A Study on Automatic Image Captioning Techniques

Athirakrishna P. R, Remya P. S, Anusree I, Harishankar P
PG Student, Ass. Professor, PG Student, PG Student
Computer Science and Engineering
Vidya Academy of Science and Technology, Thrissur, India

Abstract : Automatically generating captions of an image is a task near to the core of scene understanding which is one of the essential objectives of computer vision. To achieve the objective of image captioning, semantic information of images should be caught and described in natural language. Image captioning is a very difficult undertaking, because it requires interfacing both research areas of NLP (natural language processing) and computer vision. There have been introduced several techniques for image captioning. The recent development in the image captioning is based on the advancement of artificial intelligence. This paper gives a review on some of the milestones in the field of automatic image captioning and provides comparison of methods based on some major evaluation metrics.

Index Terms **- Image captioning, CNN, RNN, Natural language processing, Deep learning**

I. INTRODUCTION

Automatic image captioning is a much studied point in both the Natural Language Processing (NLP) and Computer Vision (CV) ranges of inquires about. The process is to recognize the visual content of the input image, and to yield a important natural language caption. Individuals can by and large successfully delineate the circumstances they are in. Given a picture, it is ordinary for a human to depict the gigantic amount of experiences around this picture with a brisk look. But it could be a challenging for computers due to the inconstancy and ambiguity of conceivable picture depictions. Automatic image captioning increase developments in areas such as, Creating characteristic human robot interactions, early childhood instruction, data retrieval, outwardly weakened offer assistance.

Image captioning may be a much more involved task than object recognition [14] or classification, since of the extra challenge of learning representations of the interdependency between the objects/concepts within the image and the creation of a compact sentential portrayal.

The description of a picture is the yield of an extremely complex process that includes:

- a detailed understanding of an image
- ability to communicate that information via natural language

Automatically creating captions of a picture could be a task exceptionally near to the heart of scene understanding which is one of the essential objectives of computer vision. Not as it were must caption generation models be capable sufficient to illuminate the computer vision challenges of deciding which objects are in an picture, but they must moreover be able of capturing and communicating their connections in a natural language. For this reason, caption generation has long been viewed as a troublesome issue. It could be a exceptionally vital challenge for machine learning algorithms, because it sums to mimicking the exceptional human capacity to compress gigantic amounts of striking visual information into descriptive language.

Given an image, the objective of image captioning is to produce a sentence that is linguistically plausible and semantically honest to the contents of this picture. So there are two fundamental inquiries associated with image captioning, i.e. visual understanding and language processing. To guarantee produced sentences are linguistically and semantically right, strategies of computer vision and NLP should be embraced to manage issues emerging from the corresponding technique and integrated appropriately. To this end, different methodologies have been proposed.

In spite of the challenging nature of image captioning, there has been a recent surge of research interest in the field of image captioning. This paper is a study on some of the important methods for automatic image captioning.

II. LITERATURE SURVEY

Initially, automatic image captioning is just attempted to yield basic descriptions for pictures is done by utilizing a predefined image, caption dataset or by using certain manually created rules. The major limitation of such methods is the obtained captions are not flexible enough. After that with the advancement in the fields of computer vision and artificial intelligence major developments are happened in the research area of image captioning. This paper goes through some of the major milestones in the area of automatic image captioning.

2.1 Description Based on Concept Hierarchy of Actions

A method proposed by Kojima et al.[1], for generating textual description of office environments which explains human behaviour in real video images by extracting semantic features of human motions and associating them with concept hierarchy of actions.

Initially the process estimates the human actions and identifying objects, and then conceptual descriptions of actions for each body part are generated. Finally integrating them into one description of the whole body, and translating it into natural language text.

2.2 Non-parametric Method for Image Captioning

Non-parametric Method for Data-driven Image Captioning by Mason and Charniak [2] was proposed to overcome the challenges such as noisy estimations of visual content and poor alignment between images and human-written captions.

