International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

A Survey of Image Captioning Models

Sanober Sultana Shaikh[1], Vinaya Bomnale[2], Kajal Gandhi[3], Nikunj Otwani[4]

[1]*Professor, Department of IT Engineering, Thadomal Shahani Engineering College, Maharashtra, India*
[2]*Student, Department of IT Engineering, Thadomal Shahani Engineering College, Maharashtra, India*
[3]*Student, Department of IT Engineering, Thadomal Shahani Engineering College, Maharashtra, India*
[4]*Student, Department of IT Engineering, Thadomal Shahani Engineering College, Maharashtra, India*

Abstract - *Image caption generation has been a challenging problem for a long time.Numerous attempts have been made at the difficult task of image captioning, which includes the complexities of both computer vision and natural language processing. Deep Learning models have the capability to perform the intricate task of image captioning. In this survey paper, we aim to give a complete review of the various image captioning techniques that have been implemented till date. We discuss the structure of the various models, their performance, advantages and limitations. The different datasets and evaluation metrics that are frequently used in image captioning models have also been discussed.*

Key Words: **Image Captioning, CNN, LSTM, RNN, Computer Vision, Deep Learning.**

1. INTRODUCTION

The human mind is capable of deducing the semantic meaning of an image without having to refer to any descriptions of the image. It is able to process large amounts of information without much effort. But for automatic image caption generation, the machine has to first identify all the key objects present in an image and generate a syntactically and semantically correct sentence for it. With the rise in popularity of artificial intelligence, the task of image captioning has become achievable.

Image Captioning has various applications in the domain of artificial intelligence such as giving recommendations in editing applications, usage in virtual assistants, robotics, social media and several other natural language processing applications. It can also be quite beneficial for the visually impaired.

The main motivation of this paper is to give the readers a comprehensive view of the various image captioning models. We first group the models into four different categories:(1)Encoder-decoder image captioning models (2)Show and tell image captioning models, (3) Attention mechanisms and (4)Others.Section 2 describes these categories.A brief overview of the various datasets and evaluation metrics used in these models has been given in section 3. Finally, we conclude the survey paper in section 4.

2. IMAGE CAPTIONING ANALYSIS

An image captioning model must not only recognise the object in the image but also express the relation between them. The main motivation behind all of these models,that have been implemented over the years, was to find an efficient and accurate way to perform the task of automatic image captioning. Following are the different architectures implemented for image captioning:

2.1 Based on Encoder-Decoder Architecture

A Work [1] (by Amritkar and Jabade) has described a model consisting of CNN and RNN which is regenerative in nature. The model generates natural sentences to describe an image. In their model, CNN is used for feature extraction and RNN is used for caption generation. The datasets used in their model are Flickr8K, Flickr30K and MSCOCO. They have used a pre-trained CNN model for image classification which acts as an image encoder. The output of the encoder is passed as input to the RNN network which acts as a decoder and generates captions. In their model they have used Visual Group Geometry (VGG) network which is a deep CNN. Their model has used a block which depends on LSTM with no peephole architecture. Their model has 3 sub-models; first, in the Image model the feature vector is repeated 28 times since the maximum number of words in the caption is 28. Second is the Language model which has a single LSTM unit whose output is a matrix. Third model merges the two matrices and passes it to another LSTM unit. The model was trained for 50 epochs and the loss observed was 3.74. The BLEU (Bilingual Evaluation Understudy) score for the dataset Flickr 8k is 0.53356, for Flickr30k it is 0.61433 and for MSCOCO it is 0.67257 by testing 1000 images of each dataset.

