INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 2057-5688

A survey on Deep Learning Image Caption Generation Models

[1]Anumalla Niharika, [2] Dr.T. Venugopal

[1]MTech Scholar, Dept. Of CSE, JNTUH University College of Engineering Jagtial

[2]Professor, Dept. Of CSE and Vice Principal, JNTUH University College of Engineering Jagtial

Abstract

Image captioning is the process of automatically creating captions for images. Since it is a new area of research it is getting increasing interest. To realize the aim of captioning images, the the semantic content of images has to be recorded and communicated in natural language. In connecting the research communities for computers and processing of natural languages captioning images is a complicated task. Many approaches have been suggested to address this issue. In this paper, we offer an overview of the latest developments in research on image captioning. Based on the method used to classify captioning methods into different categories. Examples of each method are reviewed along with their strengths and weaknesses are discussed. In this paper, we begin to examine methods employed in early research that are mostly templates and retrieval based. We then focus our attention on neural networks-based techniques, that provide state of the current results. Methods based on neural networks are further subdivided into subcategories depending on the framework they employ. Each of the subcategories of neural network-based techniques is discussed in depth. Then, the most up-to-date current methods are evaluated against benchmark datasets. Then, discussions about future research directions are discussed.

Keywords:

Image captioning, Sentence template, Deep neural networks, Multimodal embedding, Encoder-decoder framework, Attention mechanism

1. INTRODUCTION

Images are everywhere. They come from a variety of sources. As humans we can comprehend the meaning that these pictures contain without needing descriptions of images. However, computers do know as much about these images as humans do. Therefore, it is necessary for automatic generation of image explanations for images. This process is called image captioning. It employs the language model and computer vision methods. The algorithm used to caption images has to be able of focusing on the most prominent object in the image, and then create a natural language-based descriptions of the particular image. In order to retrieve the visual aspects the captioning algorithms use Convolutional Neural Networks (CNN). In contrast when it comes to sentence generation the group of Recurrent Neural Networks is employed.

Models for captioning images use a variety of types of data sets for training in testing, validating, and testing the models. The datasets are diverse from various ways like the size of collection in terms of number of the number of images, captions referenced provided per image caption format, size of the image. In order to evaluate the quality of generated captions, various evaluation metrics are used. These captions are evaluated against the reference captions. Each metric has its own benefits and drawbacks.

2. IMAGE CAPTIONING ANALYSIS

A captioning system for images should not only recognize the object that appears in the image, but also describe the relationship between them. The principal motive behind all the models that have been developed over time is to find an effective and precise method to carry out the job of automated captioning. Here are the various models used for captioning images:

2.1 Based on Encoder-Decoder Architecture A Work [11 (by Amritkar and Jabade) has presented a model comprised of CNN and RNN that is naturally regenerative. The model creates natural phrases to explain images. Their model uses CNN uses feature extraction, and RNN is employed to generate captions. The datasets they use in their model include Flickr8K, Flickr30K and MSCOCO. They've used an already trained CNN model to classify images which functions like an encoder for images. The encoder's output is then fed into the RNN network, which functions as a decoder and creates captions.

In their model, they have utilized the Visual Group Geometry (VGG) network, which is a deeper CNN. The model they have used is an element that is based on LSTM but has no peepholes. The model is composed of three models. The first is the Image model, the feature vector repeats 28 times because there is a maximum word count within the caption can be 28. The second model is the Language model, which is only one LSTM unit, whose output is an array. The third model combines two matrices, and then passes it to a different unit in the LSTM. The model was developed over 50 epochs and the loss that was observed was 3.74. Its BLEU (Bilingual Evaluation Understudy) score for the Flickr 8k dataset is 0.53356 For Flickr30k, it's 0.61433 as well for MSCOCO it's 0.67257 by analyzing 1000 images for each data set.

2.2 Based on Show and Tell Architecture

A Work [5] (by Shah, Bakrola and Pati) has implemented the Show and Tell model. In this model the image is processed by the Inception-V3 algorithm to detect all objects within the image. Once all objects are identified the result is passed to the input layer through a fully connected layer that changes the result into an embedding vector for words that is then processed through a sequence of LSTM cells. In the phase of training, they've processed captions by using tags to mark the beginning and the end in the line. In the test phase, the model is using Beam Search for finding appropriate words to create the caption. To train their model, they've employed MSCOCO. MSCOCO dataset. Tensorflow was utilized to design and create the model. They've used the BLEU score to assess their model. They have found that the average BLEU score for the model they have developed is 65.5.

