International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 03 | Mar -2020 ISSN: 2582-3930

A Survey on Image Captioning Using Neural Networks

Krusha Joshi[1], Vishal Paymode[2], Krutika Shimpi[3], Abhishek Singh[4]

*Information Technology, MET BKC IOE

Abstract:-Describing a particular image or scene with the help of natural language has attracted interests recently both because of its importance in practical applications and because it connects two major artificial intelligence fields: computer vision and natural language processing. On one hand, image caption models need to determine what objects and events are contained in an image. On the other hand, they need to express the relationships properly in a natural language. Machines can learn to abstract various relevant features including spatial attention, channel-wise attention and visual dependence. Image captioning is a very challenging task, but it has great significance such as helping the visual impaired people and building intelligent robots.

Index Terms- Computer Vision, Data Mining, Machine Learning, Artificial Intelligence, Neural Networks, Image Processing, Natural Language Processing.

I. INTRODUCTION

Generating a description of an image is called image captioning. Image captioning requires to recognize the important objects, their attributes and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep learning-based techniques are capable of handling the complexities and challenges of image captioning. In this survey paper, we aim to present a comprehensive review of existing deep learning-based image captioning techniques. We discuss the foundation of the techniques to analyze their performances,strengths and limitations.We also discuss hedatasets and the evaluation metrics popularly used in deep learning based automatic image captioning[1].

A picture may be a price thousand (coherent) words: building a natural description of pictures. It is easier for humans to caption a scene or a picture however if it's asked for the machine to explain the scene like humans will be a difficult task. Caption generation is the difficult computing drawback of generating a human-readable matter description of given a photograph. It needs each image understanding from the domain of vision and a language model. Image captioning may be a method of generating image descriptions for a close understanding of the varied parts of the image. Automatic generation of image captions may be an elementary task to make a bridge between sensory system and system. The language helps to give usable and necessary info from the scenes delineated within the pictures. This results in a stronger perceiving of the scene by generating captions out of pictures and totally understand the data from the photographs. For generation of captions from pictures, tasks that has to be performed 1. Gaining info concerning the photo. 2. Generating sentences to explain the Vision world

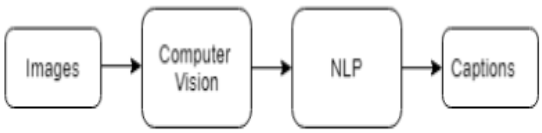

Fig. 1. General Model of an Image Captioning Process.

The first one is associated with image-based model that extracts the features of the image, and next is language-based model that interprets the options

and objects given by our image-based model to a natural sentence. Totally two different domains of Artificial Intelligence are used in which one is Computer Vision and the other is Natural Language Processing is used. The encoder-decoder framework shows promising help in image captioning[4][7]. During this framework, CNN is used for extracting features from an input image to convert into a feature vector and RNN is employed to decipher the feature vector to a English sentence. The attention models helps to boost the encoder-decoder framework by that specialize in relevant image regions. The attention models treat every image as a collection of native regions associated predict an attention chance for every native region. Then they notice regions with massive attention possibilities to calculate a context feature and afterwards feed it into RNN to predict subsequent word.

Every day,weencounteralargenumberofimagesfromvariou ssourcessuchastheinternet,newsarticles,documentdi agramsandadvertisements.Thesesourcescontainima gesthatviewerswould havetointerpretthemselves.Mostimagesdonothavea description,butthehumancanlargely understandthemwithouttheirdetailedcaptions.Howe ver,machineneedstointerpretsomeform ofimagecaptionsifhumansneedautomaticimagecapti onsfromit.
Imagecaptioningisimportantformanyreasons.Forex ample,theycanbeusedforautomaticimageindexing.I mageindexingisimportantforContent-BasedImageRetrievalandtherefore, itcanbeappliedtomanyareas,includingbiomedicine,c ommerce,themilitary,education,digital libraries,andwebsearching.Socialmediaplatformssu chasFacebookandTwittercandirectly generatedescriptionsfromimages.Thedescriptionsca nincludewhereweare(e.g.beach, cafe), whatwewearandimportantlywhatwearedoingthere.