In work [2] (by Singh and Sharma), they have also implemented an encoder-decoder model for image captioning. They have used CNN as the encoder to extract the image features and a LSTM as the decoder for generating the caption. They have implemented two different models for extracting the image features. In Model-1, they have used a pre-trained VGGNet that will help in feature extraction. In Model-2, they have created a 4-layer CNN to extract the features. The decoder part of the model contains an embedding layer and a LSTM. The model used the Flickr8K dataset for training. They have

used 6000 images for training the model, 1000 images for testing and 1000 images for validation. They have used BLEU metric evaluation for evaluating their models. The BLUE-1 score for Model-1 and Model-2 are 0.54 and 0.51 respectively. The BLUE-2 score for Model-1 and Model-2 are 0.28 and 0.25 respectively. The BLUE-3 score for Model-1 and Model-2 are 0.19 and 0.17 respectively. The BLUE-4 score for Model-1 and Model-2 are 0.082 and 0.07 respectively. The captions generated by both the models don't differ by much. Since the dataset contained a difference in number of images for different scenarios and objects, the captions generated were not that appropriate in some cases. They further believe that using a bigger and unbiased dataset will help in increasing the accuracy of the captions generated by their model for random image sets. They also believe that increasing the number of layers or implementing a pre-training step will help in increasing the accuracy of their model.

Work [3] (by Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach et al), proposes recurrent convolutional architectural models for image captioning, activity recognition and video description. They put forward a new type of neural network called LCRN which stands for Long-term Recurrent Convolutional networks. It combines convolutional layers with temporal recursion. Using this concept, they trained video recognition models by linking a convolutional model directly to the LSTM networks. They also propose a model that implements the encoder-decoder architecture, where a visual convolutional neural network is used for encoding the deep state vector and for decoding the vector into a caption, LSTM is used. Their model is successfully able to perform better than the baselines from recent work but is beaten by the VGGNet model with respect to retrieval tasks. The MSCOCO dataset was used to train their model and the evaluation metrics used are R-precision and success@k.

Traditionally the encoder and decoder are trained in two different steps but a Work [4] (by Zhu, Li et al) has introduced a concept of Joint Learning.The innovative aspect of this model is that both the CNN (encoder) and the LSTM (decoder) model learns at the same time hence it is called Joint learning. For feature extraction, they have used a pre-trained VGGNet. They had first trained their model on the MSCOCO dataset. They had later fine tuned their model by training it on different datasets. The evaluation metric used by them to calculate the accuracy of their model is METEOR. The accuracy of the captions when trained on MSCOCO dataset, Flickr30K dataset and Flickr8K dataset was 0.133, 0.117 and 0.123 respectively. They also experimented by not applying joint learning. The accuracy when joint learning was applied was 0.165 whereas when joint learning was not applied the accuracy was 0.133. For their experiments they chose a model which was trained on the MSCOCO dataset as the pretrained model since it had the highest accuracy. They have then tuned their model further on different datasets. They first fine tuned their model using a small dataset of 2000 images out of which 1500 were for training and 500 for validation. They later trained their model on a big dataset which was obtained by combining Flickr30K and Flickr8K. The accuracy of the model increased to 0.174 due to the use of a bigger dataset. They have then hand-picked 1000 more images along with Flickr8K and Flickr30K dataset for the final finetuning of the model and the accuracy of the model increased to 0.204. From their experiments, we can see that the accuracy of the model increased when they applied joint learning. We can see that by choosing a large dataset the accuracy also improved.

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 first passed through the Inception-V3 model for detecting all the objects in the image. Once all the objects have been recognised, the result is sent as input through a single fully connected layer which transforms the output into a word embedding vector which is then passed through a series of LSTM cells. In the training phase they have pre-processed the captions by adding tags to signify the start and end of the string. In the testing phase, the model has used Beam Search for finding appropriate words for generating the caption. For training their model they have used the MSCOCO dataset. Tensorflow was used to create and train the model. They have used the BLEU score to evaluate their model. The average BLEU score of their model is 65.5.

