ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

A Survey on Auto Image Captioning

D. D. Sapkal[1], Pratik Sethi[2], Rohan Ingle[3], Shantanu Kumar Vashishtha[4], Yash Bhan[5]

Professor, Department of Computer Engineering, PVG's College of Engineering and Technology, Parvati, Pune, India[1]

U.G Student, Department of Computer Engineering, PVG's College of Engineering and Technology, Parvati, Pune, India[2, 3, 4, 5]

ABSTRACT: Auto image captioning has recently gathered a lot of attention specifically in the natural language domain. There is a pressing need for context based natural language description of images, however, this may seem a bit farfetched but recent developments in fields like neural networks, computer vision and natural language processing has paved a way for accurately describing images i.e. representing their visually grounded meaning. We are leveraging state-of-the-art techniques like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and appropriate datasets of images and their human perceived description to achieve the same. We demonstrate that our alignment model produces results in retrieval experiments on datasets such as MS-COCO.

KEYWORDS: Computer Vision, Data Mining, Machine Learning, Artificial Intelligence, Neural Networks, Image Processing, Natural Language Processing.

I. INTRODUCTION

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task, even for the current state-of-art visual recognition models. The majority of previous work in visual recognition has focused on labelling images with a fixed set of visual categories (labels) and great progress has been achieved in these endeavours. However, while these mechanisms work great for a single label, blunt definition of an image, they are nowhere close to the i.e. vastly restrictive when compared to the enormous amount of descriptions a human being can compose.

Some approaches that address the above stated challenge of generating image descriptions have been developed. However, these models often rely on hard-coded visual concepts and pre-generated sentence templates, which imposes limits on their variety. Moreover, the focus of these works has been on reducing complex visual scenes into a single sentence, which proves to be an unnecessary restriction. In this work, we strive to take a step towards the goal of generating dense descriptions of images i.e. produce the natural language description of images rather than a restricted single label mechanical description.

The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their respective description in the domain of natural language. Additionally, the model should be free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data. The second, practical challenge is that datasets of image captions are available in large quantities on the internet, but these descriptions mix-up mentions of several entities whose locations in the images are unknown.

II. LITERATURE SURVEY

Convolutional Neural Networks (CNN) are biologically-inspired variants of Multi Layered Perceptrons. It is comprised of one or more convolution layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network. The architecture of a CNN is designed to take advantage of the 2D structure of an input image (or other 2D input such as a speech signal). This is achieved with local connections and tied weights followed by some form of pooling which results in translation invariant features. Another benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units.CNNhave been widely used and studied for image tasks, and are currently state-of-the art for object recognition and detection. A typical CNN looks like this:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

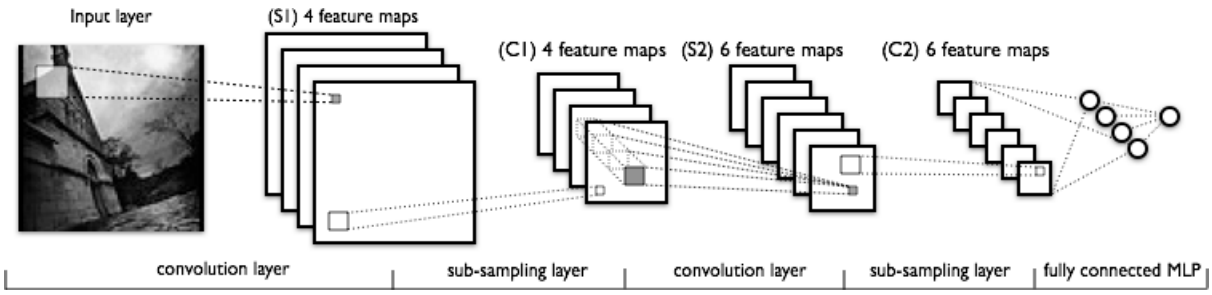

Fig. 1 A convolutional neural network and subsequent convolutional layers.

