Automatic Image Captioning Methods

Ruchitesh Malukani, Nihaal Subhash, Chhaya Zala

Abstract: ***A language known to humans is a natural language. In computer science it is the most challenging task to make the computers understand the natural languages and generating caption automatically from the given image. While a lot of work has been done, the total solution to this problem has been demonstrated daunting so far. Image captioning is a crucial job involving linguistic image understanding and the ability to generate interpretation of sentences with proper and accurate structure. It requires expertise in Image processing and natural language processing. The publishers suggest in this practice a system using the multilayer Convolutional Neural Network (CNN) to generate language describing the images and Long Short Term Memory (LSTM) to concisely frame relevant phrases using the driven keywords. We aim in this article to provide a brief overview of current methods and algorithms of image captioning using deep learning. We also address datasets and measurement criteria widely used for the same.***

Keywords: Image captioning, Deep learning, Computer Vision, Natural language processing, CNN, RNN, LSTM.

I. INTRODUCTION

Today, images are ubiquitously present in every aspect of our lives – in magazines, social media, news, books, advertisements, etc. Humans can intuitively understand images without captions, but machines cannot. Image captioning has a wide range of applications. It is used in image indexing, in social media websites, etc.

Image captioning is distinctively separate from the task of image classification and is more complex. While image captioning only focuses on object identification, captioning also involves identifying the properties of the identified objects (like location), overall context of the image, and the relationship amongst the different objects. Further it also involves generating properly structured sentences which are both syntactically and semantically correct.

Traditional machine learning approaches usually fail to meet this challenge and therefore techniques like CNNs, RNNs, Deep Learning, Adversarial Learning, Reinforcement Learning, etc. are used instead. These algorithms can handle the complexities associated with this task well. There have been many different approaches employed for this task and in this paper, we attempt to review some of the unique ones. We also give a brief overview of the common algorithms employed, datasets used and accuracies of the different approaches.

Revised Manuscript Received on May 20, 2020.
***** Correspondence Author
Ruchitesh Malukani*, Computer Engineering Department, G. H. Patel College of Engineering & Technology, Anand, India. E-mail: ruchiteshmalukani@gmail.com
Nihaal Subhash, Computer Engineering Department, G. H. Patel College of Engineering & Technology, Anand, India. E-mail: nihaal.subhash@gmail.com
Prof. Chhaya Zala, Computer Engineering Department, G. H. Patel College of Engineering & Technology, Anand, India. E-mail: chhaya20491@gmail.com

II. METHODS USED FOR IMAGE CAPTIONING

The key categories of the current image description methods are reviewed and defined in this section. Initial part of this section includes traditional image captioning methods template-based image captioning, retrieval-based image captioning and the later part gives an overview of novel approach of image captioning

A. Template-based methods

Template based image captioning is a bottom up approach of image captioning. Bottom-up approach involves the process of perceiving the individual parts and organizing them as a whole. In this method various objects, attributes and actions are first detected and blank spaces are filled out in the templates. This approach takes advantage of fact that descriptive language includes several important lexical patterns. It uses predefined templates, which will be filled based on the results of the on-scene element detection. For instance, a simple template represented in paper [1] by Kulkarni et al. can be read as “This is a photograph of <count> <object(s)>.” or “Here we see <count> <object(s)>.”. We use templates for encoding visual relationships such as “The <nth> <adjective> <object1> is <prep> <nth> <adjective> <object2>.”. Using these templates, we may create sentences in this way “This is a photograph of one sky.”, “Here we see one person and one train.” or “The first black person is by the blue sky.”.

Farhadi et al. [5] uses the triplet of picture objects for filling the template. Paper [1] by Kulkarni et al. represents Conditional Random Field (CRF) approach. This approach is used in paper [6] by Li et al. which describes generation of sentences associated with objects identified, their characteristics and their relationships.

B. Retrieval-based methods

Retrieval-based image captioning is an older approach for image captioning. In this approach captions are generated from set of existing annotations. Retrieval-based methods first find images similar in appearance along with their captions in training data collection.