In the work [6] (by Fu, Liu, Xie), the authors have implemented an image captioning mechanism called Show, Attend and Tell. It is an image captioning generator accompanied with visual attention. The five major components in their implementation are: Data Prepossessing, Convolutional Neural Network (encoder), attention mechanism, Recurrent Neural Network (RNN) as a decoder, Beam Search to find the most optimal caption. They have used Flickr8K dataset to train their model. The input images need to be preprocessed into a proper format for the CNN network and the captions for the RNN network. VGG-16 and ResNet are used as image encoders. They extract various features from the images and encode them into a vector space which is to be fed to the RNN. Following CNN, they have built a soft trainable attention mechanism (Show, Attend and Tell) which tells the network which part of the image to focus on for generating the next word in the caption. The decoder then uses a RNN called LSTM which is able to generate words sequentially. The final step consists of the Beam Search, which helps in generating a sentence with the highest likelihood of occurrence with respect to the image. They have used the BLEU score to evaluate their model.

2.3 Based on Attention Mechanism

A work [7] (by Tian and Li) aimed to implement image to sentence generation by using Flickr8K, Flickr30K or MSCOCO datasets. Attention mechanisms are able to identify what a word refers to in the image. Soft attention mechanism and hard attention mechanism are the two types of attention mechanisms. They have implemented the soft attention mechanism. To generate the captions, the model made use of Long Short-Term Memory (LSTM). In their model they had used CNN as the encoder and LSTM as the decoder. CNN is used to map features from the image and LSTM is used with attention functions. Only the decoder was trained. Their model could recognise the main parts of the image and show them in sentences quite well. The generated sentences do well in grammar too. The limitation of their model was that it did not have a good BLEU score. To get better results they needed to enlarge their model and train and tune it further.

2.3 Other Architectures

In a Work [8] (by Gan, Gan, He, Gao and Deng), the authors have proposed a novel framework called StyleNet to generate attractive captions of different styles, for images and videos. The current caption generating models use traditional LSTMs, which mainly capture the long-term sequential dependencies between the words in a sentence, but fail to factor the style from other linguistic patterns of a particular language. In StyleNet, the authors have used a variant of the traditional LSTM called the factored LSTM. In the factored LSTM model, the matrix sets which contain three different matrices are shared among different styles, factual, humorous and romantic. Another matrix set is used which is style specific and is used to distill underlying style factors in the text data. The experimentations carried out by the authors show that StyleNet is capable of generating effective and attractive captions for images. The authors have used the FlickrStyle 10K dataset to train their model. To evaluate the accuracy of the model they have used BLEU, METEOR, ROUGE and CIDEr evaluation metrics.

Work [9] (by Demirel, Cinbis, Ikizler-Cinbis) aims at generating captions for images which contain objects the model hasn't seen before in the training dataset. The proposed model makes use of a zero shot object detection model (ZSD) and a template-based sentence generator. ZSD is used to leverage all the similarities between different classes obtained from the distributed word representations. YOLO has been used as the backend architecture for the ZSD model. For image captioning, a template-based image captioning model is used that employs a recurrent neural network to generate templates containing sentences with blanks. The names of the objects detected by the ZSD are filled into these blanks. When compared to the NBT-baseline, the proposed model gave much better results since it had the capability of captioning unseen objects as well. The model was trained on the MSCOCO dataset. They evaluated their model using mAP i.e Mean Average Precision.

In a Work [10] (by Pan, Yang et al), they have implemented 4 new methods for automatic image captioning. The methods are used to create a translation table which then helps to caption the given image. Before actually creating the translation tables they have performed some preprocessing tasks on the images. They have manually annotated the image regions with labels which are called blob-tokens. The G-means algorithm is used to generate the blob-tokens. They have also used a different concept for creating their data matrix called Weighting by Uniqueness.The dataset used for performing their experiments are 10 Corel image datasets, where each dataset contains 5200 training images and 1750 test images. The evaluation metric used by them to evaluate the accuracy of their model is the percentage of correctly captioned words. Their proposed methods have around 12% absolute accuracy and around 45% relative accuracy as compared to the unweighted state of art methods over baseline. They have also used another evaluation metric called recall and precision values for each word. The average precision values of the unweighted method, correlation method, cosine similarity method, SVD correlation method and SVD cosine method are 0.0411, 0.1131, 0.1445, 0.1197 and 0.2079 respectively. Their methods have shown lesser bias to the training images and are more generalized and the weighting by uniqueness concept improves the accuracy of the model.