ImagecaptioningisapopularresearchareaofArtificial Intelligence(AI)thatdealswithimage understandingandalanguagedescriptionforthatimag e.Imageunderstandingneedstodetectand recognizeobjects.Italsoneedstounderstandscenetype orlocation,objectpropertiesandtheir interactions.Generatingwell-formedsentencesrequiresbothsyntacticandsemantic understanding ofthe language. Understandinganimagelargelydependsonobtainingi magefeatures.Thetechniquesusedfor thispurposecanbebroadlydividedintotwocategories: (1)Traditionalmachinelearningbased techniquesand(2)Deepmachinelearningbasedtechni ques.Indeepmachinelearningbasedtechniques,featur esarelearnedautomatically fromtrainingdataandtheycanhandlealargeanddivers esetofimagesandvideos[7].

II. LITERATURE SURVEY

Datasets

MSCOCO Dataset. Microsoft COCO Datasetis a very large dataset for image recognition, segmentation, and captioning. There are various features of MSCOCO data set such as object segmentation, recognition in context, multiple objects per class, more than 300,000 images, more than 2 million instances, 80 object categories, and 5 captions per image. Many image captioning methods use the dataset in their experiments.

Flickr30K is a dataset for automatic image description and grounded language understanding. It contains 30k images collected from Flickr with 158k captions provided by human annotators. It does not provide any fixed split of images for training, testing, and validation. Researchers can choose their own choice of numbers for training, testing, and validation. The data set also contains detectors for common objects, a color classifier, and a biastowards selecting larger objects.

Flickr8K Dataset.

Flickr8k is a popular dataset and has 8000 images collected from Flickr. The training data consists of 6000 images, the test and development data, each consists of 1,000 images. Each image in the dataset

has 5 reference captions annotatedbyhumans.Anumberof imagecaptioningmethodshaveperformedexperiment susingthedataset.

Convolutional neural networks(The Encoder) :-
Convolutional neural networks were used to extract features from the pictures. It's comprised of 1 or additional convolution layers (often with a subsampling step) and so followed by one or additional totally connected layers as in an exceedingly normal multilayer neural network. The design of a CNN is intended to require advantage of the 2nd structure of associate input image followed by some type of pooling which ends up in sampling of images. Another advantage of CNNs is that they're easier to train and have several fewer parameters than totally connected networks with identical variety of hidden units and CNN are wide used and studied for image tasks, and state-of-the art for seeing and detection.
CNN are widely used for feature learning, and a classifier such as Soft-max is used for classification. CNN is generally followed by Recurrent Neural Networks (RNN) in order to generate captions[1][4][5].

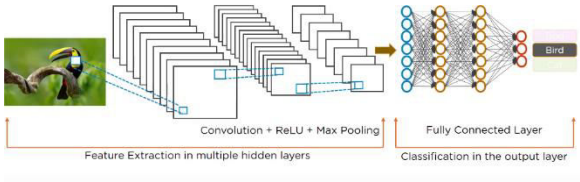

Figure 2: Layers of Convolutional Neural Network.

Recurrent Neural Networks (The Decoder)measure models that have higher performance in several natural language processing tasks. If you wish to predict subsequent word in an exceedingly sentence you've got to grasp that words came before it. RNNs is known as continual as a result of same task is performed for each component within the sequence, and also the output is relied on the previous computations. Instead, RNNs is thought of as networks that have a "memory" that captures info concerning what has been calculated[7][10].

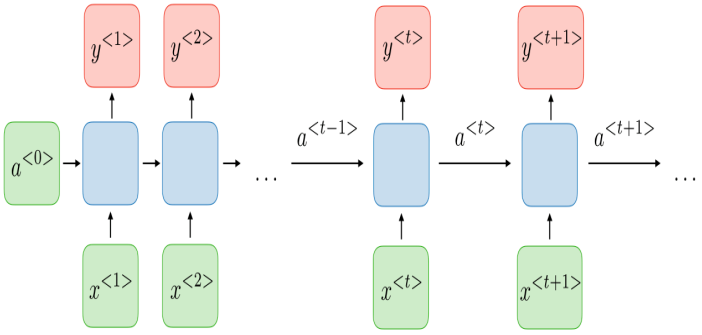

Figure 3 :An unrolled RNN Network.

Issues of Standard RNN:-

There are two major obstacles RNN's have or had to deal with.

- Exploding Gradient Problem.
- Vanishing Gradient Problem.