2.3 Based on Attention Mechanism

A study called [7] (by Tian and Li) was designed to implement an image-to-sentence generation making use of Flickr8K, Flickr30K or MSCOCO datasets. Attention mechanisms are able discern what a word is referring to in an image. The soft attention system and the hard mechanism are two different kinds of the attention mechanism. They have been able to implement an attention system that is soft. In order to generate captions the model relied on the Long Short-Term Memory (LSTM). In their model, they employed CNN as an encoder and the LSTM decoder. CNN can be used draw features to images and LSTM is employed using attention function. The decoder is the only one trained. Their model could recognize the major elements of the image and translate them into sentences with ease. The sentences they generated are excellent in grammar, too. The only drawback of the model is that it didn't have a high BLEU score. In order to get better results, they needed to increase the size of their model, train it and fine tune it.

2.3 Other Architectures

In a work [8In a Work [8.4] (by Gan, Gan, He, Gao and Deng) The authors have created a brand new framework known as StyleNet to produce appealing captions in various styles for images and videos. The existing caption generation models utilize the traditional LSTMs that mainly record the long-term, sequential relationships between words in the sentence however they fail to take into account the style with other patterns of one particular language. In StyleNet the authors have developed an alternative version of the conventional LSTM known as"the factored" LSTM. In the model of the factored LSTM the matrix sets that comprise three different matrices are distributed among three different styles: factual, humorous, and romantic. A third matrix set is employed that is specific to the style and helps to identify the key factors in style within the text data. The tests conducted by the authors demonstrate that StyleNet can produce appealing and effective descriptions for photos. The authors have utilized FlickrStyle 10K as the FlickrStyle 10K dataset as a basis for training their model. To test the accuracy of their model, they've employed BLEU, METEOR, ROUGE and CIDEr as evaluation metrics.

3. DATASETS AND EVALUATION METRICS

In this paper, most popular captioning datasets for images like Flickr8K [1Flickr8K [1], Flickr30K [2], MSCOCO [33 IAPR TC-12 [4IAPR TC-12 [4], Visual Genome [5] and FlickrStyle10 are reviewed. Additionally, the main evaluation metrics utilized by the models that use captions for images like METEOR [6], BLEU [7], ROUGE CIDEr [9] as well as SPICE [10] are reviewed. A summary of evaluation metrics and data sources that are used by modern models is given.

BLEU. BLEU is a shorthand for Bilingual Evaluation Understudy.The high-quality text is assessed using the BLEU measurement. The principal idea of BLEU refers to the fact that, if a human translator is professional and close to machine translation, then the text's quality will be considered of good quality. The works that have utilized this evaluation metric to determine precision of models include METEOR. METEOR is the acronym as Metric to Evaluate the Quality of Translation with Explicit Ordering.. It is used to evaluate the output of machines. Synonyms are also considered to be matched in this method. The shortcomings of the BLEU metrics are overcome by this metric, and it can provide a sufficient connections with human judgements. Works [4,8] have incorporated this evaluation measure.

4. IMAGE CAPTIONING

In the modern, integrated world of language-vision, automated captioning of images has been recognized as an essential aspect of research. Image captioning is the process of the taking of an image, analysing the content of the image and producing a natural language description which defines the main features of the picture. The generated natural language is usually in nature in the shape of phrases; however, this isn't always the case. Captioning images is an difficult task from a CV and NLP perspective. Image comprehension requires the identification and understanding of objects' characteristics as well as the high-level aspects of the image such as indoor or outdoor images and the action that was performed. There are some challenges for example, implied objects in which the objects or individuals who aren't in the image could be mentioned. The captions created thus require a thorough understanding of the picture and could require knowledge databases. The captions that are generated should be simple to comprehend and be able to describe the image in detail.

