Comparison of Image Captioning Methods

[1]Jeel Sukhadiya, [2]Harsh Pandya, [3]Vedant Singh
[1]Department of Information Technology,
[1]Dwarkadas J.Sanghvi College of Engineering, Mumbai, India

Abstract—**Humans can give insight descriptions of the images or the scenes presented to them. Computer vision aims at incorporating this ability of humans to provide distinctive and captious description of the images and different scenes. Thus, image captioning is a task of generating a subjective description of all the objects, their relationship with the environment around them present in the images, to effectively describe the scene. Various approaches have been described in the paper to solve image captioning tasks. A detailed analysis of these approaches has been done along with a descriptive comparison to thoroughly gain an insight into the working and the methodologies used in various approaches. Possible alternative approaches are also introduced to achieve better performance in image captioning tasks**

Index Terms—**Computer vision, image captioning, feature detection.**

I. INTRODUCTION

Image captioning is a process of generating image descriptions for a detailed understanding of the various elements of the image. The elements include the objects/person present in the image , the background or the setting of the environment in which the image is based , and the relationship of the objects and all the entities of the image with among themselves and the environmental setup in which they exist. Language or any form of communication can be used to describe the significant amount of information present around us in the world. Similarly, the language can be used to provide usable and important information from the scenes depicted in the images. This leads to a better understanding of the scene by generating captions out of images and using the captions to thoroughly understand the information from the images.

Several factors are required to get an in-depth understanding of an image such as the spatial and semantic information about the various entities present in the image, the backdrop in which the image is based and the relationships between all the elements of the image. For generation of captions from images, the two major tasks that needs to be performed on the images, are as follows:

1. Gaining information about the world.
2. Generating sentences to describe the Vision world.

So different methods of Computer Vision and Natural Language Processing (NLP) are incorporated for extracting information from the images and representing them in the form of meaningful sentences.

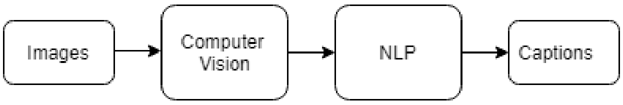

Fig. 1. General Model of an Image Captioning Process.

Generation of captions or description from images has been a wide area of research. The work in image caption generation can be traced back to the year 2010 where Ali Farhadi[1] provided an introspective about how the captions can be generated and how the images can be described with the help of sentences. Many other methods followed, but the most recent work by P.Anderson and team[2] achieved state-of-the-art performance on image captioning tasks. A deep analysis, comparison and drawbacks of various works have been discussed in the paper and a call for using alternative methods has been made to improve the performance on image captioning problems.

II. LITERATURE SURVEY

A large amount of work has been done on image caption generation task. The first significant work in solving image captioning tasks was done by Ali Farhadi[1] where three spaces are defined namely the image space, meaning space and the sentence space where mapping is done from the respective image and sentence space to the meaning space.

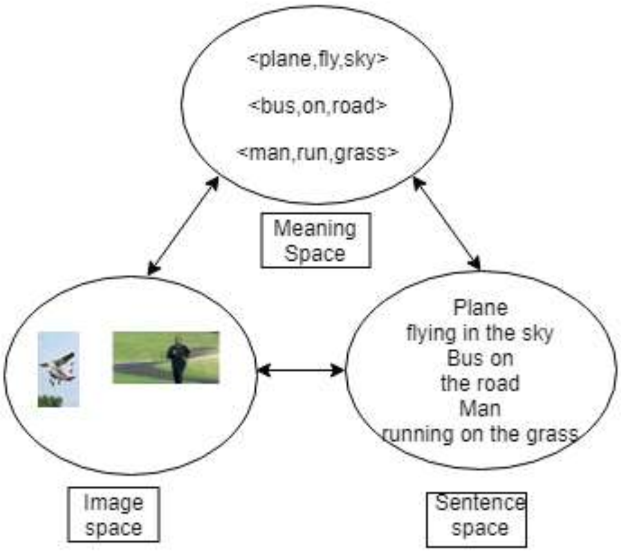

Fig 2.The three spaces defined by the work proposed by Ali Farhadi[1]

With the help of mapping, similarity between the images and the sentence is evaluated, the meanings are stored as triplets of (image, action, object) and a score is evaluated by predicting the image and sentence triplets. If an image and sentence have high level of similarity in terms of the predicted triplets then they will be highly compatible and have a high score. Thus, appropriate sentences can be generated. This model has many drawbacks such as requirement of the middle meaning space and the results obtained from it are not at all highly accurate. Various other works were introduced but more recent work use the methodology of neural networks for solving the task. With the advent of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), a good performance was achieved and found applications in various fields of study. O.Vinyals and team, in the work [3], introduced a novel approach of using (CNN) and (RNN) for image captioning tasks. Convolutional neural networks were used to extract features from the images. So, CNN acts as a encoder, first for classification of tasks and the last layer's output is provided as the input to (RNN). (RNN) acts as a decoder that generates sentences. LSTM networks (Long Short Term Memory) was the type of RNN used.