The real problem isn't backpropagation, it's the fact that long-range dependencies are really complicated. The memorization task exemplifies how the issue can arise in RNNs. It is necessary for the network to read and memorize the input sequence, then give it out again. This method forces it to discover and benefit it to know its capacity. As, it is quite hard to catch long term dependencies because of multiplicative gradient. The multiplicative gradient can be exponentially decreasing/increasing with respect to the number of layers. But to understand them, knowledge about gradient is important. A gradient is a partial derivative with respect to its inputs that is it measures how much the output of a function changes, if you change the inputs a little bit. It simply measures the change in weights with regard to the change in error. In exploding gradient problem, algorithm assigns a stupidly high importance to the weights, without much reason. This problem can be easily solved if you truncate or clip the gradient. Vanishing gradient occurs due to lower values of gradients and because of that model stops learning or takes way too long because of that. This is a major problem with RNN and is much harder to solve than exploding gradient

problem. Two strategies for dealing with exploding/vanishing gradients:

- Minimize the long-distance dependencies.
- Keep the network's transformation close to the identity function.

For minimization and too handle vanishing gradient problem the birth of architectures like the LSTM (Long Short-term memory). GRUs are further improved version of standard recurrent neural network and are used to solve the vanishing gradient problem of a standard RNN. GRU also uses, so called, update gate and reset gate. But LSTM are more powerful than GRUs. LSTM training is relatively short as it keeps the gradients steep enough and the accuracy is high[7][8][10].

LSTM (Long-short Term Memory)

LSTM networks (Long short-run Memory) was the kind of RNN used. Long remembering networks – sometimes simply known as "LSTMs" –is a special quite RNN, want to model the words dependency and generate the language sentence. they're capable of learning long-run dependencies and might handle the matter of vanishing gradient that happens in RNN. LSTMs square measure expressly designed for the aim of basic cognitive process info for long periods of your time. Through this survey, it'll be useful to supply semantically and visually grounded description of abstract pictures, the projected description of which might be in linguistic communication i.e. human perceived description. By investing the techniques like CNN, RNN, and information sets like those of MS-COCO, it strives to achieve the human level perception of given pictures[8].

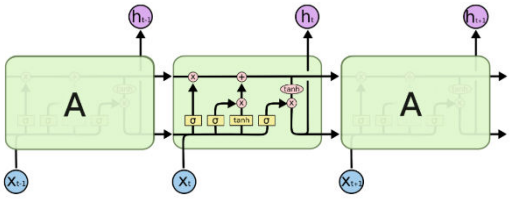

Figure 4: The repeating module in a LSTM contains four interacting layers.

III. CONCLUSION

In this survey, we discussed recent advances in automatic image description generation. This helps us to know CNN works and how it can be used for feature abstraction. The feature abstracted from the input image will be given as input to the decoder part i.e. RNN and which will further help in Image captioning. The most important layer in CNN is convolution layer which takes most of the time within the network. Network performance also depends on the number of levels within the network. But in the other hand as the number of levels increases the time required to train and test the network. Today the CNN consider as power full tool within machine learning for a lot of application such as face detection and image, video recognitions and voice recognition.

Recurrent Neural Networks are the state-of-the-art algorithm for sequential data.

REFERENCES

1) Object detections and attribute discovery", (Kulkarni, Li, Yang , Mitchell, Elliott and Keller (2013).

2) "Sequence to sequence training with neural networks for machine translation",(Cho et al; Bahdanau et al; Sutskever et al, (2014).

3) Deep captioning with multimodal recurrent neural networks (m-rnn)". In ICLR,J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille (2015).

4) One major reason image caption generation is well suited to the encoder-decoder framework" (Cho et al., 2014).

5) Show and tell: a neural image caption generator". In CVPR, O. Vinyals, A. Toshev, S. Bengio, and D. Erhan (2015).

6) Professor forcing: A new algorithm for training recurrent networks". A. Lamb, A. Goyal, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. In NIPS, (2016).

7) Deep visual-semantic alignments for generating image descriptions". IEEE Trans. Pattern Anal.A. Karpathy and L. F. Fei. Mach. Intell., 39(4):664{676, Apr. (2017).

8) LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems",KlausGreff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jurgen Schmidhuber.28(10): 2222{2232, (2017).

9) Learning to guide decoding for image captioning". W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu. In AAAI, (2018).

10) Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present", Xinpeng Chen Lin Ma Wenhao Jiang Jian Yao Wei Liu Wuhan University Tencent AI Lab (2018).