IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 4, April 2017

A Survey on Vision Based Approaches for Image Description

Akanksha P. Deshmukh[1], **Dr. A. S. Ghotkar**[2]

P. G. Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India [1]

Associate Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India [2]

Abstract: Automatic description from image is a challenging problem that contains interest from the domain like computer vision and natural language processing. In this survey, we represents deep learning concepts for describing a recent development in Computer Vision and Natural Languange Processing. This paper contains neural network appproaches for generating description of images, process of image decription, identified dataset and technolgies used for this framework and various evaluation metrics used for calculating scores.

Keywords: Computer Vision, Natural Language Processing, Machine Learning, Neural Network, Deep learning, Image Processing, Torch.

I. INTRODUCTION

Image description is sufficient for a human to point out and describe an large amount of details about visual description. Automatically generating captions of an image is a task very close to the heart of scene understanding. It requires identifying and detecting objects, people, scenes etc., reasoning about spatial relationships and properties of objects, combining several sources of information into a coherent sentence. Hence it is a complex task to define an image or a scene; which is an important problem in the field of computer vision. Even though it is a challenging one, a lot of research is going on which explores the capability of computer vision in the field of image processing and it helps to narrow the gap between the computer and the human beings on scene understanding. The purpose of this survey is to analyze various techniques used for an image caption generation using the neural network concepts. It will be dicussed in Section II.

Computer Vision task includes processing, acquiring, analysing and understanding a digital image which deal with extraction of high dimensionl data from real world in order to produce symbolical information.

Natural language generation constitutes one of the fundamental research problems in natural language processing (NLP) and is core to a wide range of NLP applications such as machine translation, summarizing, dialogue systems, and machine assisted revision[1]. Natural language generation still remains an open research problem. Most previous work in NLP on automatically generating captions or descriptions for images is based on retrieval and summarizing. In Section IV, a image description approaches will be dicussed.

Obtaining sentence level descriptions for images is becoming an important task and has many applications, such as early childhood education, image retrieval, and navigation for the blind.

In addition to the direct outputs of our system automatically generated natural language descriptions for images. There are also a number of possible related applications. These include improving accessibility of images for the visually impaired and creating text-based indexes of visual data for improving image retrieval algorithms. In addition, our work is in line with a more general research direction toward studying visually descriptive text and developing deeper into the connection between images and language that has the potential to suggest new directions for research in computer vision application. A few technologies are dicussed in Section VII.

Recent research in deep learning have inspired works which discuss a deep learning based approach inspired by recent advances in the applications of Convolutional deep neural networks and recurrent neural networks. To reduce the training time required for the Neural Image Captioning as well as integrate the decoder part into the network, while tweaking the convolutional part to adapt to the dataset[5]. The encoder part of NIC consists of a Convolutional Neural Network (CNN) called GoogLeNet, . Thus, in order to cut down on the training time, we tried to adapt the size of the network to the dataset by evaluating its performance on the dataset with multiple sizes. A few neural network approaches are dicussed in Section III.

A description must capture not only the objects contained in an image, but it also must express how these objects related 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.

DOI10.17148/IJARCCE.2017.6435

Dataset of image description is available in large quantities on the internet, but these descriptions mix up mentions of several entities whose locations in the images are unknown. Some of the standard datset used for training is discussed in section VI.

II. PROCESS OF GENERATING IMAGE DECRIPTION

The overall process for generating text based sentence description contains the steps explains as follws:

A. Dataset Collection

A dataset containing number of images needs to be collected. A dataset to train the classifier needs to be prepared. Training dataset will contain set of images along with annotations.

B. Preprocessing

The first step of pre-processing is representing json file in <Caption, ImageId> sequence format. It contains the caption for image along with id of image to which it indicates.

C. Training

Training is used for train image to reduce the complexity of accurate result for correct description of image. Train the model on the inferred correspondences and evaluate its performance on a new dataset of region-level annotations.

D. Testing

By representing images for testing wheather images gives accurate result in sentence generateed format for maintain the accuracy of dataset.