To be able to effectively caption an unseen image, the dataset must satisfy two conditions:

1) It must be large and diverse enough that a reasonably similar match to the query image can be found.
2) The captions of the training dataset must be relevant to the pictures so that transferring captions between similar images is feasible.

Conventional data driven methods work only by using labels of semantically similar images. They are prone to issues like noisy estimations of the content of the images and differences between the actual content of the images and human written captions. Mason's approach uses a nonparametric density estimation technique for generating captions to address these issues. [7] It works by estimating a word frequency representation of the content of the input image. Caption generation is then worked upon as an extractive summarization problem.

The three stages of the process employed are as follows:

1) Defining a feature space to look for visually similar images Patterson et al., 2014 define the scene attributes in their paper [9]. The images are represented using a real valued vector with a dimensional size of 102. The similarity between images is measured using Euclidean distance.

2) Using density estimation to model the words which are being used to describe the relevant images.

3) Extractive Caption Generation

KL Divergence Method is used for this as it outperforms other methods like SumBasic sentence selection method.

This method can be improved by using compression methods which can remove visually irrelevant information.

Vincent takes a two-step strategy, which begins with the retrieval of globally similar images followed by image analysis to select the most relevant image. [8]

For selecting visually similar images, two image descriptors are used. The first is called gist – which gives a low dimensional representation of the scene. Similarly, the second is also a global image descriptor which is calculated by resizing the input image into a small resolution of 32x32. The sum of both these descriptors is taken to compute similarity.

After visually similar images are located, their phrases are retrieved and reranked. This process is divided into three stages:

Step 1: Dataset Processing: This step consists of the following stages.

Object detection: Object category detections (using part models) for 89 of the common object categories are run. Since running multiple object detectors gives very noisy results, only detectors of objects present in the caption of the database image are fired. Synonyms of the objects present in the caption may also be fired.

Image parsing: Image parsing is used in estimating the region of the database image which contains background elements. Six categories are used: building, tree, road, water and sky. The detectors are given input to a sliding window based SVM classifier. Scene classification: 26 common scenes are selected and their classifiers are fired. Features, method of classification and the training data are selected from the SUN dataset.

Caption Parsing: The Berkeley PCFG parser is used to obtain a hierarchical parse tree for each caption from which we gather constituent phrases, (like noun and verb phrases) referring to each of the above kinds of image content in the database.

Step 2: Retrieving Phrases

From the input image, several types of relevant phrases like Noun Phrases, Verb Phrases and prepositional phrases are selected. 5 features are used to measure visual similarity here: colour, texture, sift shape, hog shape and scene category. The first four are locally computed and the last is computed globally.

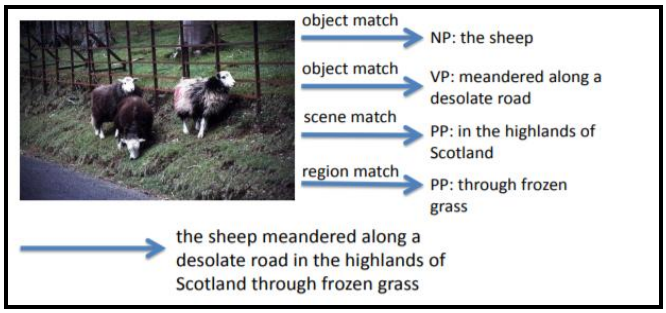

Figure 1: Retrieval of phrases [8]

Step 3 Reranking Phrases

In the last stage the relevant phrases are reranked so that the relevance of the top retrieved phrase can be increase. There is a lot of noise present in the retrieved phrases which has to be dealt with. Inaccurate image matches lead to inaccurate relevant phrases. However, if a lot of the relevant captions contain similar phrases like "black cat" the image probably contains a black cat. For reranking phrases, either PageRank based reranking using visual and/or text similarity is used or Phrase-level TFIDF based reranking.

