© 2016 IJSRSET | Volume 2 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Object Detection and Sentence Generation from Images

Anakha P. J.*, Devika Hari, Rinku Roy, Prof. Joby George

Department of Computer Science, M. G. University, Kerala, India

ABSTRACT

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. The ultimate goal is to generate descriptions of image regions. A model that generates natural language descriptions of images and their regions is thus developed. The approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Alignment model is based on a novel 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. A Multimodal Recurrent Neural Network architecture is described that uses the inferred alignments to learn to generate novel descriptions of image regions. The alignment model produces state of the art results in retrieval experiments on Flickr8K dataset. The generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations.

Keywords: Computer vision, Object detection, RNN

I. INTRODUCTION

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web. This task is significantly harder. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding. Most previous attempts have proposed to stitch together existing solutions of the above sub-problems, in order to go from an image to its description. The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories and great progress has been achieved in these endeavors.

Some pioneering approaches that address the challenge of generating image descriptions have been developed. However, these models often rely on hard-coded visual concepts and sentence templates, which impose limits on their variety. The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about the contents of the images and their representation in the domain of natural language. The second, practical challenge is that datasets of image captions are available in large quantities on the internet but these descriptions multiplex mentions of several entities whose locations in the images are unknown.

The core insight is that these large image sentence datasets can be leveraged by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular but unknown location in the image. The approach is to infer these alignments and use them to learn a generative model of descriptions. Concretely, the contributions are twofold:

- A deep neural network model is developed that infers the latent alignment between segments of sentences and the region of the image that they describe. The model associates the two modalities through a common, multimodal embedding space and a structured objective. The effectiveness of this approach are then validated

on image-sentence retrieval experiments in which the state-of-the-art is surpassed.

A multimodal Recurrent Neural Network architecture is introduced that takes an input image and generates its description in text. The model is then trained on the inferred correspondences and the performance is evaluated on a new dataset of region-level annotations. The rapid development in the field of digital image processing, motion detection and tracking are attractive research topics. In recent years, real-time video applications were inapplicable due to the expense computational time where an intelligent method to analyse the motion in a video stream line using the methods of background subtraction, frame differencing, and optical flow, methods are proposed. Organizations, commercial places and residential areas need to secure their facilities; this can be achieved by using security monitoring system with latest technology. An intelligent video sensor (Motion detector) was developed to support the monitoring security systems to detect unexpected movement without human intervention.

II. METHODS AND MATERIAL

A. Existing System

The problem of generating natural language descriptions from visual data has long been studied in computer vision, but mainly for video. This has led to complex systems composed of visual primitive recognizers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports.

The problem of still image description with natural text has gained interest more recently. Leveraging recent advances in recognition of objects, their attributes and locations, allows us to drive natural language generation systems, though these are limited in their expressivity. Farhadi et al. use detections to infer a triplet of scene elements which is converted to text using templates. Similarly, Li et al. start off with detections and piece together a final description using phrases containing detected objects and relationships. A more complex graph of detections beyond triplets is used by Kulkani et al., but with template-based text generation. More powerful language models based on language parsing have been used as well. The above approaches have been able to describe images "in the wild", but they are heavily hand designed and rigid when it comes to text generation. A large body of work has addressed the problem of ranking descriptions for a given image. Such approaches are based on the idea of co-embedding of images and text in the same vector space. For an image query, descriptions are retrieved which lie close to the image in the embedding space. Most closely, neural networks are used to co-embed images and sentences together or even image crops and subsentences but do not attempt to generate novel descriptions. In general, the above approaches cannot describe previously unseen compositions of objects, even though the individual objects might have been observed in the training data. Moreover, they avoid addressing the problem of evaluating how good a generated description is. In this work we combine deep convolutional nets for image classification with recurrent networks for sequence modeling, to create a single network that generates descriptions of images. The RNN is trained in the context of this single "end-to-end" network. The model is inspired by recent successes of sequence generation in machine translation, with the difference that instead of starting with a sentence, we provide an image processed by a convolutional net. The closest works are by Kiros et al. who use a neural net, but a feedforward one, to predict the next word given the image and previous words. A recent work by Mao et al. uses a recurrent NN for the same prediction task. This is very similar to the present proposal.

B. Proposed System

The ultimate goal of the model is to generate descriptions of image regions. During training, the input to the model is a set of images and their corresponding sentence descriptions. A model is first presented that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. These correspondences are then treated as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets.

Recurrent Neural Networks (RNNs) are popular models that have shown great promise in many NLP tasks. The idea behind RNNs is to make use of sequential

information. In a traditional neural network all inputs (and outputs) are independent of each other. But for many tasks this is not practical. If we want to predict the next word in a sentence we better know which words came before it. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps.

III. RESULTS AND DISCUSSION

Implementation

Dataset. We use the Flickr8K[3] dataset in our experiment which contains 8000 images. A JSON file which contains 5 sentences about each image in Flickr 8K is used as input to the RNN.

Data Preprocessing. We convert all sentences to lowercase, discard non-alphanumeric characters. We filter words to those that occur at least 5 times in the training set, which results in 2538 words for Flickr8k dataset.

A. Feature Extraction From Images

We extract a 4096-dimensional feature vector from each region proposal using the Caffe implementation of the CNN. Features are computed by forward propagating a mean subtracted 256x256 RGB image through five convolutional layers and two fully connected layers.

B. Training

In the training stage, the images are fed as input to RNN and the RNN is asked to predict the words of the sentence, conditioned on the current word and previous context as mediated by the hidden layers of the neural network. In this stage, the parameters of the networks are trained with backpropagation

C. Prediction

In the prediction stage, a withheld set of images is passed to RNN and the RNN generates the sentence one word at a time. The result is illustrated in Fig.1

Figure 1. Prediction Results

IV. CONCLUSION

A model that generates natural language descriptions of image regions is introduced based on weak labels in form of a dataset of images and sentences, and with very few hardcoded assumptions. The approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding. It has been showed that the model provides state of the art performance on image-sentence ranking experiments. Second, Multimodal Recurrent Neural Network architecture is described that generates

descriptions of visual data. The performance is evaluated on both full frame and region-level experiments and showed that in both cases the Multimodal RNN outperforms retrieval baselines.

V. REFERENCES

[1] Andrej Karpathy Li Fei-Fei "Show and Tell: A Neural Image Caption Generator", In CVPR, 2015

[2] Mike Schuster and Kuldip K. Paliwal, "Bidirectional Recurrent Neural Networks ", Member IEEE,June 2006

[3] M. Hodosh, P. Young, and J. Hockenmaier,"Framing image description as a ranking task: data, models and evaluation Metrics" ,Journal of Artificial Intelligence Research, 2013

[4] D. Elliott and F. Keller,"Image description using visual dependency representations" In EMNLP, pages 1292–1302, 2013.