E. Sentence Generation

Neural network model 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.

III.NEURAL NETWORK APPROACHES

There are two main approaches used in sentence generation are Convolution Neural Network and Recurrent Neural Network based is also called as LSTM are :

A. Convolution Neural Network

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. 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. CNN have been widely used and studied for image tasks, and are currently state-of-the art for object recognition and detection[4].

B. Long Short Term Memory(LSTM)

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. 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[4].

The goal of 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. Image

DOI10.17148/IJARCCE.2017.6435

description using visual dependency representations, where in 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.
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 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)[3] and UT, Austin, there research too aims at generating captions of images.

IV. IMAGE DESCRIPTION APPROACHES

Andrej Karpathy et al (2014) [3] presented a model that generates natural language descriptions of images and their regions. This 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. Then used a Multimodal Recurrent Neural Network (MRNN) architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. Used the Flickr8K, Flickr30K and MSCOCO datasets for the experiment. The Multimodal RNN model is subject to multiple limitations. First, the model can only generate a description of one input array of pixels at a fixed resolution. A more sensible approach might be to use multiple the image to identify all entities, their mutual interactions and wider context before generating a desription.
Junhua Mao et al (2014)[7] proposed a multimodal Recurrent Neural Network (m-RNN) model for generating novel image captions. It directly models the probability distribution of generating a word given previous words and an image. Image captions are generated according to this distribution. The model consists of two sub-networks: a deep recurrent neural network for sentences and a deep convolutional network for images. These two sub-networks interact with each other in a multimodal layer to form the whole m-RNN model.
Kelvin Xu et al (2015) described an approach to caption generation that attempt to incorporate a form of attention with two variants: a hard attention mechanism and a soft attention mechanism. The hard stochastic attention mechanism is trainable by maximizing an approximate variation lower bound while the soft deterministic attention mechanism is trainable by standard back propagation methods. The main attention of the framework is the visualization of Where and What the attention is focused on. CNN act as an encoder and it extracts a set of features called convolution features of the input image. In order to obtain a correspondence between the feature vectors and portions of the 2-D image, features are extracted from a lower convolutional layer. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors. Then used a long short-term memory (LSTM) network that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. Two alternative mechanisms are used for learning as stochastic attention and deterministic attention. Stochastic hard attention represents location variables as where the model decides to focus attention when generating a particular word. Learning stochastic attention requires sampling the attention location while taking the direct expectation of the context vector can formulate deterministic soft attention model. Finally, quantitatively validated the usefulness of attention in caption generation with state of the art performance on three benchmark datasets: Flickr8k, Flickr30k and the MS COCO dataset.
Oriol Vinyals et al (2015)[4] In this paper, a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image is presented. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. This model is often quite accurate, when verified both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 scores on the Pascal dataset is 25, this approach yields 59, to be compared to human performance around 69. Also showed BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, it achieved a BLEU-4 of 27.7, which is the current state-of-the-art.

V. EVALUATION METRICS

Evaluating the output of a natural language generation (NLG) system is a fundamentally difficult task. The most common way to assess the quality of automatically generated texts is the subjective evaluation by human experts[3]. Some evaluation metrics used to calculate score are as follows:

A. BLEU

(Bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and

DOI10.17148/IJARCCE.2017.6435

that of a human: "the closer a machine translation is to a professional human translation, the better it is" this is the central idea behind BLEU. BLEU was one of the first metrics to achieve a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics. Scores are calculated for individual translated segments generally sentences by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness are not taken into account.

B. CIDEr

Automatically describing an image with a sentence is
a long-standing challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition,etc there is renewed interest in this area. However, evaluating the quality of descriptions has proven to be challenging.

C. ROUGE

Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

D. METEOR

(Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

VI. IDENTIFIED DATASET

Machine learning is a method of analyzing data to build a neural network model. We use Machine Learning Algorithms to iteratively learn from data[10]. Well known standard datasets are as follows:

A. MS COCO

MS COCO is a large image dataset designed for object detection, segmentation and caption generation. The Microsoft COCO dataset contains 82,783 training images and 40,504 validation images, each With 5 human generated descriptions. We used the training set and validation set to train our model in our experiments and uploaded our generated captions on the testing set (40,775 images) to the COCO server for evaluation.

