SCENE DESCRIPTION FROM IMAGES TO SENTENCES

Khushali Acharya , Abhinay Pandya

[1,]Computer Engineering
LDRP-ITR
Gandhinagar, India
[2]Prof & HOD, Information Technology
LDRP-ITR
Gandhinagar, India

---***---

Abstract— *People exchange their views using language, whether spoken, written, or typed. A notable amount of this language describes the environment around us, especially the visual scenario in our surroundings or depicted in images or video. Scene description aims to generate the sentences from given set of input images. It links the visual perception with the language space. All present approaches are purely found in supervised machine learning setup. However, owing to the dearth of training data, this seldom achieves desired accuracy. We present a model that uses "Distributed Intelligence" as the prevalent theme in the artificial intelligence literature. Rather than only relying on the training dataset (PASCAL VOC 2012 containing 11530 images, FLICKR8K, FLICKR30K & MSCOCO), we harness the power of internet in order to generate more precise sentences related to the images.*

Keywords—Image Processing, Opencv, Distributed Intelligence, Object Detection

1. INTRODUCTION

"A picture is worth a 1000 words" WRONG! Why?
We see in our daily life how important words are to us. If a picture is enough then everyone must wonder why these memes we see on Instagram do and Facebook daily have sentences written down. There are hashtags, descriptions, comments etc. to make the post more interesting. Thus, description of an image makes it more compelling and captivating. We understand image in our brains with processes that we do not know enough about. For example, have a glance at an image below.

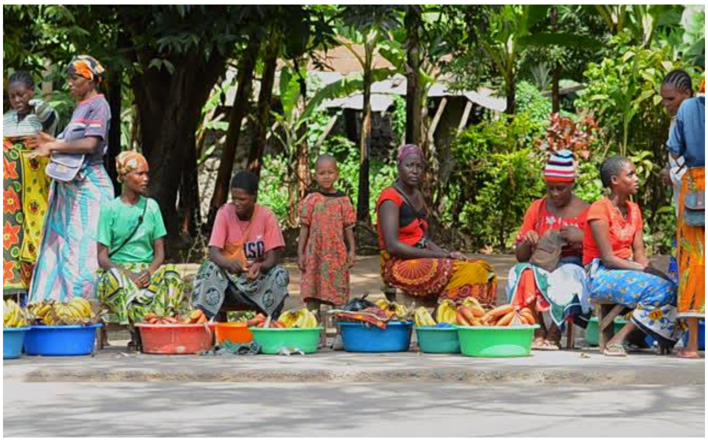

Fig 1: A group of tribal people selling vegetables in market.

One can describe this image as:

- A group of people sitting on land.
- A group of people selling bananas/fruits.
- A group of tribal people selling fruits in a local market.

From above sentences, the third sentence seems the most acceptable. Will the answer be same if asked to a 7-year old? Of course no. How would you communicate with Aliens? How would you teach them your language? The way we teach children right? By showing them pictures of various objects and their names. In order to make a meaningful sentence some previous knowledge is must in case of humans as well as computers. Thus in order to transfer information unambiguously, we must attach textual cues to support its correct interpretation/view-point.

Scene Description can be helpful in various fields. Some of them are listed below:

- Robotics
- Content-based image search
- For visually impaired people
- Soccer game analysis
- Criminal Act Recognition
- Agricultural Sector

2. RELATED WORK

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015) [1]:

This paper trains the model in a deterministic manner using standard back propagation techniques. Two attention-based caption generation are introduced under a common framework: 1) a "soft" deterministic attention mechanism trainable by standard back-propagation methods and 2) a "hard" stochastic attention mechanism trainable by maximizing an approximate variational lower bound or equivalently by REINFORCE (Williams, 1992).

Deep Visual-Semantic Alignments for Generating Image Descriptions (2015) [2]:

Their approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. It is a combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. They first present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. Then they treat these correspondences as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets.

Choosing Linguistics over Vision to Describe Images (2012) [3]:

Given a collection of images and their corresponding human-generated descriptions, they address the problem of automatically generating human-like descriptions for unseen images. This method captures the semantics of an image using information coded in its description and provide extensive evaluations to test the applicability of their model on IAPR TC-12 benchmark.

Baby Talk: Understanding and Generating Simple Image Descriptions (2013) [4]:

This system automatically generates natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from computer vision. This is accomplished by detecting objects, modifiers (adjectives), and spatial relationships (prepositions), in an image, smoothing these detections with respect to a statistical prior obtained from descriptive text, and then using the smoothed results as constraints for sentence generation. Sentence generation is performed either using a n-gram language model or a simple template based approach.