Work [11] (by Fariha) suggests use of multi-task learning for captioning images automatically. The author has developed a system, in a multi-task framework, that solves two tasks simultaneously. Initially, all the essential features of an image will be extracted. These are then forwarded as input to both the caption generation task and activity detection which is an auxiliary task. The auxiliary task is a classification problem of the simple multi-class and multi-label type, for detecting any action occurring in the image. Even though the BLEU score for this model is quite subpar and the author hasn't implemented several parameters like batch-normalisation, drop-out and attention, the author proposes that if the auxiliary task is related to the original task then multi-task learning can improve performance. The BLEU evaluation metric has been used to evaluate the model and the author uses MSCOCO dataset to train the model.

3. DATASETS AND EVALUATION METRICS

3.1 Datasets

1) MS COCO Dataset. The Microsoft COCO is a sizable dataset used for object detection and image captioning. COCO (Common Objects in Context) implies that the images in the dataset are everyday images or common images captured from day to day life. COCO has 300,000

images out of which 200,000 are annotated. There are 80 object categories with 2 million images and every image has 5 captions associated with it. Image captioning methods [1,3,4,5,6,7,9,11] have used this dataset for their implementation.

2) Flickr 30K Dataset. The Flickr30k dataset contains 244,000 coreference chains and 276,000 bounding boxes which have been manually annotated for each of the 31,783 images and 158,915 captions in English. Each image has 5 captions. This dataset is one of the most widely used datasets after MS COCO. This dataset is used in different image captioning models [1,4,6,7].

3)Flickr 8K Dataset. The Flickr8K dataset consists of 8000 images wherein each image has 5 different captions. All the images have been chosen from six different Flickr groups and each image has been selected manually. A number of image captioning models [1,2,4,6,7] have been implemented using this dataset.

4) ImageNet Dataset. ImageNet dataset has 15 million annotated images out of which 1 million images have an annotated bounding box for object detection. It also has more than 22,000 categories. This dataset is quite popular among researchers due to its high quality images.

5) Pascal VOC Dataset. The PASCAL Visual Object Classes (VOC) dataset has 20 object categories which includes vehicles, household, animals, and others. The usage of this dataset is widely seen in object detection, segmentation, and classification tasks. In this dataset, each image is assigned with pixel level segmentation annotations, bounding box annotations, and object class annotations. The whole dataset is divided into 3 subsets consisting of a training set of 1464 images , validation set of 1449 images and a private testing set.

6) Visual Genome Dataset. The Visual Genome dataset contains 101,174 images from the MSCOCO dataset with 1.7 million question answer pairs. On an average there are 17 questions per image. The questions asked are of six types: What, Where, When, Who, Why and How. This dataset also presents 108,000 images with annotated objects, attributes and relationships.

3.2 Evaluation Metrics

1) *BLEU.* BLEU stands for Bilingual Evaluation Understudy.The quality of text is evaluated using the BLEU metric. The main ideology behind BLEU is that if the professional human translation is close to the machine translation then the quality of the text is considered to be good. Works that have used this evaluation metric to calculate the accuracy of their models are [1,2,5,6,7,8,11]

2)*METEOR.* METEOR stands for Metric for Evaluation of Translation with Explicit Ordering.. It is used for evaluating machine translated output. Even synonyms are considered for matching in this approach. The limitations of the BLEU metric are overcome by this metric and is capable of making adequate correlation with human judgements. Works [4,8] have implemented this evaluation metric.