Recurrent Neural Networks (RNNs) are models that have shown great promise in many NLP tasks. The concept of RNNs is to make use of sequential information. In a traditional neural network we assume that all inputs (and outputs) are independent of each other. But for many tasks that's not effective. If you want to predict the next word in a sentence you have to 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. Alternatively RNNs can be thought of as networks that 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. Here is what a typical RNN looks like:

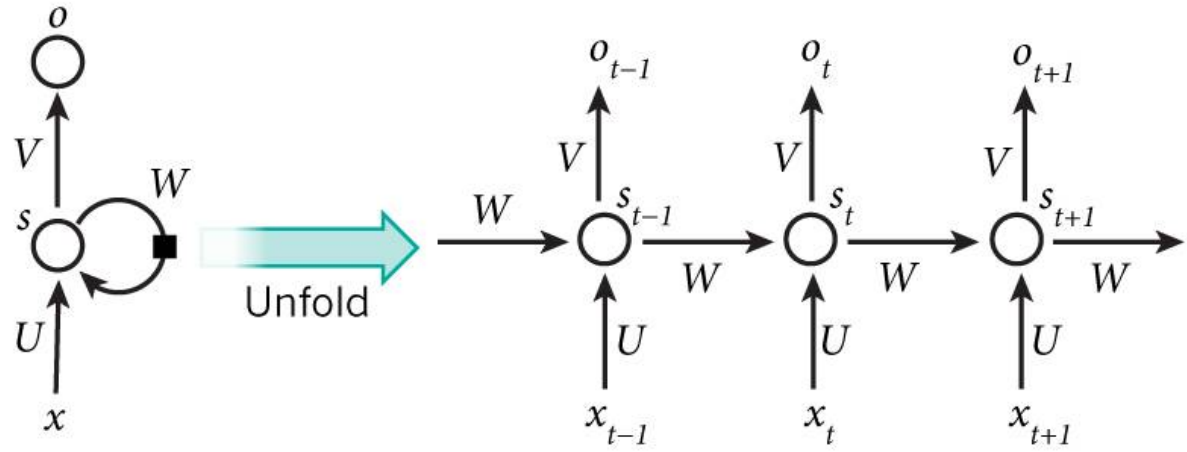

Fig. 2 A recurrent neural network and the unfolding in time of the computation involved in its forward computation. Source: wildml.com

The above diagram shows an RNN being unfolded into a full network. By unrolling we mean that we write out the network for the complete sequence. For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer neural network, one layer for each word

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

Our Model: Deep CNN (for computer vision) and RNN (for NLP):

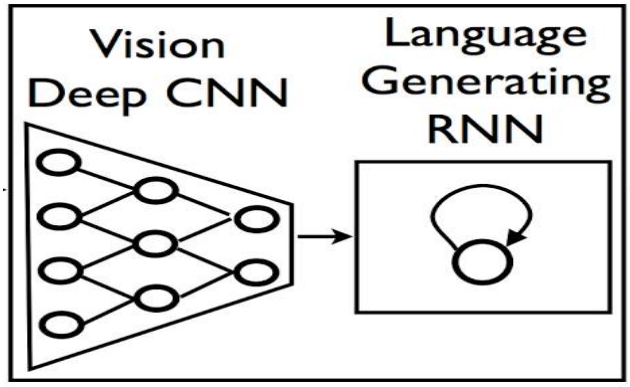

Fig. 3The model combines a vision CNN with a language-generating RNN so it can take in an image and generate a fitting natural-language caption.

