Interactive Object Identification using Image Recognition

Jayanthi Palanichamy[1] Duggi Meghana[2] Harshini. S[3]
[1,2,3]Department of Computer Science & Engineering
[1,2,3]R.M.K Engineering College Chennai – 601206

Abstract— This project is proposed to help visually challenged people. It manipulates to image processing and natural language processing techniques to simulate human vision. The proposed system takes images and other multimedia files and forms logical sequence between them and help the users understand it. We focus on recognizing human actions in still images, which is done by analyzing human poses and their interaction with objects. A systematic approach of recognizing objects and their relationship is developed through this project. In the image processing technology, automatically generating a natural language description of an image is an important task. A multi-model neural network system is used in this paper which describes the content of images automatically. This multi-model neural network is divided into an object detection and localization model, which extract the information of objects and their spatial relationship in images respectively. Sentences generation is done by using long short-term memory (LSTM) units with attention mechanism in a deep recurrent neural network (RNN). This can also be used to provide highly sophisticated search and in forensic systems since mining data from multimedia files greatly improves the search and data analytics.

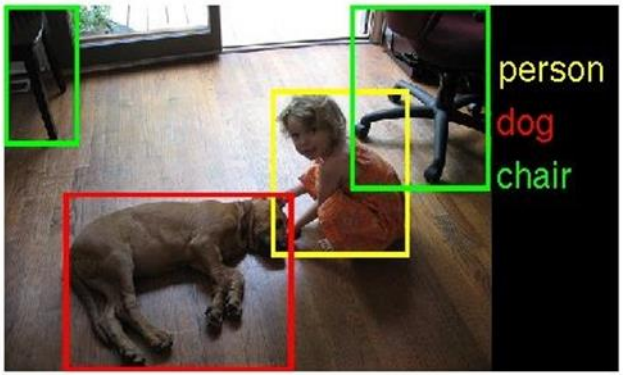
Key words: Image Processing; Neural Networks; Machine Learning; Natural Language Processing; Computer Vision; Convolution Neural Network; Recurrent Neural Network; Deep Learning; Long Short Term Memory Units

I. INTRODUCTION

The idea is to paint the scene in front of the user. Instead of relying on humans to guide them, visually challenged people can use Computer vision. Computer vision is used to identify the objects in the scene and describe it to the user. This software takes photo from the smartphone and uploads into cloud to process information in the image. Our ultimate goal is deep understanding of image i.e Understanding the whole relationship between images thus identifying the scenario not individual objects. Thus Image captioning follows the following path: extracting the complete detail of individual object and identifying their associated relationship from image. Finally, sentences are generated automatically by the system to describe the image. This problem is extremely significant, as well as hard because it involves two major artificial intelligence fields: computer vision and natural language processing. When the machine learning system identifies the objects, actions and relationship between them, it returns the result in text and audio format. This system can identify multiple objects in the image and also the relationship between them. User can also ask information about color, shape and other properties of the object. This has wider range of applications even if it aimed at visually challenged people. This can be used to improve search engines as mining data from multimedia like images and videos greatly improves the search and data analytics. It can be used to identify real world objects around us from children satiating their curiosity when they encounter unfamiliar objects to offering field guide experience for bird watchers.

II. EXISTING SYSTEM

The existing image recognition system by google detects individual objects and faces within images and finds and reads printed content within images. This system cannot recognise relationships between objects in the image. It can only identify the objects. Example it can be able to specify a person is riding a bike. The existing system can spew out words like 'person' and 'bicycle'. Though it has many features like extracting data from images and identifying the color of the objects. It has no proper implementation to aid visually challenged people. The other challenges are it takes long time time to process information from cloud as processing information locally on the device is very slow. Many computer vision-based assistive systems for the blind have tackled the problem of environment sensing and understanding [27], e.g. in the system reported in [28] semantic maps for indoor spaces were used to support high level localization, navigation and context awareness. However, very few of the assistive systems consider the pervasiveness aspect [8, 22, 29] and work either in indoor or outdoor environments.