B. Flickr30k

Flickr30k dataset has become a standard benchmark for sentence-based image description. Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding.

C. Pascal1k

The Pascal1K sentence dataset is a dataset which is commonly used as a benchmark for evaluating the quality of description generation systems. This medium-scale dataset, consists of 1,000 images that were selected from the pascal 2008 object recognition dataset and includes object from visual classes such as humans, animals and vehicles.

VII. TECHNOLOGIES

Deep learning neural network processing requires certain technologies to store, analyze and visualize results. Serial processing of every record takes longer time to produce meaningful conclusions. A distributed processing approach on commodity hardware should be considered.
Following technologies supports Machine Learning Algorithms and used for writing neural network applications-

A. Torch

Torch is a scientific computing framework with wide support for machine learning algorithms that puts GPUs first. It is easy to use and efficient, thanks to an easy and fast scripting language, LuaJIT. Torch is to have maximum flexibility

DOI10.17148/IJARCCE.2017.6435

and speed in building your scientific algorithms while making the process extremely simple. Torch comes with a large ecosystem of community-driven packages in machine learning, computer vision, signal processing, parallel processing, image, video, audio and networking among others, and builds on top of the Lua community.

B. Cuda

It is a parallel computing platform and applicationprogramming interface (API) model created by Nvidia. It allows software developers and software engineers to use a CUDA-enabled graphics processing unit (GPU) for general purpose processing – an approach termed (GPGPU) General-Purpose computiting on Graphics Processing Units. The CUDA platform is a software layer that gives direct access to the GPU's virtual instruction set and parallel computational elements, for the execution of compute kernels.

VIII. CONCLUSION

In this paper, a few approaches of neural network and image description are discussed. Neural network approaches for feature extraction are effective and provides accurate predictions by image description based approaches. To generate sentence from images, a deep learning neural network model is preferred for accurate result. 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.

REFERENCES

[1] Anurag Kishore and Sanjay Singh ”Natural langauage image descriptor”IEEE Recent Advances in Intelligent Computational Systems (RAICS)10-12 December 2015.

[2] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T.L.Berg, Baby talk: Understanding and generating image descriptions, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 35, NO. 12, DECEMBER 2013.

[3] Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CVPR, 2015.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, CoRR, vol. abs/1411.4555, 2014.[On-line].Available:http://arxiv.org/abs/1411.4555

[5] Frome, Andrea, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. ”Devise: A deep visual-semantic embedding model.” InAdvances in Neural Information Processing Systems, pp. 2121-2129. 2013.

[6] Karpathy, Andrej, Armand Joulin, and Li Fei-Fei. ”Deep fragment embeddings for bidirectional image sentence mapping.”arXiv preprint arXiv:1406.5679(2014).

[7] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille.”Deep captioning with multimodal recurrent neural networks (m-rnn)”. ICLR, 2015.

[8] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg,Tamara L. Berg and Yejin Choi”Collective Generation of Natural Image Descriptions”..

[9] Kelvin Xu, X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654, 2014.

[10] 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 CoRR,vol. Abs/1411.4389, 2014. [Online]. Available: http://arxiv.org/abs/1411.4389

[11] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation.In EMNLP, 2014

[12] 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

[13] M. Hodosh, P. Young, and J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell.Res.(JAIR), vol. 47, pp. 853899, 2013.

[14] Mitchell, Margaret, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos,and Hal Daum III. ”Midge: Generating image descriptions from computer vision detections.” InProceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 747-756. Association for Computational Linguistics, 2012.

[15] Kiela, Douwe, and Lon Bottou. ”Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics.” InProceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 36-45. 2014.

[16] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. ”Im2text: Describing images using 1 million captioned photographs” In Advances in Neural Information Processing Systems, pp. 1143-1151. 2011.

DOI10.17148/IJARCCE.2017.6435