3. OUR APPROACH

1) Object detection and labelling

The very first step in our approach is to detect the objects from the image. Our goal in this section is to get the bounding boxes on the recognized objects along with the co-ordinates of it. For this purpose we are going to use the YOLO [5] Detection system which uses the Darknet [6] framework.

YOLO has 24 convolutional layers followed by 2 fully connected layers. While GoogleNet uses inception modules, YOLO simply use 1×1 reduction layers followed by 3×3 convolutional layers, similar to Lin et al [7]. A single convolutional neural network performs feature extraction, bounding box prediction, nonmaximal suppression, and contextual reasoning all concurrently. Instead of static features, the network trains the features in-line and optimizes them for the detection task.

2) Find Relative Position of these Objects

The next step is to find the relative position of the object with respect to another object and background. If there are N objects detected in our image, then there will be nC2 relations between these objects. Simple prepositional functions that evaluate the spatial relationships between bounding boxes are designed which provide a score for each of 16 preposition terms (e.g. above, under, against, beneath, in, on etc.). For example, the score for 'above (a, b)' is computed as the percentage of region 'a' that lies in the image rectangle 'above' the bounding box around region 'b'. The potential for 'near (a, b)' is computed as the minimum distance between region 'a' and region 'b' divided by the diagonal size of a bounding box around region 'a'. Similar functions are used for the other preposition terms.

3) Find a text label for the background scene

We need to find the text label for the background scene in our next step. This is done by eliminating the detected objects from the image. The remaining scene in the image will be considered as a background. This process is done by using Multiclass Support Vector Machines trained by using various background scenes.
The primary process of eliminating the objects is done by applying the inverse operation to the output of the grabcut algorithm. Originally grabcut is a foreground extraction algorithm but we transpose its result in order to get the background. In this algorithm, user needs to draw rectangle around the foreground area. Then algorithm segments it iteratively to get the best result.

After separating the foreground and background, its transpose operation is performed which leads to the background image. Next we are going to use the concept of support vector machine. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier's detection.

4) Populate a tuple for that image

As of now, we have the three things with us, the detected objects, the label of the background and spatial relationship between them. These three entities will form a tuple of the form, <obj 1, obj 2…obj n, Scene, Relations>. There can be multiple detected objects in the scenario and Scene is the Background labelling. Relations here represent attributes such as directions, background-foreground relations, touching, non-touching etc.

5) Find from the parallel corpora or word co-occurrence data, the most likely set of sentences containing these words

By now we have got the tuple of the objects, scene and the relationship between them. But these individuals are not enough to generate a meaningful sentence. In order to generate a relevant sentence, we are going to use the notion of "Distributed Intelligence".

We are going to infuse the knowledge source to the problem statement in order to solve it. Infusing knowledge means providing dataset to the computer. The dataset used here is Wikipedia. Based on this dataset, an ontology will be populated. This idea is not hand crafted, it is automatic. So this is statistically probabilistic graphical model. It is an undirected graphical model much in the lines of Conditional Random Field (CRF) or Markov Random Field (MRF).

The tuple <obj 1, obj 2…obj n, Scene, Relations>, is compared with the Wikipedia dataset. Top 5 sentences with the combination of words in the tuple are extracted. We have prior knowledge in the form of Wikipedia sentences. So we will have likelihood to be calculated based on the given prior information and after integrating them in the Bayesian framework, the final answer is obtained.

6) Perform Google image search and find relevant images for each such sentence

There are 5 most relevant sentences extracted from the Wiki dataset containing the words of the tuple. The Google Image Search API provides a JavaScript interface to embed Google Image Search results in your website or application. For each one of the 5 sentences, the following steps are performed:

- Give input as the sentence retrieved from above step to Google Image API.
- Download top 10 images for each such sentence.
- Compare each of the 10 images with the original image and generate a similarity score.
- The highest matching image with the actual image is taken and kept in the table along with its score.
- This process is repeated for all 5 sentences and thus we get 5 images with its score in the form of a table.

7) Compare retrieved images with the input image and assign similarity score

The downloaded 10 images for each sentence are compared with the actual image. Various algorithms are used for comparison such as key point matching, Histogram methods, feature matching techniques with decision trees etc

One can also use distance metrics such as ***Euclidean*** distance, ***Manhattan*** (also called City block) distance, and the ***Chebyshev*** distance. A similarity measurement must be selected to decide how close a vector is to another vector. The problem can be converted to computing the discrepancy between two vectors x, y ∈R^d. The Euclidean distance between x, y ∈ Rd is computed by:

$$\delta(x,y) = \|x-y\| = \sqrt{\sum_{j=1}^{d}(x-y)^2}$$

The Manhattan (city block distance), which takes fewer operations, is computed by:

$$\tau(x,y) = \|x-y\| = \sum_{j=1}^{d}|x-y|$$

8) Find the image which has the highest similarity score with the input image. The sentence corresponding to this image is most likely the one for our image too.