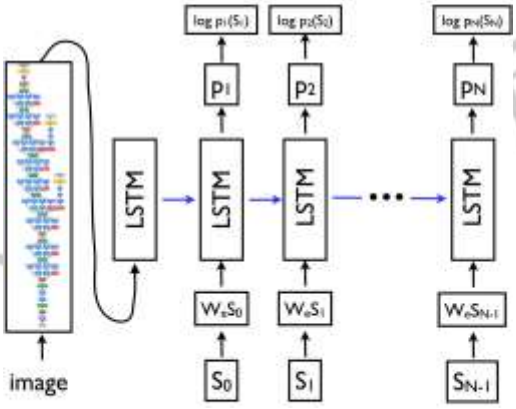

Fig 3.The figure represents the pipeline of the CNN and LSTM network as illustrated in [3, Fig. 3].

A Similar approach was used by Donahue et al. (2014)[4] which used LSTM's on videos instead of images. Works of [3][4][5] proposed a model on which detectors were trained to extract various features out of the images and a translation model was trained on a set of captions to generate the appropriate descriptions for the image at hand. On the contrary other novel work carried out by [6] used a novel approach for generating captions of images with neural networks and visual attention. In this approach, attention is given to the most important part of the image and producing a sentence around it. In real world scenarios there is noise or clutter present in the images so unlike the traditional methods, not all the features are fed into the (RNN) but only the important and salient features are fed into the (RNN).

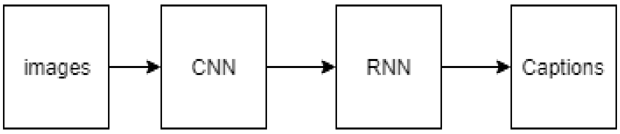

Fig 4.Traditional model used to generate captions.

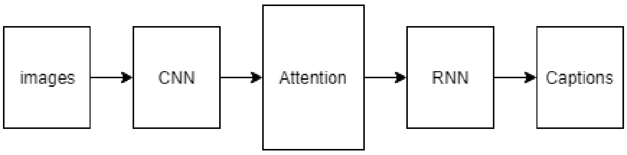

Fig 5.Generating Descriptions with the help of visual attention.

A better performance framework was achieved by [7] by taking visual attention as the basis of the proposal. Most of the attention models for image captioning and visual question answering [8][9][10] attended to images at every step.

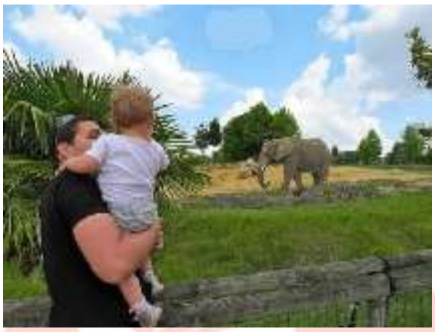

Fig 6.Caption generated by using attention models is as follows- A man carrying a child while standing near the fence of a elephant zoo.

Considering the caption generated in Fig.6 the attention models proposed by [8][9][10] attend to images even while predicting words like 'A' ,'while', 'near', 'the', 'of', which have no visual signal in the image. A novel approach was proposed by [7], where an adaptive attention model via a visual sentinel was used for image captioning tasks. The proposed model by [7] was proposed keeping in mind as to when to rely more on visual signals and when to rely on language models for generating captions for images. While relying on the visual signals the proposed model learned "where" to rely for the more important and salient features in the image. The visual sentinel is used by the proposed model to determine as to when to rely on input signals. The visual sentinel is the representation of what the network already knows. The information is stored in the decoder of the framework for both short term and long term use. Other works such as by [11] for image captioning were based on the combination of reinforcement learning and sequence based training as a basis for generating descriptions. In this, a self-critical sequence training is done using reinforcement learning as a baseline, given the rewards from the inference algorithm of reinforcement learning, the metrics can be updated and the rewards can be normalized. As the proposed framework is self-critical it does not depend on a third-party framework or algorithm to evaluate and normalize the rewards of the inference algorithm. All the proposed frameworks were based only on objects present in the already existing image captioning datasets. Most of the work in image captioning was limited to only the image captioning datasets, thereby limiting the object categories for which the captions could be generated, The novel approach of the work proposed by Subhashini Venugopalan [12] took advantages of not only the image captioning datasets but also the external sources of datasets such as object recognition datasets. Thus, a large variety and diversity of the object categories were used in the approach. A Novel Object Captioner (NOC) network was proposed which could generate captions from images with diverse objects. This approach outperformed prior works with a large number of object categories and a considerably better performance.