The goal of our model is to generate captions or descriptions of images automatically. In the past there have been researches carried out by different groups, belonging to both, industry and academia that bear a resemblance to or are based on a topic similar to what we are doing. Many aspects of our model take references from these researches. The research papers that we utilized are cited in the bibliography respectively, like the work of D. Narayanswamy et al [1] which aims at generating labels that define the video frames, or that of D. Elliott and F. Keller[9], Image description using visual dependency representations, wherein the authors aim at identifying the different elements of an image. However the research stated above are generally concerned with and focused more on using image processing to detect and identify various objects in an image. They don't deal with identifying the context of these objects. These research papers have allowed us to greatly understand the concept of image processing and segmentation that plays a major role in our model. In addition to this, we try to obtain the contextual description of the image.

A lot of present research is also being carried out in the above mentioned area of understanding the context of images in a variety of fields especially in core aspects of industry and academia. The current status quo consists of extensive research by groups associated with computer vision and NLP, the ones that we surveyed are those at Stanford (A. Karpathy, Li Fei Fei) and UT, Austin, there research too aims at generating captions of images.

Apart from academia notable industrial developments include those by Microsoft (Cortana), Cortana is an Artificial Intelligence entity which is being developed by Microsoft, according to their recent conference Microsoft Build they are integrating the AI into their search service Bing, which would allow a user to interact more naturally with the UI. One of the first practical implementation of image processing and captioning was by Reddit (Reverse Image Search), which allowed a user to upload an image and their algorithm would analyze the image and display images with a similar context, this was followed by Google, although it is to be noted that both these projects are still in beta stages and under development. Another notable application of Computer Vision is by Facebook (Image Tagging).

Through our model we aim to produce semantically and visually grounded description of abstract images, the proposed description of which would be in natural language i.e. human perceived description. By leveraging the techniques like CNN, RNN, and data sets such as those of MS-COCO, we strive to attain the human level perception of given images. Our core insight is that we can leverage these large image-sentence datasets by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image.

Our approach is to infer these alignments and use them to learn a generative model of descriptions. We develop a deep neural network model that infers the alignment between segments of sentences and the region of the image that they describe. We introduce a Recurrent Neural Network architecture that takes an input image and generates its description in text. Our experiments show that the generated sentences produce sensible qualitative predictions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

We then train the model on the inferred correspondences and evaluate its performance on a new dataset of region-level annotations.

III. CONCLUSION

We have introduced a model that generates natural language descriptions of image regions based on obtained labels in form of a dataset of images and their respective sentence description, and with few assumptions.
Our approach features a novel sentence ranking model that identifies and generates various possible description of given image based on the labels acquired from processing it. This model aims at providing state of the art performance on image-sentence generation and ranking experiments.
Second, we described a Recurrent Neural Network (RNN) architecture that generates the natural language descriptions i.e. “human perceived" description of visual data. We evaluated its performance on various datasets of images and showed that in most cases the Multimodal RNN performs efficiently with respect to the status quo retrieval baselines.
In future work we hope to develop better sentence label representations, incorporate spatial reasoning, and move beyond bags of labels.

We have presented a neural network system that can automatically process an image and generate a reasonable description in natural language i.e. plain English. The model is based on a Convolutional Neural Network that encodes an image into a compact plain English representation, followed by a Recurrent Neural Network that generates a corresponding sentence.
The model is trained using a labelled dataset prior to deployment. BLEU, a metric used in machine translation to evaluate the quality of generated sentences, is used to evaluate and test the viability of our model. It is quite evident from these experiments that, as and when the size of the datasets (used for training) for image description increases, so will the performance of the model.

REFERENCES

1. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. Uai, 2012.
2.]K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. JMLR, vol. 3,1107–11352003.
3. Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning. Springer, vol. 3, 2006.
4. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server.uai, 2015.
5. X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. 2014.
6. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. CVPR, 2009.
7. M. Denkowski and A. Lavie. Meteor universal: Language speci c translation evaluation for any target language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014.
8. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. EECS, 2014.
9. D. Elliott and F. Keller. Image description using visual dependency representations. EMNLP, 2013.
10. J. L. Elman. Finding structure in time. Cognitive science, 1990