Till now, we have achieved similarity scores of all the retrieved relevant images. Of all the 5 sentences and their relevant 10 images per sentence, those image with highest similarity score or least distance metric will be picked up. The sentence for the highest scorer image will be the one for actual image too.

The RGB Color Histogram of actual image and all other images are generated. Then Euclidean distance between all the images is calculated. The image having least distance with actual image is taken as intermediate result. The corresponding sentence with this intermediate result is the original sentence which describes the scene or the image taken as input.

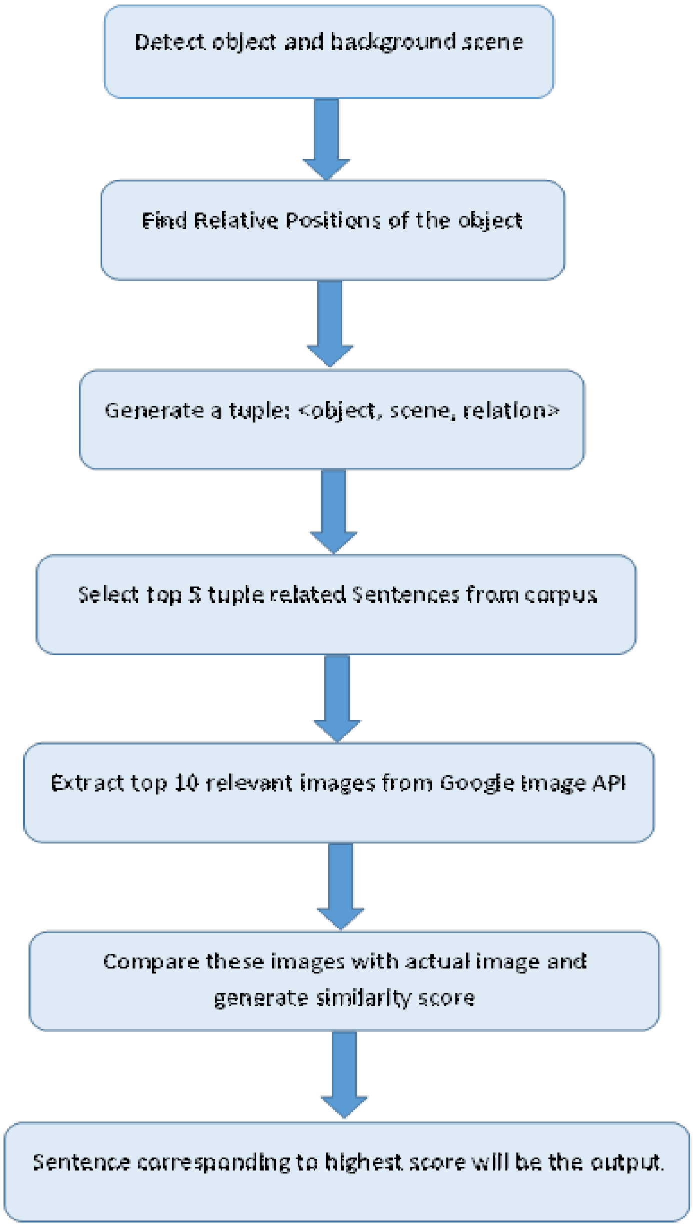
Detect object and background scene
Find Relative Positions of the object
Generate a tuple: <object, scene, relation>
Select top 5 tuple related Sentences from corpus
Extract top 10 relevant images from Google Image API
Compare these images with actual image and generate similarity score
Sentence corresponding to highest score will be the output.

Fig 2: Architecture of our approach

4. EXPERIMENTS AND RESULTS

As the first step of our implementation, we start with detecting objects from the image. According to [5], ImageNet 1000-class competition dataset and PASCAL VOC detection dataset is used.