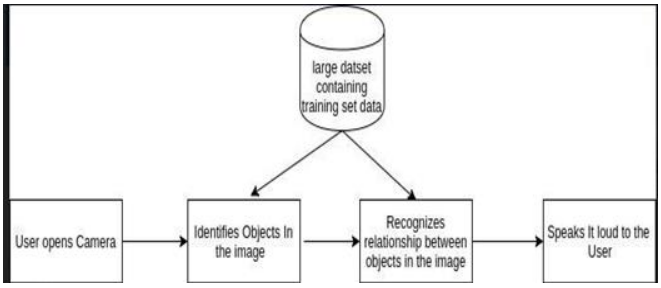
The above image describes the working of the existing system. It correctly identifies various objects in the image. But it can't identify the relationship among them. Thus the following system is proposed,

III. PROPOSED SYSTEM

The proposed system recognises the logical relationship between the objects and it can be used to tell stories with images. It speaks out loud the objects and their relationship to the user. The user can converse with the module and ask questions. The system can discover multiple objects in the scene such as table,dog and the child,it can also understand the relationship between them.

The various modules are

- Object Identification and Localization
- Relationship Identification
- Voice Interaction module



A. Architecture of Modules

The above "Fig 1" represents the work-flow of the project. The user opens the camera and takes the picture of the scene to be recognised. The object recognition module uses training set data to recognize objects in the module. The relationship Identification module finds relationship between various objects. Finally the voice interaction module forms UI for the visually challenged people.

1) Object Identification and Localization

It identifies the objects in the scene and their physical properties like color , shape,etc. It uses neural networks and training data set to identify the objects in the image.Neural Network (NN) is chosen as a classifier tool due to its well-known technique as a successful classifier for many real applications. Neural Network is selected as classifier because it gives high accuracy and basically used for nonlinearity detection.

Significant progress has been made in object detection in last few years. The success in RCNN (region-based convolutional neural networks) and Recurrent Neural Networks helped in this significant progress [3]. We use Faster R-CNN for object Detection because of its efficiency and effectiveness. Faster R-CNN is composed of two modules. Deep fully convolutional network that propose regions is the first module, the second module is the Fast R-CNN detector [3] which uses the proposed regions to detect objects. The last shared convolutional layer outputs convolutional feature map by sliding a small network over it to generate the proposed regions [7]. This small network maps for each sliding window of the input convolutional feature map, small network to a lower-dimensional feature. Two sibling fully-connected layers: a box-classification layer(cls) processes these features and a box-regression layer (reg). After training, the Faster R-CNN produce a set of rectangular object proposals, with an objectness score from an input image (of any size). To choose the top-n boxes as the regions of objects in the input image, the above produced rectangular boxes are sorted according to their scores in descending order. The Fast R-CNN model has a fully connected layers which maps feature vectors to each rectangular object region. More explicitly, n objects are detected in every input image and each object is represented as a d-dimension vector:

$$\{ obj_1 , obj_2 , \dots , obj_n \}\ obj_i \in d^d$$

To identify the spatial relationships between objects, object Localization part is designed to extract the information of objects spatial locations. Junqi etal. [4] has also used the locations of different localized regions to derive the annotations. They have just added the boxes central's with width, height x location, y location and area ratio with respect to the entire image's geometry to the end of the vector of each localized regions. In this paper, the implementation of extracting information of each object location is completely different from Junqi etal. [4]. In object detection part, we know that for each input image, the output is n rectangular object regions, each with an objectness score. For each object in this image, we keep the region of its bounding box unchanged and set remaining regions to mean value of the training set. So we get a new image, which has the same size as the original image but just consists the bounding box region of one object as shown in Figure 1. As we detected n objects for each image therefore, we get n new images for individual image. Each new image will then be fed into the VGG net [2]and the feature vector of its 'fc7' layer will be extracted, which yields to the vectorized representation of object location. Furthermore, we get another n vectors of t-dimension in which each vector represents the information of spatial location of each object:

$$\{ loc_1 , loc_2 , \dots , loc_n \},\ loc_i \in d^t .$$

Each annotation vector A i consists of two parts: First, vector obj_i represents the feature of object which is used to particularly describe the contents of image. Second, vector loc_i represents the feature of object location which gives us information about the location of individual object.