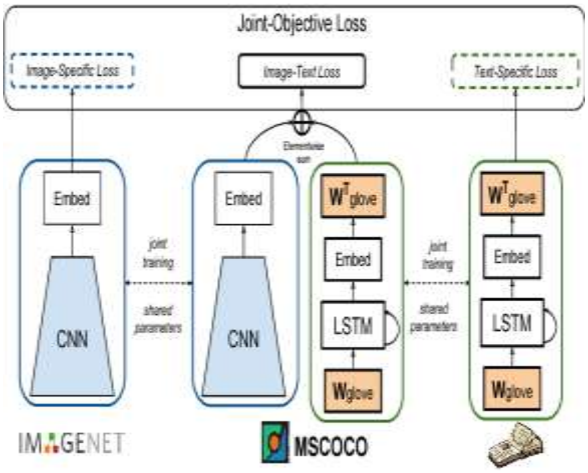

Fig 7.The Novel Object Caption network as illustrated in [3, Fig 2]

A substantial amount of work has been done in solving image captioning and visual question answering tasks using visual attention mechanism by [11,7,13,14] which were based on top down visual attention mechanisms. However a novel framework was proposed by [2] which emphasized not only on the top down and bottom up mechanisms to detect and extract all the salient features of the image. A state-of-the-art performance was achieved by using this approach. This approach uses a combination of both top-down and bottom-up attention for image captioning. For the bottom-up approach a faster-RCNN network was used to detect the objects belonging to various classes and by representing them by drawing boxes around them. The approach uses a soft top-down attention mechanism. The captioning model consists of two different LSTM layers, the first one is the top-down LSTM layer and the second LSTM layer is the language model. The model achieved a state-of-the-art performance by achieving a BLEU-4/Spice/Cider scores of 36.9,21.5,117.9 respectively as illustrated in [2]. Most of the work done in image captioning is based on combining CNN with other type of models. A capsule network can be used as mentioned by Geoffrey Hinton in his proposed work [15]. The reasons behind using capsule networks in place of a CNN network are explained in the section IV.

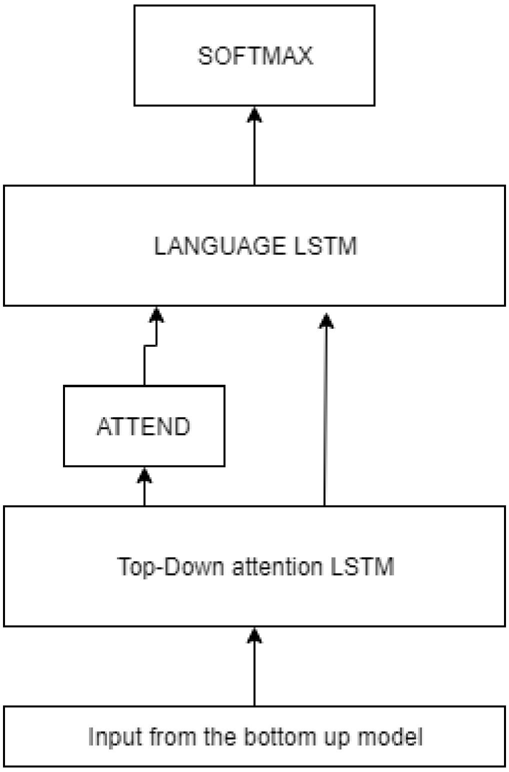

Fig 8.The captioning model as illustrated **in [2, Fig 3]**

III. COMPARISON OF RESULTS

IV. A comparative analysis of various methods is done. The comparisons of the results obtained are made with the help of two novel approaches [16] & [17]. The quality of captions are evaluated by the standard evaluation metric BLEU [16] and METEOR [17] respectively. The BLEU and METEOR scores of various approaches are mentioned below:

V.