3)*ROUGE.* ROUGE stands for Recall Oriented Understudy for Gisting Evaluation. Automatic summarization and machine translation are evaluated using this set of metrics. It compares the automatic summary against a set of reference summary generated by humans. Different metrics related to the ROUGE metric are ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU. ROUGE-N is overlap of N-grams states, ROUGE-1 is the overlap of 1-gram, ROUGE-L is the longest common subsequence based statistics, ROUGE-W is the weighted longest common subsequence, ROUGE-S is the skip bigram based co-occurrence statistics, ROUGE-SU is the skip bigram plus unigram based co-occurrence statistics. Work [8] has determined the accuracy of their model using this evaluation metric.

4)SPICE. SPICE stands Semantic Propositional Image Caption Evaluation. This metric is based on semantic concepts. In human caption evaluation, semantic propositional content is an important factor thus this metric is based on graph representations called scene-graphs. Details of various objects, attributes and their relationship from the description of the image can be extracted using this graph.

5)CIDEr. CIDEr stands for Compatibility-Based Image Analysis. It is an automated metric compatibility test used for image definitions. The data that is mostly available, only has five captions for each image. The test metrics, previously used, work with few sentences but they are not sufficient to measure the correlation between the captions automatically generated and the human judgment. This evaluation metric is used in [8].

Table-1: A summary of the datasets and evaluation metrics used.

Dataset Used	**Image Captioning Architecture Using them**	**Evaluation Metric Used**
MS COCO	[1,3,4,5,6,7,9,11]	BLEU [1,5,6,7,11] R-precision[3] success@k[3] Meteor[4] mAP[9]
Flickr8K	[1,2,4,6,7]	BLEU[1,2,6,7] Meteor[4]
Flickr30K	[1,4,6,7]	BLEU[1,6,7] Meteor[4]

Corel Image	[10]	Percentage of correctly captioned words [10]
FlickrStyle 10K	[8]	BLEU[8] METEOR[8] ROUGE[8] CIDEr[8]

4. CONCLUSIONS

In this paper, we have given a brief survey about the various techniques used to implement the task of image captioning including their advantages and disadvantages. We have surveyed several image captioning models based on different deep learning concepts. We have also discussed the different datasets and evaluation metrics used in these models and summarised them in a table.

Even though deep learning has made it possible for machines to perform the complex tasks of image captioning, more research needs to be done to improve the accuracy of the captions generated.

REFERENCES

[1] C. Amritkar, V. Jabade, "Image Caption Generation using Deep Learning Technique", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

[2] R. Singh, A. Sharma, "Image captioning using Deep Neural Networks", 10.13140/RG.2.2.36410.08644, May, 2018.

[3] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2015, pp. 2625-2634.

[4] Y. Zhu, X. Li, X. Li, J. Sun,X. Song, S. Jiang, "Joint Learning of CNN and LSTM for Image Captioning", CLEF (Working Notes), pp. 421-427, 2016.

[5] P. Shah, V. Bakrola, S. Pati, "Image Captioning using Deep Neural Architectures", 2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS), 2017, pp. 1-4, doi: 10.1109/ICIIECS.2017.8276124.

[6] Q. Fu, Y. Liu, Z. Xie, "EECS442 Final Project Report", University of Michigan.

[7] Y. Tian, T. Li, "Project Final Report", Stanford University, 2005.

[8] C. Gan, Z. Gan, X. He, J. Gao, L. Deng, "StyleNet: Generating Attractive Visual Captions with Styles", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2017, pp. 3137-3146.

[9] B. Demirel, R. Cinbis, N. Ikizler-Cinbis, "Image Captioning with Unseen Objects",30th British machine vision conference, 2019, arXiv:1908.00047.

[10] J. Pan, H. Yang, P. Duygulu, C. Faloutsos, "Automatic Image Captioning", 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), 2004, pp. 1987-1990 Vol.3, doi: 10.1109/ICME.2004.1394652.

[11] A. Fariha, "Automatic Image Captioning using Multi-task Learning", 29th Conference on Neural Information Processing Systems, Barcelona, Spain, 2019.