$$A_i = [obj_i ; loc_i],\ A_i \in d^D ,\ D = d + t.$$

B. Relationship Identification

This module identifies relationship between various objects in the given image. This is done by using natural language processing and further image processing. To recognize motions of objects.(verbs). Take a specific dance sequence. Have the machine see and learn many dance sequences. Make sure the differences between each of them is learned and also the differences between similar motions which are not dances such as walking and running etc. are known. The point where the machine can recognize a new dance that it never saw it understands the concept of dance. The concept of generalization is more completely and thoroughly understood. It is possible a machine might generalize concepts we cannot understand.

The first subproblem involves on how to delineate the detailed shape of human-object interaction regions (i.e., the action mask). Subsequently the second subproblem concentrates on proper feature representation for the recognition task.

The input to our model is a single image I, while the output is a descriptive sentence S consists of K encoded words:

$S = \{w_1 , w_2 , \dots , w_K\}$ .In the encoding part, firstly, we present a model that recognizes objects in the input image followed by a deep CNN to extract their locations, which reflect the spatial relationship associated. All the information will be represented as a set of feature vectors referred as annotation vectors. The encoding part produces L annotation vectors, each of which is a D-dimensional representation corresponding to an object and also its spatial location in the input image: $A = \{A_1 , A_2 , \dots , A_L\}\ A_i \in R^D$

This section describe the training of proposed model. The training data for each image consists of input image features $\{A_i\}$ and output caption words sequence $\{w_k\}$

Parameters of the proposed encoding part is fixed, so we only need to learn the parameters of the proposed decoding part, which are all the attention model parameters $\Theta_{Att} = \{W_{.}, U_{.}, Z_{.}, b_{.}\}$ jointly with RNN parameters Θ_{RNN}. We train our model using maximum likelihood with a regularization term on the attention weights by minimizing a loss function over training set. The loss function is a negative log probability of the ground truth words $w = \{w_1, w_2, \ldots, w_K\}$:

$$LOSS = -\sum_t \log(p(w_j)) + \lambda \sum_i (1 - \sum_t \alpha_{i,j})$$

Where w_j is the ground truth word and $\lambda > 0$ is a balancing factor between the cross entropy loss and a penalty on the attention weights. We use stochastic gradient descent with momentum 0.9 to train the parameters of our network.

we describe a decoding part based on an LSTM network with attention mechanism. Attention mechanism was first used in neural machine translation area by [5]. Following the same mechanism, the authors of [6,8,9] introduced it into image processing domain whereas, [9] was the first to apply it in image captioning task. The key idea of attention mechanism is that when a sentence is used to describe an image, not every word in the sentence is "translated" from the whole image but actually it just has relation to a few subregions of an image. It can be viewed as a form of alignment from words of the sentence to subregions of the image. The feature vectors of these subregions are referred to as annotation vectors. Here in our implementation, subregions are referred to as the bounding box of objects and annotation vectors are referred to as $\{A_i\}$,which is already discussed in the encoding part.

C. Voice Interaction Module

This module forms the user interface of the project. It helps to converse with user using voice control. The user can hear the object before them, they can also ask questions about it. We use WaveNet to build the interactive system because it facilitates modelling the raw waveform of the audio signal directly, one sample unit at a time. This means the wavenet can model any kind of audio data including music using raw waveforms. It can also produce more natural-sounding speech. It is shown that WaveNets can be able to generate speech which mimics any human voice and accent. It also sounds more natural than the other best existing Text-to-Speech systems, reducing the gap with human performance by over 50%

IV. CONCLUSION

In this paper, we present an interactive system which can identify not only the objects but also the relationship between them when the live image data is fed into it. It designed to aid visually challenged people by including voice interactive system. This model uses multi model Neural Network which can automatically learn and describe the content of given images. The spatial locations of the objects in an image and other information about the objects is extracted and given to a deep recurrent neural network (RNN) which uses LSTM units to generate descriptive sentences with attention mechanism. When the description is generated each word in it is automatically aligned to different objects in the input image. The user can prompt the system using voice commands to get detail about scene before them.The proposed model is more optimized compared to other benchmark algorithms on the ground that its implementation is totally made on human visual system.

REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.
[4] Jin J, Fu K, Cui R, et al. Aligning where to see and what to tell: image caption with region based attention and scene factorization[J]. arXiv preprint arXiv:1506.06272, 2015.
[5] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
[6] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in neural information processing systems. 2014: 2204-2212.
[7] Ren, S.; He, K.; Girshick, R.; Sum, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Neural Information Processing System (NIPS), Montreal, QC,Canada, 7–12 December 2015; pp. 1–9.

[8] P. Chippendale, V. Tomaselli, V. D'Alto, G. Urlini, and C. Modena. Personal shopping assistance and navigator system for visually impaired people. In Proc. of the CVPR 2014 Workshop", 2014
[9] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv preprint arXiv:1412.7755, 2014.
[10] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. 2015: 2048-2057.
[11] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks(m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.
[12] Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448
[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997,9(8): 1735-1780.
[14] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.
[15] Bahadur, A.K.; Tripathi, N. Design of Smart Voice Guiding and Location Indicator System for Visually Impaired and Disabled Person: The Artificial Vision System, GSM, GPRS, GPS, Cloud Computing. IJCTER 2016, 2, 29–35
[16] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision. Springer International Publishing, 2014: 740-755.
[17] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
[18] M. Bujacz, P. Skulimowski, and P. Strumillo. Naviton - a prototype mobility aid for auditory presentation of three dimensional scenes to the visually impaired. Journal of the Audio Engineering Society, 60(9):696-708, 2012.
[19] S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, A. Serpa, M. Raynal, B. Oriola, M.-M. Mac, M. Auvray, M. Denis, S. Thorpe, P. Truillet, B. Katz, and C. Jouffrais. Navigation and space perception assistance for the visually impaired: The {NAVIG} project. {IRBM}, 33(2):182 – 189, 2012. {ANR} {TECSAN Technologie} pour la santet l'autonomie.
[20] D. E. King. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10:1755–1758, 2009.
[21] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In Intelligent Vehicles Symposium (IV), 2010.
[22] T. Kurata, M. Kourogi, T. Ishikawa, Y. Kameda, K. Aoki, and J. Ishikawa. Indoor-outdoor navigation system for visually impaired pedestrians: Preliminary evaluation of position measurement and obstacle display. In Wearable Computers (ISWC), 2011 15th Annual International Symposium on, pages 123–124, June 2011.
[23] Y. H. Lee, T.-S. Leung, and G. Medioni. Real-time staircase detection from a wearable stereo system. In 21st International Conference on Pattern Recognition (ICPR 2012), pages 3770–3773, 2012.
[24] M. Auvray, S. Hanneton, and J. K. O'Regan. Learning to perceive with a visuo - auditory substitution system: Localisation and object recognition with 'the voice'. Systems Journal, IEEE, 36(3):416–430, 2007.
[25] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.
[26] Van de Sande, K.E.A.; Uijlings, J.R.R.; Gevers, T.; Smeulders, A.W.M. Segmentation as Selective Search for Object Recognition. In Proceedings of the International Conference Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1879–1886.
[27] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. Farinella. Computer vision for assistive technologies. Computer Vision and Image Understanding, 154:1 – 15, 2017.
[28] B. Li, P. Muoz, X. Rong, J. Xiao, and Y. T. nad Aries Arditi. Isana: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind. In Lecture Notes in Computer Science, volume 9914, pages 448–462, 2016.
[29] L. Ran, S. Helal, and S. Moore. Drishti: An integrated indoor/outdoor blind navigation system and service. In IEEE International Conference on Pervasive Computing and Communications, pages 23–32, 2004.