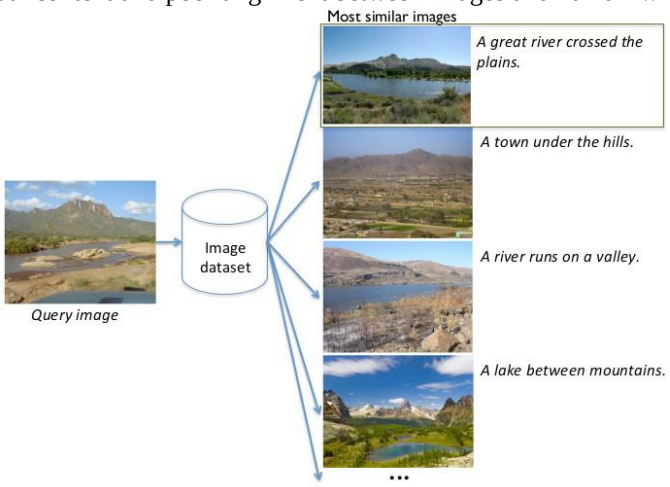

Fig.1. Non-parametric Method for Image Captioning

Rather than selecting an output caption agreeing to a single noisy estimate of visual similarity, this framework uses a word frequency model to discover a smoothed assess of visual content over numerous captions. It then creates a depiction of the query image by extracting the caption which best speaks to the mutually shared content. For a query picture I_q, create a important depiction by selecting a single caption from C, a huge dataset of pictures with human-written captions. At first define the feature space for visual similarity, and after that define a density estimation problem with the point of modelling the words which are utilized to portray visually comparable images to I_q.

2.3 Corpus-Guided Sentence Generation for Natural Images

Corpus-guided sentence generation is presented in [3] by Yang et al. , that depicts images by anticipating the foremost likely nouns, verbs, scenes and prepositions that make up the core sentence structure. This hypothesis is based on the assumption that natural images precisely reflect common ordinary scenarios which are captured in language.

The approach is summarized in Fig. 2. At first recognize objects and scenes utilizing trained detection algorithms [4], [5]. To keep the system computationally tractable, they constrain the components of the quadruplet (Nouns-Verbs- Scenes-Prepositions) to come from a limited set of nouns N , actions V, scenes S and prepositions P classes that are commonly used. Then utilize a language model [6] trained over the huge corpus to predict verbs, scenes and prepositions that will be utilized to compose the sentence. With probabilities of all components computed, the most excellent quadruplet is obtained by utilizing Hidden Markov Model (HMM) inference. At last, the image description is created by filling the sentence structure given by the quadruplet.

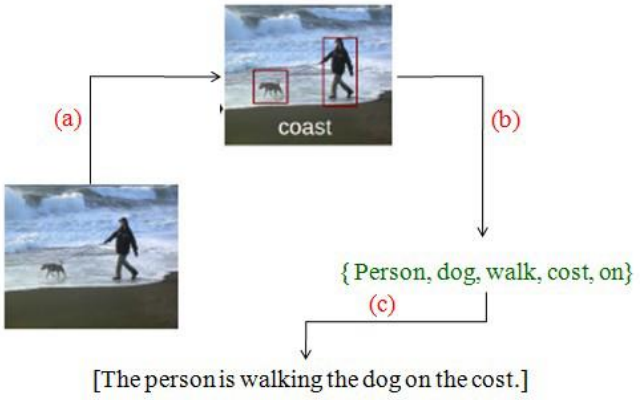

Fig. 2. Corpus-Guided Sentence Generation of Natural Images

2.4 Captioning with Multimodal Recurrent Neural Networks

Multimodal neural network is one of the methods that depend on immaculate machine learning to create image captions. Mao et al. [7] proposed a multimodal Recurrent Neural Networks (m-RNN) model to address both the task of creating novel sentences descriptions for images, and the task of image and sentence retrieval. The entire m-RNN demonstrates contains a language model portion, a vision portion and a multimodal portion. The language model portion learns a dense feature embedding for each word within the dictionary and stores the semantic temporal context in recurrent layers. The vision portion contains a deep Convolutional Neural Network (CNN) which produces the image representation. The multi-modal portion interfaces the language model and the deep CNN together by a one-layer representation.

Under their structure, a deep CNN [9] is utilized to extract visual features from images, and a RNN [10] with a multimodal part is utilized to model word probabilities conditioned on image features and context words. For the RNN language model, every unit comprises of an input word layer w, a recurrent layer r and an output layer y. At the t_{th} unit of the RNN language model, the calculations performed by these three layers are shown below:

$$x(t)=[w(t)\ r(t1)] \quad (2.4.1)$$

$$r(t)=f(U \cdot x(t)) \quad (2.4.2)$$

$$y(t)=g(V \cdot r(t)) \quad (2.4.3)$$

Where f (·) and g (·) are element-wise non-linear functions, and U and V are matrices of weights to be learned. The multimodal part calculates its layer activation vector m(t) by using the equation below:

$$m(t)=g_m\,(V_w \cdot w(t) + V_r \cdot r(t) + V_I \cdot I) \quad (2.4.4)$$

Where g_m is a non-linear function. I is the image feature. V_w, V_r and V_I are matrices of weights to be learned. The multimodal part fuses image features and distributed word representations by mapping and adding them. To train the model, a perplexity based cost function is minimized based on back propagation.