Novel caption generation techniques are recent techniques for image annotation. In these methods captions are generated by combining image features and a language model instead of matching with the captions that already exist. Semantically more accurate captions can be generated by using this approach. These methods use deep learning techniques for caption generation. The next section talks about deep-learning based approach of image captioning.

III. DEEP LEARNING BASED IMAGE CAPTIONING TECHNIQUES

Recent work begins to focus on deep neural networks for automated image captioning due to great success in the field of deep learning. In this section we will discuss about following methods Encoder-Decoder, Attention Mechanism, Novel Object based mechanism and Semantic Concept-Based mechanism.

A. Encoder Decoder Mechanism

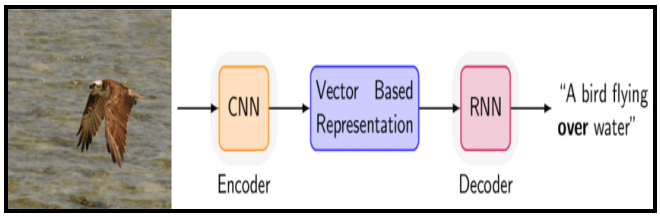

Figure 2: Encoder Decoder Mechanism [2]

Figure 2 shows a general structure of the encoder-decoder based image captioning methods. In this method a neural network acts as an encoder which encodes an image into intermediate representation of feature vectors.

This vector is given to the language generator which generates appropriate caption word by word. Here RNN is used as a decoder for caption generation.

Ryan Kiros et al. proposed an early work in this field in their paper. [10] This method uses CNN as an encoder for extracting features from the image for caption generation. It also introduces various multimodal neural language models. No additional templates, structures or constraints are used in approach which represents its biggest advantage.

Yuan et al., Yan et al. and Li et al. use encoder-decoder model for caption generation. [11, 13, 14]. They make use of LSTM as language translator. Aneja et al. use CNN for language translation.[12] Further details are mentioned in table 1 given in the next section.

B. Attention Mechanism

The concept of image captioning with attention mechanism was introduced by Kelvin Xu et al. [2] in 2015. Their work is inspired by the application of attention in image identification and sequence problems. In this context, attention as a technique refers to the ability to differ in weight regions of the image. In general, these techniques provide the allocation of available processing materials to the parts which provide more knowledge of the input signal. Instead of summarizing the image as a whole, the network can make the relevant and important parts of the image more weightful. Kelvin Xu et al. [2] represents a method in which the image is broken into a grid after the extracting the features from the image and generate a vector of features for each element.

As listed in their paper [3] by Luong et al. and presented in their paper [2] by Xu et al. there are two different ways of providing attention: i) Soft attention and ii) Hard attention

Soft attention is given by multiplying attention map over the image feature map and summing up it. Every region's feature vector receives weight in the soft attention variant at every step of decoding with RNN. This provides relative weights to each region of image. Multilayer perceptron followed by SoftMax function is used to calculate the weights. Hence this method is deterministic, and learning can be done by a standard backpropagation.

In Hard attention every step of generating the output word, only one region is sampled from the feature vectors. The focusing is done by random sampling hence it is not deterministic. Because of stochasticity of sampling, this avoids backpropagation of the network training.

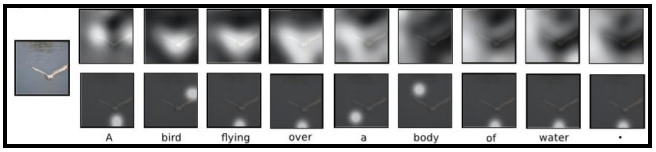

Figure 3: Soft Attention and Hard Attention [2]

For the given image as shown in the Figure 1 the top row indicates on which part of the image model is focusing on using soft attention method. The bottom row shows the same indication with hard attention.