The most significant issues faced by the problem of missing objects in the process of making the description of the image and the possibility that an object might be identified as belonging to a different category. in [15], a new method referred to as global-local attention (GLA) for the generation of description of an image is discussed. It is proposed that the GLA model makes use of an attention mechanism that combines local representation at the object level as well as global representation of the image. This allows for the processing of every object as well as context information simultaneously.

CNN to extract features of images in addition to RNN to predict the next word, making the resulting features inflexible to the word that is being produced at the present time. A new parallel system that is time-varying RNN (TVPRNN) that can solve this problem is described in [12]. The system employs two classic CNNs specifically VggNet and Inception V3 to extract the global image features, working with RNN to produce a time-varying feature in each time step which is used to represent the current state of the words. The representation of text and visual in a multimodal environment is combined. In addition, a visual attention is used to help guild this proposed system.

5. CONCLUSION

Image Caption Generation techniques based on features and data are discussed within the article. Direct techniques convert an image directly into sentences or by copying the caption from the related image or by computing the caption using captions derived from relevant images. These methods are heavily dependent on the image they extract their accuracy, precision and correctness. the captions that accompany them. Initial research into feature-based methods utilizes images parsing. Then, the images are identified by with visual primitive recognizers which are paired with graphs or logic systems which are then converted into natural language using rules-based systems. These systems are typically designed by hand and have only been demonstrated on a small number of areas. New techniques that combine the use with CNN and RNN for the generation of captions on images prove to provide superior results in caption generation.

6 REFERENCES:

[1] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, “I2T: Image parsing to text description,” in Proc. IEEE, vol. 98, no. 8, pp. 1485– 1508, Aug. 2010.

[2] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013.

[3] G. Kulkarni et al., "BabyTalk: Understanding and Generating Simple Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891-2903, Dec. 2013.

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg, “Baby Talk: Understanding and Generating Image Descriptions,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1601-1608, 2011.

[5] A. Farhadi “Every picture tells a story: Generating sentences from images,” in Proc. 11th Eur. Conf. Comput. Vis.: Part IV, 2010, pp. 15– 29.

[6] R. Kiros and R. Z. R. Salakhutdinov, “Multimodal neural language models,” in Proc. Neural Inf. Process. Syst. Deep Learn. Workshop, 2013.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, “Explain images with multimodal recurrent neural networks,” in arXiv:1410.1090, 2014.

[8] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, April 1 2017.

[9] Y. Yang, C. L. Teo, H. Daume III, and Y. Aloimonos, “Corpusguided sentence generation of natural images,” in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 444–454.

[10] D. Elliott and F. Keller, “Image description using visual dependency representations,” in Proc. Empirical Methods Natural Language Process., 2013, pp. 1292–1302.

[11] A. Gupta and P. Mannem, “From image annotation to image description,” in Neural Information Processing. Berlin, Germany: Springer, 2012

[12] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, April 1 2017.

[13] P. He´de, P.A. Moellic, J. Bourgeoys, M. Joint, and C. Thomas, “Automatic Generation of Natural Language Descriptions for Images,” Proc. Recherche d ’Information Assistee par Ordinateur, 2004.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” in Neural Computation. Cambridge, MA, USA: MIT Press, 1997.

[15] A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 619-632, Feb. 2017.

[16] K. Fu, J. Jin, R. Cui, F. Sha and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2321-2334, Dec. 1 2017.

[17] Z. Shi and Z. Zou, "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 6, pp. 3623-3634, June 017.

[18] A. Muscat and A. Belz, "Learning to Generate Descriptions of Visual Data Anchored in Spatial Relations," in IEEE Computational Intelligence Magazine, vol. 12, no. 3, pp. 29-42, Aug. 2017.

[19] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in Proc. 16th Conf. on Empirical Methods in Natural Language Processing, Edinburg, Scotland, July 27-31, 2011, pp. 444–454.

[20] D. Elliott and F. Keller, “Image description using visual dependency representations,” in Proc. 18th Conf. on Empirical Methods in Natural Language Processing, Seattle, Oct. 18-21, 2013, pp. 1292–1302.

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi , T. Berg, K. Stratos, and H. Daumé, III, “Midge: Generating image descriptions from computer vision detections,” in Proc. 13th Conf. of the European Chapter of the Association for Computational Linguistics, Avignon, France, Apr. 23-27, 2012, pp. 747–756.