2.5 Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

Motivated by later progresses in multimodal learning and machine translation, Kiros et al.[11] presented an encoder decoder framework into image captioning, which successfully binds together joint image-text embedding models and multimodal neural language models. They argue the image caption generation can be considered as a translation problem.

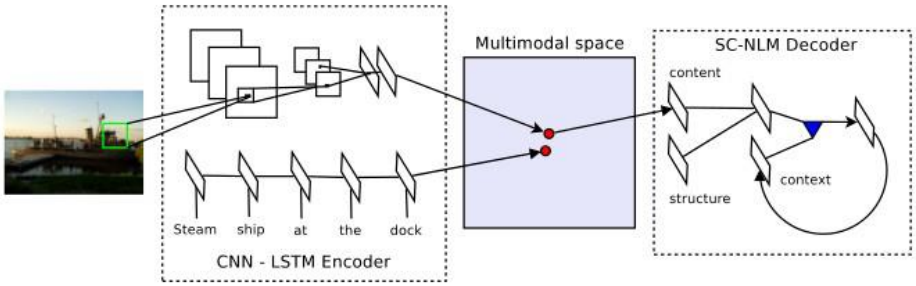

Fig. 3. Encoder Decoder Framework

Fig.3. appears architecture of strategy, where encoder is a deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) for learning a joint image sentence embedding; decoder is a new neural language model that combines structure and content vectors for creating words one at a time in sequence.

They use Long Short-Term Memory (LSTM) RNN to encode literary information [12] and a deep CNN to encode visual information. After that, by utilizing a pair wise ranking loss, encoded visual information is anticipated into an embedding space spanned by LSTM hidden states that encode textual information. Within the embedding space, a structure-content neural language model is utilized to decipher visual features conditioned on context word feature vectors, permitting sentence generation word by word.

2.6 Neural Image Caption Generation with Visual Attention

Motivated by later work in machine translation and object detection, Xu et al.[13] presented an attention based model that automatically learns to describe the content of images. They portray how a model can be trained in a deterministic way using standard back propagation methods and stochastically by maximizing a variational lower bound, also show how the model is able to automatically learn to fix its look on salient objects while creating the corresponding words within the output sequence.

An image is given to a deep Convolutional Neural Network and features are extracted from a lower convolutional layer of the network, then encode the picture as a set of feature vectors. Within the translating stage, a Long Short-Term Memory network with a context vector is utilized to dynamically represent image parts that are important for caption generation.

They proposed a stochastic hard attention and a deterministic soft attention for image captioning. The stochastic hard attention system chooses a visual feature from one of the N locations as the context vector to produce a word, while the deterministic soft attention instrument consolidates visual features from all N locations to obtain the context vector to create a word.

III. COMPARISON

According to the above section we can classify the image captioning techniques into methods using neural networks and methods without neural networks. Obviously the methods using neural networks are more efficient than other methods. Since the methods without neural networks are based on some predefined rules and strategies like templates the captions produced by such techniques are less natural, and also such methods can not able to handle a image that is not in its database.

The most instinctive approach to decide the quality of produced sentence depicts the content of a picture is by direct human judgments. But human assessment requires a lot of unwanted human efforts; moreover it varies from person to person. This paper includes strategy comparison based on evaluation metrics. The main evaluation metrics for image captioning methods are; BLEU, METEOR and CIDEr.

BLEU (Bilingual Evaluation Understudy): An algorithm [17] for assessing the quality of content. Quality is considered to be the correspondence between output of a machine and human caption: "the closer a machine translation is to a proficient human translation, the better it is" – typically the central idea behind BLEU. BLEU was one of the primary metrics to claim a tall relationship with human judgements of quality. It makes utilize of variable lengths of phrases of candidate sentence to compare with reference sentences composed by human to decide their closeness. There's distinctive variance of BLEU like BLEU-1, BLEU-2, BLEU-3, and BLEU-4 according to the length of phrases utilized to compare.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): Metric [16] based on the harmonic mean of unigram precision and recall with recall weighted higher than precision. It too has a few highlights that are not found in other measurements, such as stemming and synonym coordinating, together with the standard exact word matching. The metric was outlined to settle a few of the issues found within the more well known BLEU metric, additionally deliver great relationship with human judgment at the sentence or segment level. This varies from the BLEU metric in that BLEU looks for relationship at the corpus level.