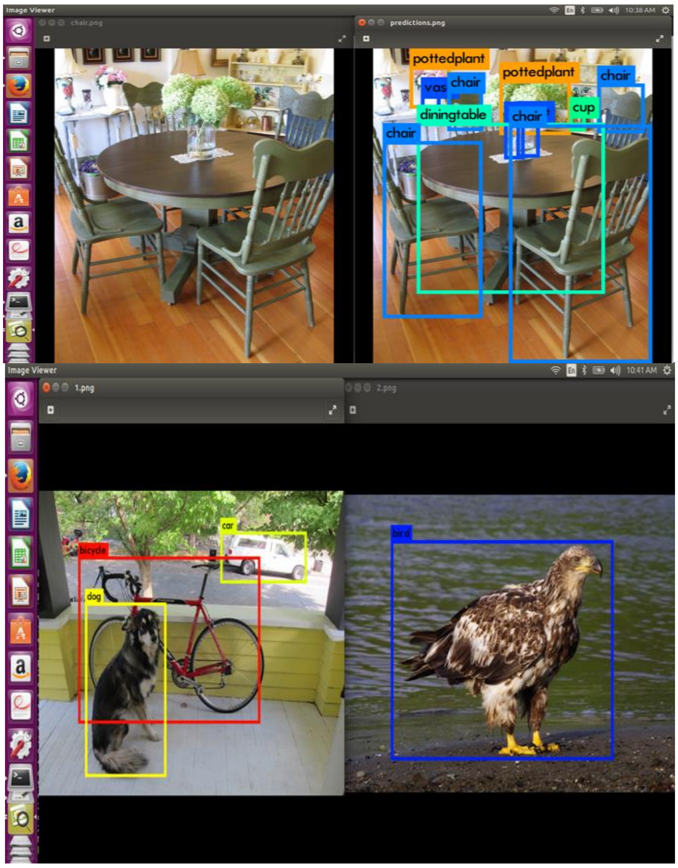

Fig 3: Object Detection and Labelling

There are multiple bounding boxes detected in the above step. (x,y) coordinates of all the bounding boxes are detected along with the probability of existence of that object. From its coordinates, the spatial relationship between the objects can be established.

We are going to compare our result with our base paper that is Babytalk. Comparison is based on two systems: BLEU and ROUGE.

BLEU Score Measured for Generated Descriptions versus the Set of Descriptions Produced by Human Annotators

No. of Images	Babytalk [4]	Our Approach
100	0.40	0.42
200	0.45	0.37
500	0.38	0.37

ROUGE Score Measured for Generated Descriptions versus the Set of Descriptions Produced by Human Annotators

No. of Images	Babytalk [4]	Our Approach
100	0.43	0.44
200	0.41	0.33
500	0.36	0.38

5. DISCUSSION ON RESULTS OBTAINED

We compare our results with only [babytalk] since their work comes closest to ours. The research community uses BLEU score or ROUGE score which are essentially a measure of lexical similarity of sentences in the ground truth data with the sentences our algorithm generates. This is not quite proper since BLEU and ROUGE miss the semantic closeness of the sentences. We find that many of sentences our algorithm generated are matching with human judgment but did not match well with the ground truth sentences exactly.

Some of the images for which our algorithm fails to find description are such that their scene backgrounds were not found in our training images. Our algorithm relies on transferring acquired knowledge to new unseen images and hence it fails when a new scene background is found. However, this problem can easily be overcome if we expand our training dataset to include many possible backgrounds.

Another source of error was misidentification of objects. This too can be resolved if more training data samples are used. Lighting conditions, quality of images, and other noise factors were also responsible for somewhat poor object recognition.

6. CONCLUSION AND FUTURE WORK

Thus we are proposing an approach that generates sentences from image by understanding the entities in the image and their correlations among each other, which further uses Distributed Intelligence in order to generate proper sentences.

The series of activities that are carried out in entire process are: Object and scene recognition, tuple generation, sentence extraction, searching images and take sentence with highest similarity score as output. Based on further experience, the work can be expanded in future so as to increase the efficiency.

- Here, we record the top 5 sentences from the web depending on the likelihood of the words from the tuple. This reduces the chances of detecting unusual activities being captured from image. So in future, this discrepancy can be eliminated.
- The object detection technique fall short in detecting smaller objects in the background. This can also be improved so that minute objects get recognized.
- Instead of using Wiki dataset, one can go for much larger dataset that covers wide range of sentences with unusual activities also.

7. REFERENCES

[1] Kelvin Xu ,Jimmy Lei Ba ,Ryan Kiros ,Kyunghyun Cho ,Aaron Courville ,Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio ;Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015

[2] Andrej Karpathy, Li Fei-Fei Department of Computer Science, Stanford University; Deep Visual-Semantic Alignments for Generating Image Descriptions, 2015

[3] Ankush Gupta. Yashaswi Verma, C.V. Jawahar; Choosing Linguistics over Vision to Describe Images, 2012

[4] Girish Kulkarni,Visruth Premraj,Sagnik Dhar,Siming Li,Yejin Choi,Alexander C Berg,Tamara L Berg,Stony Brook University,Stony Brook University, NY 11794, USA Baby Talk: Understanding and Generating Simple Image Descriptions, 2013

[5] Joseph Redmon∗, Santosh Divvala∗†, Ross Girshick¶, Ali Farhadi∗† University of Washington∗, Allen Institute for AI†, Facebook AI Research¶; YouOnlyLookOnce: Unified, RealTime Object Detection, 2016

[6] Joseph Redmon, Darknet: Open Source Neural Networks in C, http://pjreddie.com/yolo/, 2013-2016

[7] M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013.