C. Caption Generation for novel objects

In the paper [15] Lisa Anne Hendricks et al. illustrate a method that can identify and annotate new objects and their interactions with other objects. Their approach does this by utilizing broad object reorganization data sets, external text corpora and by maintaining flow of information between semantically related concepts. They evaluate their model on MSCOCO dataset. Their approach can also be extended to video caption generation. They represent concept of Deep Compositional Captioner (DCC) for caption generation for new objects.

Yao et al. also mentions novel object-based image captioning method. In this paper [16], they introduce a new design called LSTM with Copying Mechanism that combines copying into the Convolutional Neural Networks. In the next step this mechanism is integrated with a decoder RNN with copying mechanism in order to generate captions for novel objects.

D. Image captioning using semantic attention

This section will describe about the image captioning using semantic attention. Semantic attention is basically capable of dealing with a semantically important element in an image and of weighing the relative strength of attention given to multiple constructs. [18] Song et al. proposes [17] a hierarchical LSTM with adaptive adaption for both image captioning - hLSTMat. It uses spatial temporal attention for selecting specific regions of the image to predict the relevant words and adaptive attention to decide if it should focus on the visual information or the language context information. HLSTMat. can also simultaneously focus on both low-level visual information and high-level language context information for captioning.

Conventional methods based on decoder models applied the attention mechanism for every word including visual words and words like stopwords (‘the’,’ a’). But the other words can simply be predicted using language models and do not need to be a part of the visual attention model. When they are included, they reduce the performance of the overall system. Also, hierarchical LSTMs enable more complex representations of the image data and can capture information at varying scales

There are three major components in this system:

1. CNN Encoder

A pretrained ResNet-101 is used to extract global features, and a Fast R-CNN is used to generate the bounding boxes.

2. Attention Based Hierarchical LSTM Decoder

This consists of two LSTM blocks; one prepares the hidden states and visual attention which is used to generate an initial version of the captions and the second LSTM is used for proofreading.

3. Losses Training

Training is split into two different stages. First, the parameters of the model are pretrained by minimizing the value of MLE loss. Then reinforcement learning is used for fine-tuning. Here the model is treated as the agent and the words, global visual features and region visual features are treated as the external environment. Rewards are given after the complete caption generation. The reward function is prepared by combining a contrastive loss function with CIDEr and used to optimize the parameters of the model.

IV. EVALUATION METRICS

In these papers, for evaluating the quality of the captions, many evaluation metrics like BLUE, METEOR have been used. Here we represent a table describing comparison of performance of some approaches mentioned above.

Conventionally, these metrics are used for evaluating the quality of machine translations of text to human translations. An output score of '1' means the caption perfectly matches and '0' means it is completely unrelated to the actual human translation. For image captioning purposes, the image caption is treated as the 'translated text'.

BLUE (Bilingual Evaluation Understudy) works by comparing n-grams of the generated caption with n-grams of the reference caption and counting the number of matches.[4] It calculates the scores of the individual components than averages the score over the entire corpus.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is designed to address some of the drawbacks of BLEU. It achieves a higher correlation than BLEU. It computes the score based on the combination of unigram precision, unigram recall and a measure of how well ordered the matched words are in the generated caption compared to the actual caption. [19]

F1 Scores are used to measure a model's ability to integrate new vocabulary. It is calculated as the harmonic mean of recall and precision.

CIDEr is a new automatic evaluation metric to measure image descriptions with consensus.

Table- I

PAPER	B-1	B-2	B-3	B-4	METEOR	CIDEr	F1	DATASET
[2]	67.0	45.7	31.4	21.3	20.3	-	-	FLICKR8K
	66.9	43.9	29.6	19.9	18.4			FLICKR30K
	71.8	50.4	35.7	25.0	23.0			MSCOCO
[11]	71.6	52.2	39.0	28.9	23.8	-	-	MSCOCO
[12]	71.0	53.5	38.9	28.1	24.4	89.9	-	MSCOCO
[13]	71.6	51.8	37.1	26.5	24.3	-	-	MSCOCO
[14]	72.0	55.1	41.5	31.4	24.7	95.6	-	MSCOCO
[15]	64.40	-	-	-	21.00	-	39.78	MSCOCO
[16]	-	-	-	-	23	-	55.66	MSCOCO
[17]	73.8	55.1	40.3	29.4	23.0	66.6	-	MSCOCO
[18]	91.0	79.0	65.0	53.0	34.0	168.0	-	MSCOCO

V. RESULT AND DISCUSSION

Techniques based on templates can produce captions that are grammatically correct, but templates are pre-set and cannot produce captions of variable size.

