large scale applications like robotic vision. Robots can effectively interact with humans by being more aware about the surrounding using this technology. Also this can be used for efficient searching of visual data from large databases.

The proposed system consists of 4 main stages:

1. Detecting number of Objects
2. Object classification
3. Localization
4. Sentence Generation

In the first step the system collects machine learned holistic image features from large scale Convolutional Neural Networks (CNN) that are computed at test time to classify the scene . In second step, object detectors are implemented with the use of deep CNN to classify people and objects within an image, and provide bounding boxes and an object detection label. The network consists of 3x3 Conv net with Batch normalization and ReLU. In the third step we find the location of these objects, each having startX, startY, and endX, endY points. In the last step a template-based sentence generation method is applied to concatenate the previous outputs into one or multiple natural language sentences.

II LITERATURE SURVEY

The two major techniques for Object Detection are SVM[Support Vector Machine] and Neural Networks.

SVM:

Support Vector Machine is often considered as the state-of-art technique for image classification. For binary classification, it works in the following way: Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The hypothesis is set such that it has maximum distance from the nodes in either of the groups. This hypothesis is then used for the binary classification. For multiple classes, one v/s all binary method is applied, which results in many hypothesis, thus dividing the space into multiple hyperplanes. This method requires very little training time, but it continues to grow as we add more and more training examples. The speed and size of SVM is a major limiting factor. They can sometimes perform very poorly on large scale classification in terms of speed.

CNN:

During the recent years, the quality of image classification and object detection has been dramatically improved due to the deep learning method. Convolutional neural networks (CNNs) brought a revolution in the computer vision area. It not only have been continuously advancing the image classification accuracy, but also play an important role for generic feature extraction such as scene classification, object detection, semantic segmentation, image retrieval and image caption .Convolutional neural network (CNNs) are one of the most powerful classes of deep neural networks in image processing tasks. It is highly effective and commonly used in computer vision applications.

III METHODOLOGY

The main stages of the system are discussed below:

1. Object Detection

When building object detection networks we normally use an existing network architecture, such as VGG or ResNet, and then use it inside the object detection pipeline. The problem is that these network architectures can be very large in the order of 200-500MB. Network architectures such as these are unsuitable for resource constrained devices due to their sheer size and resulting number of computations. This limits our applications as some of these require real time, fast detection of objects, as in the case of automated driving. Hence, we use MobileNets[1]. We call these networks “MobileNets” because they are designed for resource constrained devices such as your smartphone. MobileNets differ from traditional CNNs through the usage of depth wise separable convolution. The general idea behind depth wise separable convolution is to split convolution into two stages:

1. A 3×3 depth-wise convolution.
2. Followed by a 1×1 point-wise convolution.

This allows us to actually reduce the number of parameters in our network. However, this comes at a cost. This reduces the accuracy of MobileNets. MobileNets are usually not as accurate as other big Networks. We thus trade accuracy for speed. The MobileNet SSD was first trained on the COCO dataset[2] (Common Objects in Context) and was then fine-tuned on PASCAL VOC reaching 72.7\% mAP (mean average precision) By combining both the MobileNet architecture and the Single Shot Detector (SSD) framework, we arrive at a fast, efficient deep learning-based method to object detection.

With this network, we can detect 20 objects [+1 background], namely: airplanes, bicycles, birds, boats, bottles, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, people, potted plants, sheep, sofas, trains, and tv monitors.

2. Object Localization and Sentence Generation After obtaining the tags[classes] for various objects in the above step, we find their start and end points[co-ordinates].

This information is used to locate the objects in the image with respect to each other. Using their start and end points, we draw bounding boxes around them. These bounding boxes help us to determine if the objects are overlapping or separate from each other. If they are overlapping, then we can find the relation between them, as well as the preposition. If not, then we consider them as separate objects. After detecting the objects, and finding suitable relation between them, we then proceed to generate the required sentence. We have assigned the objects priorities ranging from 1-3. Background has lowest priority (ie, 3). The objects that come below other objects have priority 2(car, train, bicycle, etc), while human, cats, dogs have priority 1. Based on priority and localization of objects, we generate the required sentence.

IV RESULTS

This model is a Caffe version of the original Tensor Flow implementation by Howard et al. and was trained by chuanqi305. This model achieves an accuracy of 72.7% mAP. Though this model has limited functionality, this is a fast implementation for Object detection and sentence generation. Also, this has various applications in low power devices such as our smartphones, that have limited performance and battery associated with them.

V CONCLUSION AND FUTURE SCOPE

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard back propagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence.

In which we can Automatically generating captions for an image is a task close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language.

REFERENCES

[1] Andrew Howard, Menglong Zhu, et.al, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017

[2] Tsung-Yi Lin, Michael Maire et.al, Microsoft COCO: Common Objects in Context, 2014

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In ICML, 2014. 3

[5] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 2013. 2

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010. 1, 2, 5

[7] R. Gerber and H.-H. Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996. 2

[8] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014. 2, 5

[9] A. Graves. Generating sequences with recurrent neural networks. arXiv:1308.0850, 2013. 3

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8), 1997. 2, 3