Models	MS-COCO				
	B1	B2	B3	B4	Meteor
Image cap with semantic attention	0.709	0.537	0.402	0.304	0.243
Hard Attention	0.718	0.504	0.357	0.250	0.230
Adaptive attention via visual sentinel	0.742	0.580	0.439	0.332	0.266
SCST:Att2in	-	-	-	0.313	0.26
SCST:att2all	-	-	-	0.30	0.259
ResNet	0.745	-	-	0.334	0.261
Bottom up and top down attention	**0.772**	-	-	**0.362**	**0.27**

Table 1.Comparision of different Image Captioning Methods.

VI. FUTURE WORKS

Most of the work done in generating captions from images includes the use of Convolutional Neural Network as an important component of the framework. Convolutional Neural Network has been extensively used as a Encoder in works such as [2][3][4][7][11] and many more. As suggested by the work proposed by Geoffrey Hinton [15], [18], he points out the drawbacks of Convolutional Neural Networks described as follows:

1. CNN does not take into account the orientational and the spatial relationship of the features. It can be illustrated with the help of an example:

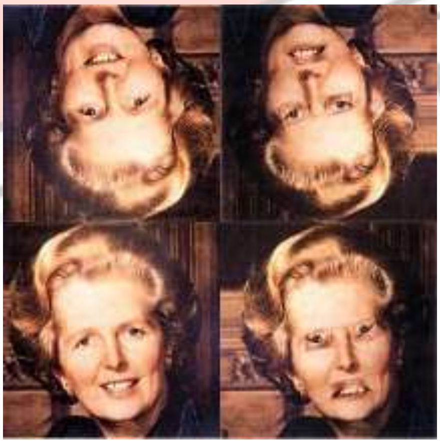

Fig 9.A image showcasing drawbacks of CNN

   In the above figure 9, the convolutional neural network will identify all the four images mentioned as a face, as compared to identifying and extracting both, the orientation and spatial relationship of the faces in the four images.

2. The approach used by convolutional neural networks to solve the above difficulty is to use max pooling which leads to a certain amount of loss of information from the image.

We generally start with a representation of a feature using instantiation parameters and then those instantiation parameters are rendered to produce images whereas CapsNet (Capsule Network) makes use of inverse Graphics. Inverse Graphics is a complete opposite method wherein the capsule predicts the instantiation parameters through the inverse rendering of the images. Capsule Networks take more time as compared to Convolutional neural networks to be trained. Apart from the required time, Capsule Networks can be used to achieve state-of-the-art results for image captioning tasks. Apart from Capsule Network, many other methodologies can be used to achieve the state-of-the-art performance on image captioning tasks.

VII. Conclusions

Thus, we have studied and compared the various approaches in image captioning and a complete analysis of these approaches have been made. Alternative methods for image captioning are described which have the potential to replace commonly followed approach of using Convolutional Neural Networks. Thus, an in depth comparison of various methods is made.

VIII. References

[1] Farhadi A. et al. (2010) Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis K., Maragos P., Paragios N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in CVPR, 2018.

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652-663, 1 April 2017.

[4] onahue *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, 1 April 2017.

[5] Andrej Karpathy, Li Fei-Fei- Deep Visual-Semantic Alignments for Generating Image Descriptions,In CVPR 2015

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio ; Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015

[7] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning, 2017 IEEE Conference on Computer Vision and Pattern Recognition.

[8] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In NIPS, 2016

[9] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In ICML, 2016.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016.

[11] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-Critical Sequence Training for Image Captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1179-1195.

[12] Venugopalan, S., Hendricks, L.A., Rohrbach, M., Mooney, R.J., Darrell, T., & Saenko, K. (2017). Captioning Images with Diverse Objects. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1170-1178.

[13] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen. Review networks for caption generation. In NIPS, 2016.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[15] Hinton G.E., Krizhevsky A., Wang S.D. (2011) Transforming Auto-Encoders. In: Honkela T., Duch W., Girolami M., Kaski S. (eds) Artificial Neural Networks and Machine Learning – ICANN 2011. ICANN 2011. Lecture Notes in Computer Science, vol 6791. Springer, Berlin, Heidelberg

[16] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, 2002.

[17] M. Denkowski and A. Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014

[18] Sabour, S., Frosst, N., & Hinton, G.E. (2017). Dynamic Routing Between Capsules. *NIPS*.