As stated by Jiang et al. in paper [20] that in encoder decoder method, the vector containing objects and their semantic information is less likely to catch all the configurational information needed for the caption generation.

The drawback of attention mechanism is they can lose some knowledge that is useful for producing detailed subtitles. When weighted pooling is calculated only on attentive feature map, these methods slowly lose spatial details. They make use of information only received from the last layer of CNN i.e., the attention method uses characteristics from the lower convolutional layer rather than the fully connected layer in order to preserve details.

Retrieval based approaches produce syntactically correct captions, however since the image descriptions are constrained to sentences that already exist in the training dataset, they cannot adapt well to a mix of identified objects or novel scenes compared to other approaches.

While novel object image captioning produces state of the art results, a few common errors like not mentioning new objects, producing grammatically incorrect sentences, object hallucination, and irrelevant errors persist.

Image captioning is a challenging problem and different from conventional image recognition tasks in the sense that the generated caption has to not only identify the entities in the image but also recognize the relationships among them. Consequently, many models face the issue of high variance. One solution to this problem is to develop a model using an ensemble-based approach which produces the final caption by using the knowledge base of possible captions generated by different models. While the proposed approach is computationally intensive, with modern powerful and dedicated hardware for neural networks, it is still feasible.

VI. CONCLUSION

In this paper, we have reviewed various image captioning methods. We have briefly explained some techniques and discussed their advantages and disadvantages and their future potentials. We have also discussed the different metrics and datasets used in image captioning. We have also prepared a brief summary of the results. While image captioning methods have made considerable progress in the last few years, no model is yet to close to perfection and the subject is still open to different implementations. With the development of newer algorithms, newer deep learning architectures and with the increase in computational power, image captioning still has an exciting future as a research area.

REFERENCES

1. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C. and Berg, T.L., 2013. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), pp.2891-2903.
2. Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.
3. Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
4. Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
5. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In European conference on computer vision. Springer, 15–29.
6. Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 220–228.
7. Mason, Rebecca, and Eugene Charniak. "Nonparametric method for data-driven image captioning." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014.
8. V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, et al. Large scale retrieval and generation of image descriptions. International Journal of Computer Vision (IJCV), 2015.
9. Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The sun attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision.
10. Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14). 595–603.
11. Yuan, Aihong, Xuelong Li, and Xiaoqiang Lu. "3G structure for image caption generation." Neurocomputing 330 (2019): 17-28.
12. Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
13. Yan, Shiyang, et al. "Image captioning using adversarial networks and reinforcement learning." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.
14. Li, Jiayun, et al. "Image captioning with weakly-supervised attention penalty." arXiv preprint arXiv:1903.02507 (2019).
15. Anne Hendricks, Lisa, et al. "Deep compositional captioning: Describing novel object categories without paired training data." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
16. Yao, Ting, et al. "Incorporating copying mechanism in image captioning for learning novel objects." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
17. Gao, Lianli, et al. "Hierarchical LSTMs with adaptive attention for visual captioning." IEEE transactions on pattern analysis and machine intelligence (2019).
18. You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
19. Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005.
20. Jiang, Wenhao, et al. "Learning to guide decoding for image captioning." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

AUTHORS PROFILE

Ruchitesh Malukani, UG Student, G.H. Patel College of Engineering & Technology.

Nihaal Subhash, UG Student, G.H. Patel College of Engineering & Technology.

Chhaya Zala, Assistant Professor, G.H. Patel College of Engineering & Technology.