CIDEr (Consensus-based Image Description Evaluation): A recent metric proposed for assessing the quality of image portrayals. It measures the consensus between candidate image description and the reference sentences given by human annotators. This metric is aiming to assess delivered sentences in terms of grammaticality, saliency, importance and accuracy.

The Table I shows method comparison on Microsoft COCO Caption. In this table, B-n, MT, CD stands for BLEU-n, METEOR, and CIDEr, respectively.

Category	B-1	B-2	B-3	B-4	MT	CD
Multimodal learning	0.625	0.450	0.321	0.230	0.195	0.660
Encoder decoder framework	0.670	0.491	0.358	0.264	0.227	0.813
Attention Guided	0.724	0.555	0.418	0.313	0.248	0.955

Table1. Comparison on MICROSOFT COCO caption dataset

Table 2 shows examples of image captioning results obtained based on different approaches to give readers a straight forward impression for different kinds of image caption methods.

Category	
Captioning with Multimodal Recurrent Neural Networks	A street sign on the side of the road.
Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models	A stop sign on the side of the road.
Neural Image Caption Generation with Visual Attention	A stop sign on road with trees
Human caption	A floodway sign sitting on the side of road in a field.

Table2. Captions obtained using different approaches

IV. CONCLUSION

Image captioning infers automatically creating a caption for a image. To achieve the objective of image captioning, semantic information of pictures should be caught and communicated in natural languages. Associating both research ranges of computer vision and natural language processing, image captioning may be a exceptionally difficult task. With the headway in deep neural network, utilizing more effective network structures as language models and/or visual models will undoubtedly move forward the performance of image description generation. Since picture comprise of objects distributed in space and image captions are arrangements of words, examination on presence and order of visual concepts in image captions are vital for picture captioning. Be that as it may, to describe images at a human level and to be applicable in real life situations, image description should be well grounded by the elements of the images. Therefore, image captioning grounded by image regions will be one of the longer term research directions.

REFERENCES

[1] A. Kojima , T. Tamura , K. Fukunaga , Natural language description of human activities from video images based on concept hierarchy of actions, Int. Comput. Vis. 50 (2002) 171–184.

[2] R. Mason , E. Charniak , Nonparametric method for data driven image captioning, in: Proceedings of the Fifty Second Annual Meeting of the Association for Computational Linguistics, 2014 .

[3] Y. Yang, C.L. Teo, H. Daume , Y. Aloimono , Corpus-guided sentence generation of natural images, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 4 4 4–454.C.

[4] P.F. Felzenszwalb, R.B. Girshick, D. McAllester , D. Ramanan , Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[5] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001)145–175.

[6] T. Dunning, Accurate methods for the statistics of surprise and coincidence, Comput. Linguist. 19 (1) (1993) 61–74.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Explain images with multimodal recurrent neural networks, arXiv: 1410.1090v1 (2014).

[8] J. Mao , W. Xu , Y. Yang , J. Wang , Z. Huang , A. Yuille , Deep captioning with multimodal recurrent neural networks, in: Proceedings of the International Conference on Learning Representation, 2015 .

[9] A. Krizhevsky , I. Sutskever , G.E. Hinton , Imagenet classification with deep convolutional neural networks, in: Proceedings of the Twenty Fifth International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[10] J.L. Elman, Finding structure in time, Cognit. Sci. 14 (2) (1990) 179–211 .

[11] R. Kiros, R. Salakhutdinov, R. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv: 1411.2539 (2018).

[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image captiongeneration with visual attention, arXiv: 1502.03044v3 (2016).

[14] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, YouTube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: Proceedings of the International Conference on Computer Vision, pp. 2712–2719.

[15] K. He, X. Zhang, S. Ren , J. Sun , Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[16] C.-Y. Lin , F.J. Och , Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in: Proceedings of the Meeting on Association for Computational Linguistics, 2004 .

[17] A. Lavie, A. Agarwal, METEOR : an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.